PhD-FSTM-2025-056
The Faculty of Science, Technology and Medicine

DISSERTATION

Defence held on 27/05/2025 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *Mathématiques*

by

CÉDRIC NOEL
Born on 2 March 1978 in Thionville (France)

ADVANCED FINITE MIXTURE MODELING WITH TRAJER: METHODS AND APPLICATIONS FOR TRAJECTORY ANALYSIS

Dissertation defence committee

Dr Jean SCHILTZ, dissertation supervisor
Professor, Université du Luxembourg

Dr Yannick BARAUD, Chairman
Professor, Université du Luxembourg

Dr Mark PODOLSKIJ
Professor, Université du Luxembourg

Dr Mohamed NADIF
Professor, Université Paris Cité

Dr RAFIK ABDESSELAM
Professor, Université Lumière Lyon 2

Acknowledgments

I would like to express my deepest gratitude to those who have supported me throughout the journey of completing this thesis. This work would not have been possible without the encouragement, guidance, and love of the following individuals.

First and foremost, I am profoundly thankful to my beloved wife, Aurore, whose unwavering support, patience, and understanding have been my rock during this challenging process. Your endless encouragement and belief in me kept me going, even in the most difficult moments. I am so lucky to have you by my side.

To my wonderful children, Louise, Soren, and Elisa, thank you for bringing joy and light into my life. Your smiles, laughter, and boundless energy gave me the strength to persevere. I hope one day you will be proud of this accomplishment, as it is as much for you as it is for me.

I am also deeply grateful to my thesis advisor, Professor Jang Schiltz, for his invaluable guidance, expertise, and dedication. Your insightful feedback, encouragement, and willingness to share your knowledge have been instrumental in shaping this research. Working under your supervision has been a privilege and an honor.

Finally, I extend my thanks to all the family, friends, and colleagues who have supported me along the way. Your encouragement and belief in me have made all the difference.

Abstract

This thesis presents **`trajeR`** , an innovative R package designed for advanced finite mixture modeling in longitudinal trajectory analysis. **`trajeR`** addresses the challenge of identifying latent subgroups in heterogeneous longitudinal data by integrating specialized distributions, including Zero-Inflated Poisson (ZIP), Censored Normal (CNORM), Logit, and Beta models, to capture diverse trajectory patterns. A dedicated chapter on multivariate finite mixture models extends the framework to handle complex, multi-dimensional longitudinal data, enabling joint analysis of multiple outcomes and their interdependencies. Methodological contributions include enhanced Expectation-Maximization (EM) algorithms, robust standard error estimation, and rigorous identifiability criteria for mixture models, supported by numerical techniques such as Iteratively Reweighted Least Squares and quasi-Newton optimization. Applied to real-world datasets in fields like criminology, medicine, and finance, **`trajeR`** uncovers meaningful subgroups and predictors of trajectory group membership. Model selection criteria, including AIC and BIC, ensure optimal clustering, while techniques like dynamic time warping enhance trajectory analysis accuracy. **`trajeR`** provides a flexible and computationally efficient tool for researchers, with broad applications in epidemiological studies, behavioral trajectory modeling, and multivariate longitudinal data analysis.

Contents

List of Figures

List of Tables

Acronyms

AIC Akaike Information Criterion

AvePP Average Posterior Probability

BFGS Broyden Fletcher Goldfarb Shanno

BHHH Berndt Hall Hall Hausman

BIC Bayesian Information Criterion

CH Calinski and Harabasz

CHDTW Calinski-Harabasz-DTW

CHF Calinski-Harabasz-Fréchet

DTW Dynamic Time Warping

EM Expectation Maximization

GBTM Group Based Trajectory Model

GMM Growth Mixture Modeling

IRLS Iteratively Reweighted Least Squares

LCGA Latent Class Growth Analysis

LGM Latent Growth Model

MLE Maximum Likelihood Estimation

MVL Multivariate Logit

OCC Odds of Correct Classification

SH Slope Heuristics criterion

SVD Singular Value Decomposition

ZIP Zero Inflated Poisson

CHAPTER 1

Introduction

Time series are commonly encountered in various fields. For instance, in finance, researchers may track employees' salaries over multiple time periods Schiltz (n.d.), while in criminology, the frequency of physical aggression may be measured over time Daniel S. Nagin and Richard E. Tremblay (2005). In medicine, electroencephalography readings can be collected at various time points Elmer, Bobby L. Jones, Zadorozhny, et al. (2019).These longitudinal data often exhibit a typical evolution over time for the whole sample or different subgroups. Many existing models aim to understand the variability of the data of individuals in relation to these average trends over time.

In some cases, natural clusters may exist within the data, while in others, researchers might create clusters based on their prior knowledge. However, forming clusters based on prior assumptions can lead to errors and overlook interesting behaviors or groups within the data. The method developed in this thesis does not assume specific groups but instead identifies them while estimating the average trajectories.

The family of Latent Growth Model (LGM) is used to study both inter-individual (between subjects) and intra-individual (within subjects) patterns of change over time. These patterns may be represented by time trends or latent trajectories and can be influenced by one or several outcomes.

A challenge in this field is that a lot of concepts have different names and notations throughout the literature. The partial glossary in Figure 1.1, inspired by Van Der Nest et al. (2020), provides a reference for some of these terms, and readers seeking more detailed information can refer to the original article.

In this work, we primarily focus on the Latent Class Growth Analysis (LCGA) and Group Based Trajectory Model (GBTM) families developed by Nagin and colleagues (Daniel S Nagin and Richard E Tremblay (2001)). The figure 1.1 show a representation of the LGM familly. GBTM can be seen as a special case of LCGA where the variances are equal across time and

groups. Another noteworthy family of LGM is Growth Mixture Modeling (GMM), developed by Muthen and colleagues (Muthén (2006) orMuthén (2004)). In GMM, within-class variation of individuals is allowed for latent trajectory classes, whereas in LCGA, no variation across individuals is allowed within classes. In a way, LCGA aligns with the classical definition of a group, where each individual within a group is expected to look the same.
GBTM introduces K latent classes, and an individual's assignment to a particular class is based on the degree of similarity in their developmental course compared to other individuals. This model does not account for between-subject variability within a class. Nagin's original formulation assumes that the error variance is the same in each class. For more flexibility, we allow the variance to vary independently in each class. This can provide a more realistic representation of heterogeneity within classes. For a detailed discussion of the differences between these two methods, you can refer to Nagin's work (Daniel S. Nagin and Richard E. Tremblay (2005)).

The present thesis is organized as follow.

In Chapter 2, we will explore essential methods that play a foundational role in this dissertation. While some of these methods are considered classics, it is always valuable to revisit them for the benefit of the reader.

In Chapter 3, which focuses on distributions, we will provide detailed definitions of mixture models for each distribution. We will also present the likelihood formulas and discuss parameter estimation techniques, utilizing both quasi-Newton methods and the Expectation Maximization (EM) algorithm. It's important to note a significant departure from Nagin's approach: first, we employ both likelihood methods and the EM algorithm, and second, we provide explicit formulas used in our software. For likelihood-based methods, the Hessian will be calculated through differentiation or extracted from the algorithm used. In the case of the EM algorithm, we will determine the Hessian following the method outlined by Louis, Louis (1982).

In Chapter 4 will delve into a fundamental issue in statistical modeling: the identifiability of parameters. Building upon two key theorems, the first established by Teicher (1963a) and the second applied by C. Hennig (2000), we will generalize these theorems to the longitudinal context and apply them to the distributions used in this dissertation.

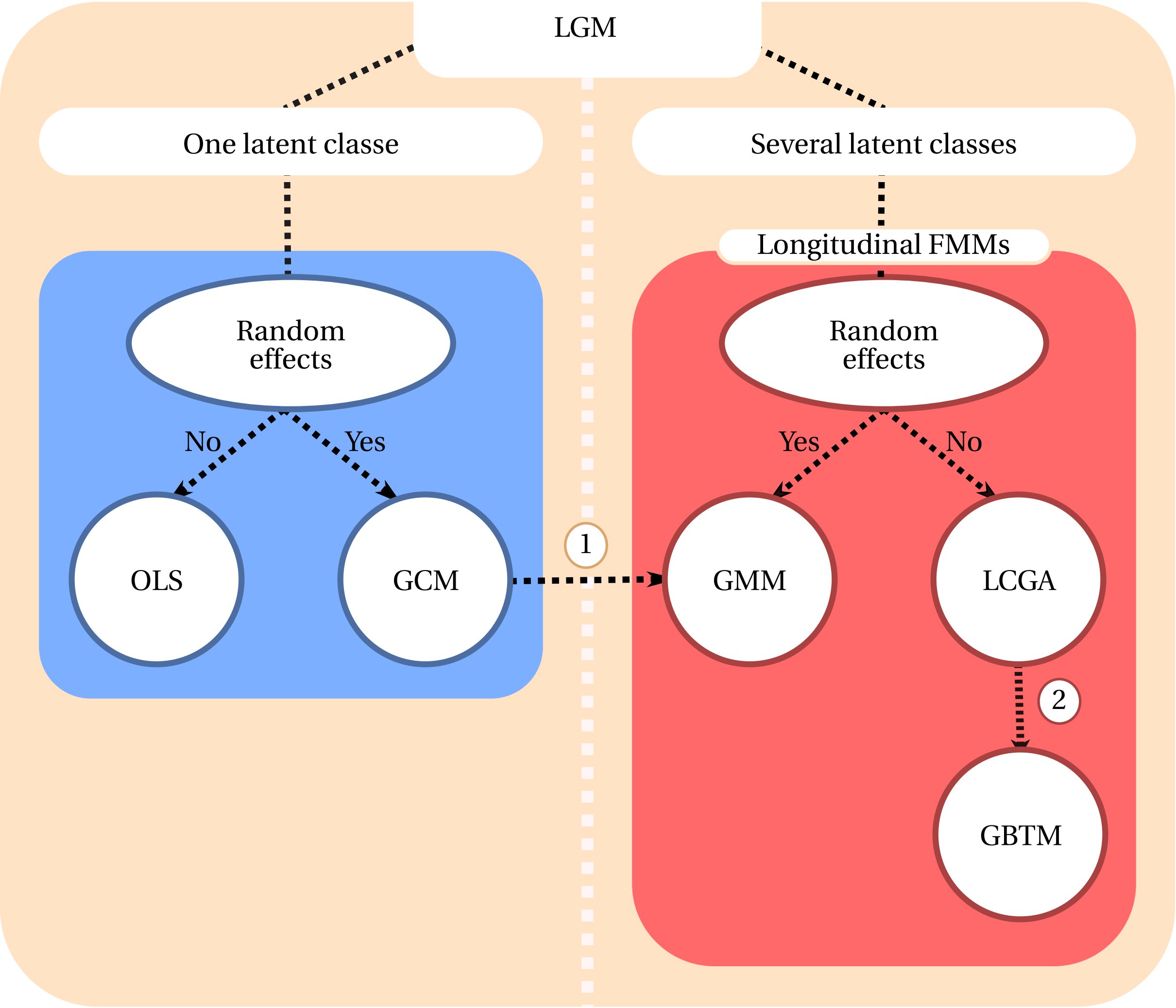


Figure 1.1: LGM: Latent Growth Model – OLS: Ordinary Least Squares – GCM: Growth Curve Model also know as latent growth curve model (LGCM), mixed model, multilevel model, hierarchical model, latent growth factor model, random effect model, latent curve model – FMM: Finite Mixture Model, also know as latent class models, unsupervised learning models – GMM: growth mixture model also know as latent class growth models (LCGM), latent class linear mixed models (LCLMM), latent class mixture model (LCMM) and generalized growth mixture modeling (GGMM) – LCGA: Latent Class Growth Analysis – GBTM: Group-Based Trajectory Models.
① at least 2 latent classes, each with its own GCM.
② Restriction: equal variance across time and classes

In Chapter 5, we will discuss the critical topic of selecting appropriate starting values for modeling. Inadequate starting values can hinder algorithm convergence or lead to incorrect outcomes. While there is no universal method to determine the best starting values, we will elucidate the rationale behind our default choices, while also providing users the flexibility to choose their own starting values.

In Chapter 6 will present various methods for model selection. In addition to traditional techniques like Schwarz (1978) and AIC Akaike (1974), we will introduce and analyze alternative approaches, including the Ratio Test, Neyman and Pearson (1928) or Wald (1941), Slope heuristics, Baudry et al. (2012) or Birgé and Massart (2007), and non-parametric indices like the Calinski and Harabasz criterion, Calinski and Harabasz (1974). Once a model is chosen, we will provide insights into assessing its quality through specific calculations.

In Chapter 7, we will introduce a significant contribution to this dissertation: a novel method for considering multiple trajectory outcomes. This is especially relevant in practical situations where multiple outcomes are measured for each individual, necessitating the simultaneous consideration of all these trajectories.

In Chapter 8, we will introduce our R package, called **`trajeR`** , which implements all the necessary functions for practical use.

Concluding the dissertation, in 9, we will provide examples of how to apply this package to real-world data, demonstrating its utility and effectiveness in real data analysis.

CHAPTER 2

Finite Mixture Models

2.1 Finite Mixture Models

Consider a time-varying variable of interest Y defined in a population of size N. Let $Y_i = (y_{i_1}, \cdots, y_{i_T})$ be T measures of the variable Y, taken at times $t_1, \cdots, t_T$ for subject number i belonging to a sample of size n and suppose that population is divided into K homogeneous sub-populations $C_1, ..., C_K$. We note $C_i = k$ the fact that individuals i belongs to cluster C_k. Furthermore, we suppose conditional independence for the sequential realizations of the elements y_{i_t} over the T periods of measurements, i.e. that a value y_{i_t} doesn't dependent on the past values $y_{i_{t'}}$, $t' < t$. Thus the individual trajectories only depend on the typical trajectory of a given group and not on the past values of the individuals.

Let $P^k(Y_i)$ be the probability of Y_i given membership in group C_k and $P(Y_i)$ the unconditional probability of observing the realization Y_i of Y. Then,

$$P(Y_i) = \sum_{k=1}^{K} P\left(C_i = k\right) P^k(Y_i) \tag{2.1}$$

$$P(Y_i) = \sum_{k=1}^{K} P\left(C_i = k\right) \prod_{t=1}^{T} P^k(Y_{it}) \tag{2.2}$$

and finally

$$P(Y) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} P\left(C_i = k\right) \prod_{t=1}^{T} P^k(Y_{it}) \right) \tag{2.3}$$

We can rewrite equation 2.3 in terms of densities. The density f of Y can be described as a mixture model, see Daniel S. Nagin (2005):

$$f(y; \psi) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} g_k(y_{it}; \Theta_k) \right) \tag{2.4}$$

where g_k represents the conditional density of Y_i given membership in cluster C_k

Here, the group sizes $\pi_k > 0$ denote the probability of a given subject belonging to group k, and they satisfy the condition $\sum_{k=1}^{K} \pi_k = 1$. The parameters θ_k describe the shape of the trajectories in group k, making the model dependent on the parameters
$\psi = (K, \pi_1, \ldots, \pi_{K-1}, \Theta_1, \ldots, \Theta_K)$.

Y depends on time measurements. For an individual i, these time measurements are denoted as a_{it}. Let A be the matrix of measurements for each individual and each time, $A = (a_{it})_{it}$, where $1 \leq i \leq n$ and $1 \leq t \leq T$.

If we also assume that the trajectories of Y are influenced by a static set of risk variables $X = (X_1, \ldots, X_{n_\theta})$, as well as by a time-dependent covariate $W = \left(W^1, \ldots, W^{n_\delta}\right)$ which is independent of X, we can denote n_θ and n_δ as integers. For $1 \leq j \leq n_\theta$, $X_j = (x_{1j}, \ldots, x_{nj})'$ for some real values x_{ji}, where $1 \leq i \leq n$. Therefore,

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1n_\theta} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nn_\theta} \end{pmatrix}$$

and for $1 \leq j \leq n_\delta$, W^j is a real matrix composed by $\left(w_{it}^j\right)_{it}$ where $1 \leq i \leq n$ and $1 \leq t \leq T$ and $w_{it}^j \in \mathbb{R}$. Thus,

$$W = \left(\overbrace{\begin{matrix} w_{11}^1 & \cdots & w_{1T}^1 \\ \vdots & & \vdots \\ w_{n1}^1 & \cdots & w_{nT}^1 \end{matrix}}^{W^1} \quad \cdots \quad \overbrace{\begin{matrix} w_{11}^{n_\delta} & \cdots & w_{1T}^{n_\delta} \\ \vdots & & \vdots \\ w_{n1}^{n_\delta} & \cdots & w_{nT}^{n_\delta} \end{matrix}}^{W^{n_\delta}} \right)$$

We can write the probability 2.3 as

$$\begin{aligned} P(Y) &= \prod_{i=1}^{n} \left[\sum_{k=1}^{K} P\left(Y_i = y_i | X_i = x_i, W_i = w_i, C_i = k\right) P\left(C_i = k | X_i = x_i, W_i = w_i\right) \right] && (2.5) \\ &= \prod_{i=1}^{n} \left[\sum_{k=1}^{K} P\left(Y_i = y_i | W_i = w_i, C_i = k\right) P\left(C_i = k | X_i = x_i\right) \right] && (2.6) \\ &= \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \left(P\left(C_i = k | X_i = x_i\right) \prod_{t=1}^{T} P\left(Y_{i_t} = y_{i_t} | X_i = x_i, W_i = w_i, C_i = k\right) \right) \right] && (2.7) \end{aligned}$$

We can modelize the effect of the variable X_i on the probability of group membership by a generalized LOGIT function.

$$P\left(C_i = k | X_i = x_i\right) = \frac{e^{\theta_k x_i}}{\sum\limits_{k=1}^{K} e^{\theta_k x_i}} \tag{2.8}$$

where $\theta_k = \left(\theta_{k_1}, \cdots, \theta_{k_{n_\theta}}\right)$ denotes the effect of $x_i = (x_{i_1}, ..., x_{i_{n_\theta}})$ on the probability of group membership for subject i.

The complete conditional density becomes

$$f(y; \psi) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \frac{e^{\theta_k x_i}}{\sum\limits_{k=1}^{K} e^{\theta_k x_i}} \prod_{t=1}^{T} g_k(y_{it}; \Theta_k, w_i) \right) \tag{2.9}$$

and the log-likelihood becomes

$$l(\psi; y) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \frac{e^{\theta_k x_i}}{\sum\limits_{k=1}^{K} e^{\theta_k x_i}} \prod_{t=1}^{T} g_k(y_{it}; \Theta_k, w_i) \right) \tag{2.10}$$

Daniel S. Nagin (2005) introduced three different distributions to model the probability of Y :

- the censored normal distribution ;
- the logistic distribution ;
- the ZIP (Zero Inflated Poisson) distribution.

In this work, among other things, we generalized the Nagin's model by

- allowing the variability to vary inside each group.
- adding beta distribution
- adding non linear link

We summarize the model by the figure 2.1.

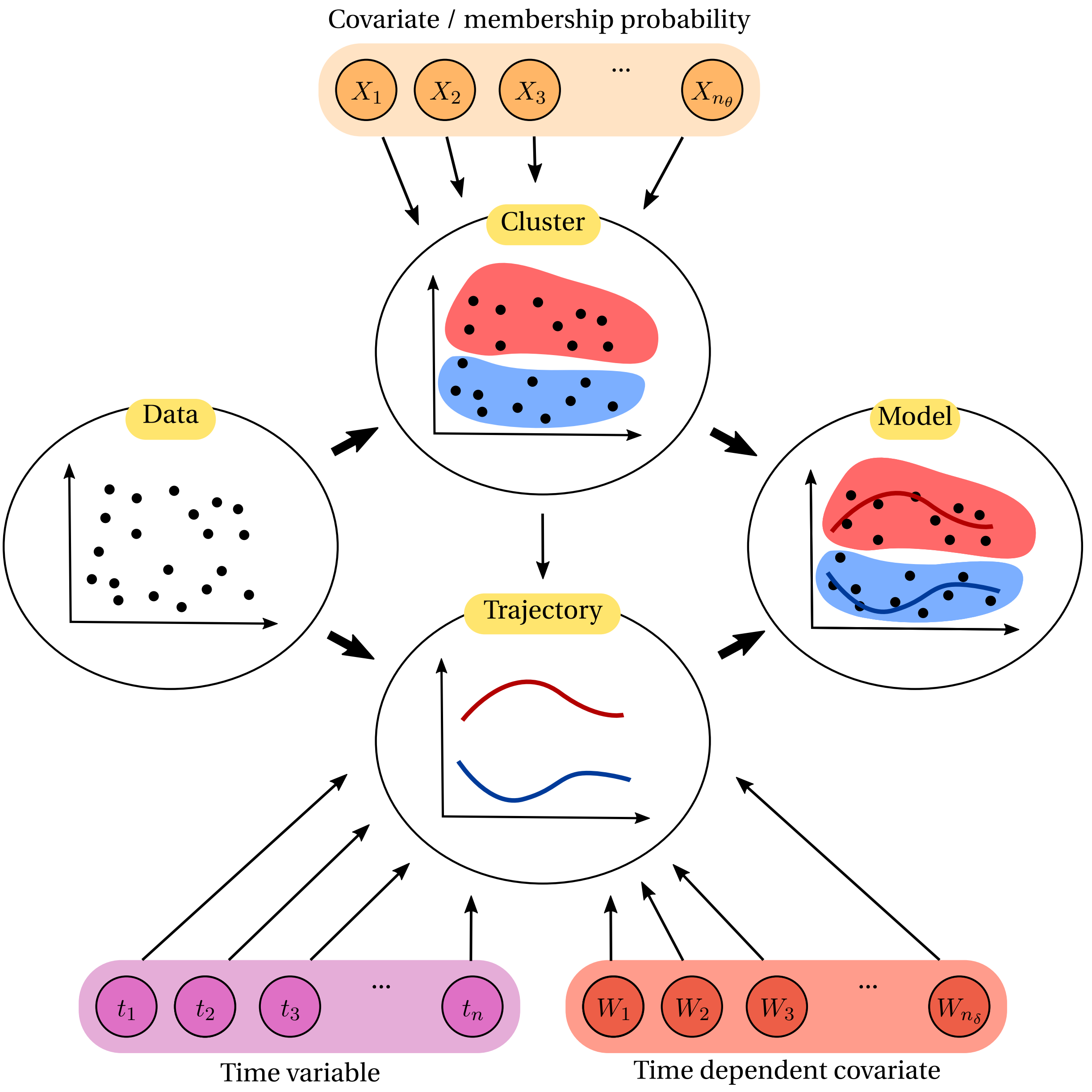


Figure 2.1: The global model

To estimate the parameters of the mixture model, we maximize the likelihood, denoted as L, which is defined by the following equation:

$$L(\psi; y) = \prod_{i=1}^{n} f_i(y_i; \psi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k) \tag{2.11}$$

where $f_i(y_i; \psi) = \sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k, w_i)$, For simplicity, we uses the following notation: $g_k(y_i; \Theta_k) = \prod_{t=1}^{T} g_k(y_{it}; \Theta_k, w_i)$.

Hence, the corresponding log likelihood function is given by

$$l(\psi; y) = \sum_{i=1}^{n} \log f_i(y_i; \psi) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k) \right) \tag{2.12}$$

Finally, in order to maximize the log likelihood function, we define the score function S as follows:

$$S(y; \psi) = \frac{\partial l(\psi; y)}{\partial \psi} \tag{2.13}$$

2.2 Quasi-Newton procedure maximum research routine

We use the classical Maximum Likelihood Estimation (MLE) method W. H. Greene (2003). The first step is to determine the derivative of the log-likelihood function.

$$\frac{\partial l(\psi; y)}{\partial \psi} = \sum_{i=1}^{n} \frac{\partial \log \left(\sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k) \right)}{\partial \psi} \tag{2.14}$$

$$= \sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \left(\frac{\partial \pi_k}{\partial \psi} g_k(y_i; \Theta_k) + \pi_k \frac{\partial g_k(y_i; \Theta_k)}{\partial \psi} \right)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k)} \tag{2.15}$$

As $g_k(y_i; \Theta_k)$ does not depend on θ_k and π_k does not depend on Θ_k, the equations above simplify to:

$$\frac{\partial l(\psi; y)}{\partial \theta_k} = \sum_{i=1}^{n} \frac{\frac{\partial \pi_k}{\partial \theta_k} g_k(y_i; \Theta_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k)} \tag{2.16}$$

$$\frac{\partial l(\psi; y)}{\partial \Theta_k} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial g_k(y_i; \Theta_k)}{\partial \Theta_k}}{\sum_{k=1}^{K} \pi_k g_k(y_i; \Theta_k)} \tag{2.17}$$

We need to compute each part of the equations above. The second part will be discussed in Chapter 3, where we will define each function g_k based on the chosen density. However, the first part remains the same regardless of the density used. Equation (2.8) implies that we can write the log likelihood function as:

$$l(\psi; y) = \sum_{i=1}^{n} \log f_i(y_i; \psi) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \frac{e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}} g_k(y_i; \Theta_k) \right) \tag{2.18}$$

$$= \sum_{i=1}^{n} \log \left(\frac{1}{\sum_{k=1}^{K} e^{\theta_k x_i}} \sum_{k=1}^{K} e^{\theta_k x_i} g_k(y_i; \Theta_k) \right) \tag{2.19}$$

Consequently,

$$\frac{\partial l(\psi; y)}{\partial \theta_j} = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \theta_j} \log \left(\frac{1}{\sum_{k=1}^{K} e^{\theta_k x_i}} \sum_{k=1}^{K} e^{\theta_k x_i} g_k(y_i; \Theta_k) \right) \right) = \sum_{i=1}^{n} \frac{\frac{\partial u}{\partial \theta_j}}{u} \tag{2.20}$$

where $u = \frac{1}{\sum_{k=1}^{K} e^{\theta_k x_i}} \sum_{k=1}^{K} e^{\theta_k x_i} g_k(y_i; \Theta_k)$.

We deduce,

$$\frac{\partial u}{\partial \theta_j} = \frac{-x_i e^{\theta_j x_i} \sum_{k=1}^{K} e^{\theta_k x_i} g_k(y_i; \Theta_k)}{\sum_{k=1}^{K} (e^{\theta_k x_i})^2} + \frac{1}{\sum_{k=1}^{K} e^{\theta_k x_i}} \times x_i e^{\theta_j x_i} g_j(y_i; \Theta_j) \tag{2.21}$$

$$= \frac{-x_i e^{\theta_j x_i} \sum_{k=1}^{K} e^{\theta_k x_i} g_k(y_i; \Theta_k) + x_i e^{\theta_j x_i} g_j(y_i; \Theta_j) \sum_{k=1}^{K} e^{\theta_k x_i}}{\left(\sum_{k=1}^{K} e^{\theta_k x_i} \right)^2} \tag{2.22}$$

$$= \frac{x_i e^{\theta_j x_i} \left(-\sum_{k=1}^{K} e^{\theta_k x_i} g_k(y_i; \Theta_k) + g_j(y_i; \Theta_j) \sum_{k=1}^{K} e^{\theta_k x_i} \right)}{\left(\sum_{k=1}^{K} e^{\theta_k x_i} \right)^2} \tag{2.23}$$

$$= \frac{x_i e^{\theta_j x_i} \left(\sum_{k=1}^{K} e^{\theta_k x_i} \left(g_j(y_i; \Theta_j) - g_k(y_i; \Theta_k) \right) \right)}{\left(\sum_{k=1}^{K} e^{\theta_k x_i} \right)^2} \tag{2.24}$$

Hence

$$\frac{\partial l(\psi;y)}{\partial\theta_j}=\sum_{i=1}^{n}\frac{x_i e^{\theta_j x_i}\sum_{k=1}^{K}e^{\theta_k x_i}\left(g_j(y_i;\Theta_j)-g_k(y_i;\Theta_k)\right)}{\sum_{k=1}^{K}e^{\theta_k x_i}\sum_{k=1}^{K}e^{\theta_k x_i}g_k(y_i;\Theta_k)} \tag{2.25}$$

This algorithm has remarkable properties W. H. Greene (2003). The maximum likelihood estimator $\hat{\psi}$ of the parameters ψ has the following characteristics:

- It is consistent, meaning that the sequence of MLEs converges in probability to the true value of the parameter.

- It is efficient, achieving the Cramér–Rao lower bound as the sample size tends to infinity. This implies that no other consistent estimator is better in terms of having a lower asymptotic mean square error than the MLE.
  We can demonstrate that the MLE satisfies $\sqrt{n}(\hat{\psi}-\psi) \overset{d}{\to} \mathcal{N}(0,\, I^{-1})$, where I represents the Fisher information matrix defined as:

  $$I_{jk}=\mathrm{E}\left[-\frac{\partial^2\ln\left(\prod_{i=1}^{n}f_k(y_i;\psi)\right)}{\partial\psi_j\,\partial\psi_k}\right]$$

  In this expression, ψ represents the parameter vector, and n is the sample size.

Unfortunately, these properties are not guaranteed for finite samples. The log likelihood function is sometimes challenging to express in a simple form. Finding a closed-form solution for the MLE is not always possible, and we often have to resort to Quasi-Newton methods to approximate a solution. Additionally, the derivative of the log of the score function can exhibit multiple local maxima. A poor choice of the initial values can lead to convergence to one of these local maxima rather than the global maximum.

2.2.1 Estimation of the standard errors

In this case, standard errors of the parameters are directly derived from the maximum likelihood method. The Fisher information matrix is defined as:

$$I_\psi=-E\left(\frac{\partial^2\log L}{\partial\psi\partial\psi^t}\right)$$

and the covariance matrix of the parameter ψ is calculated as:

$$Cov=(I_\psi)^{-1}$$

However, there are instances where this matrix may become numerically singular, making it impossible to compute its inverse. In some cases, it may not even be negatively definite due to numerical inaccuracies. In such situations, we propose different approaches:

- Using a generalized inverse of the information matrix, known as the Moore–Penrose inverse Penrose (1955). This is represented by a matrix G that satisfies the condition $I_{\psi}GI_{\psi} = I_{\psi}$. If the original information matrix I_{ψ} is non-singular, the Moore–Penrose inverse is unique and equivalent to the inverse of I_{ψ}.
- Employing an alternative approximation for the Hessian matrix H_{ψ}, such as:
  - The outer product of the gradient, commonly referred to as the Berndt Hall Hall Hausman (BHHH) estimator Berndt et al. (n.d.).
  - The approximate matrix of the inverse of the Hessian, utilized in the Broyden Fletcher Goldfarb Shanno (BFGS) algorithm Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970) (see equation 2.2.2).

Another classical method for estimating the variance of an estimator defined by a function is the delta method Oehlert (1992). We will utilize this method later in the dissertation, and it is worthwhile to revisit some essential definitions and properties.

Theorem 1. *Let be $X_1, \dots, X_n$ a sequence of random variables and $g : \mathbb{R}^d \longrightarrow \mathbb{R}^s$ differentiable in θ.*
If $\sqrt{n}[X_n - \theta] \xrightarrow{L} \mathcal{N}_d(0, \Sigma)$ where $\mathcal{N}_d(0, \Sigma)$ is the centered d-dimensional normal distribution with variance-covariance matrix Σ. We have

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{L} \mathcal{N}_s\left(0, Dg(\theta)\Sigma Dg(\theta)^T\right)$$

where $Dg(\theta)$ is the Jacobian matrix of g in θ.

We apply this theorem to the defintion of the membership probability $\pi_k = \frac{e^{\theta_k}}{\sum_{k=1}^{K} e^{\theta_k}}$ for $\theta_k \in \mathbb{R}$. Let $g : \mathbb{R}^K \longrightarrow \mathbb{R}^K$ the function :

$$g : \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix} \longmapsto \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_K \end{pmatrix}$$

If the parameters θ_k are estimated using MLE, we can compute the variance-covariance matrix of π_k by applying the theorem mentioned earlier and taking into account the fact that:

$$\frac{\partial \pi_k}{\partial \theta_{k'}} = \begin{cases} \pi_k(1-\pi_k) & \text{if } k' = k \\ -\pi_k\pi_{k'} & \text{if } k' \neq k \end{cases}$$

and

$$D(g) = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \cdots & -\pi_1\pi_K \\ \vdots & & & \vdots \\ -\pi_K\pi_1 & \cdots & -\pi_K\pi_{K-1} & \pi_K(1-\pi_K) \end{pmatrix}$$

This formula will be particularly valuable in the algorithms developed in this thesis, especially for determining the standard deviation of the membership probability in cases where covariates are not available.

2.2.2 Newton Raphson Algorithm

In this section, we will review the key concepts of the Newton-Raphson Algorithm and the Quasi-Newton Algorithm, both of which are concretely used in our algorithm.

Let $f : \mathbb{R}^N \to \mathbb{R}$ be a function that is at least twice continuously differentiable. We are seeking the root of the function f. Suppose $x \in \mathbb{R}^N$ and $\Delta x \in \mathbb{R}^N$. Then the first-order Taylor approximation to f at x is given by:

$$f(x+\Delta x) \simeq f(x) + \nabla f(x)^t \, \Delta x$$

The second-order Taylor approximation to f at x is given by:

$$f(x+\Delta x) \simeq f(x) + \nabla f(x)^t \, \Delta x + \frac{1}{2}\Delta x^t H_f(x) \, \Delta x,$$

where ∇f represents the gradient of f and H_f denotes the Hessian of f (i.e., $\nabla^2 f(x) = H_f(x)$).

Taking derivatives with respect to vectors and matrices, and considering the symmetry of H_f, we have:

$$\nabla f(x+\Delta x) \simeq \nabla f(x) + H_f(x)\Delta x \tag{2.26}$$

Since the gradient at the value that maximizes f is a vector of zeros, we can conclude that the maximizer $\hat{\Delta x}$ satisfies:

$$0 = \nabla f(x) + H_f(x)\Delta x$$

This implies:

$$\Delta x = -H_f(x)^{-1}\nabla f(x)$$

In other words, the vector that maximizes the second-order Taylor approximation to f at x is:

$$\Delta x = -H_f(x)^{-1}\nabla f(x)$$

$$x + \hat{\Delta x} = x - H_f(x)^{-1}\nabla f(x) \tag{2.27}$$

With this in mind, we can specify the Newton-Raphson algorithm for N-dimensional function optimization by iteratively using this method, starting with an initial value of $x_0 \in \mathbb{R}^N$.

$$\begin{cases} x_0 \\ x_{k+1} = x_k - H_f(x_k)^{-1}\nabla f(x_k),\ k > 0 \end{cases} \tag{2.28}$$

Newton's method ensures that the sequence of iterates, denoted as x_k, converges. In general, the convergence of this method is quadratic.

However, there are some challenges associated with it. For instance, the choice of the starting value is crucial, as a poor choice may result in non-convergence. Additionally, computing the derivative functions can be difficult, especially in cases where the Jacobian or Hessian matrices are unavailable or too expensive to compute. In such situations, we can turn to the Quasi-Newton method.

In the Quasi-Newton method, we replace the calculation of the Hessian matrix by an approximate matrix denoted as B. This approximation is chosen to satisfy Equation 2.26.

$$\nabla f(x_k + \Delta x) \simeq \nabla f(x_k) + B\,\Delta x,$$

Most methods use a symmetric matrix B and an iterative procedure. At each step of the algorithm, they construct an update matrix B_{k+1} that is close to B in some norm, which can be formulated as:

$$B_{k+1} = \arg\min_{B} \|B - B_k\|$$

To find the next iterate x_k, the Newton's step is utilized, using the approximate Hessian matrix B_k in place of H_f:

- $\Delta x_k = -\alpha_k B_k^{-1}\nabla f(x_k)$, with α chosen to satisfy the Wolfe conditions Wolfe (1969) and Wolfe (1971);

- $x_{k+1} = x_k + \Delta x_k$;

- The gradient computed at the new point $\nabla f(x_{k+1})$, and $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$ is used to update the approximate Hessian B_{k+1}, or directly its inverse $H_{k+1} = B_{k+1}^{-1}$ using the Sherman–Morrison formula Lipovetsky (2009).

The BFGS update the matrix B_k and the approximation of the hessian H_k by

$$B_{k+1} = B_k + \frac{y_k y_k^{\mathrm{t}}}{y_k^{\mathrm{t}}\Delta x_k} - \frac{B_k\Delta x_k(B_k\Delta x_k)^{\mathrm{t}}}{\Delta x_k^{\mathrm{t}} B_k\,\Delta x_k}$$

and

$$H_{k+1} = B_{k+1}^{-1} = \left(I - \frac{\Delta x_k y_k^{\mathrm{t}}}{y_k^{\mathrm{t}} \Delta x_k}\right) H_k \left(I - \frac{y_k \Delta x_k^{\mathrm{t}}}{y_k^{\mathrm{t}} \Delta x_k}\right) + \frac{\Delta x_k \Delta x_k^{\mathrm{t}}}{y_k^{\mathrm{t}}\, \Delta x_k}$$

In the algorithm used in the following, we utilize the BFGS method to find the minimum of f.

2.2.3 IRLS

Iteratively Reweighted Least Squares (IRLS) is a classical method used to solve optimization problems in the form

$$\underset{\beta}{\arg\min}\, ||y - X\beta|| = \underset{\beta}{\arg\min} \sum_{i=1}^{n} \left|y_i - f_i(\boldsymbol{\beta})\right|^p$$

for some p-norm by using iterative step which involves solving a weighted least squares:

At step $t+1$

$$\beta^{(t+1)} = \underset{\beta}{\arg\min} \sum_{i=1}^{n} w_i(\boldsymbol{\beta}^{(t)}) \left|y_i - f_i(\boldsymbol{\beta})\right|^2$$

In the case of f is a linear function, the step $t+1$ become

$$\beta^{(t+1)} = \underset{\beta}{\arg\min} \sum_{i=1}^{n} w_i^{(t)} \left|y_i - X_i \boldsymbol{\beta}\right|^2 = (X^{\mathrm{t}} W^{(t)} X)^{-1} X^{\mathrm{t}} W^{(t)} \mathbf{y}$$

where $W^{(t)}$ is the diagonal matrix of weights.

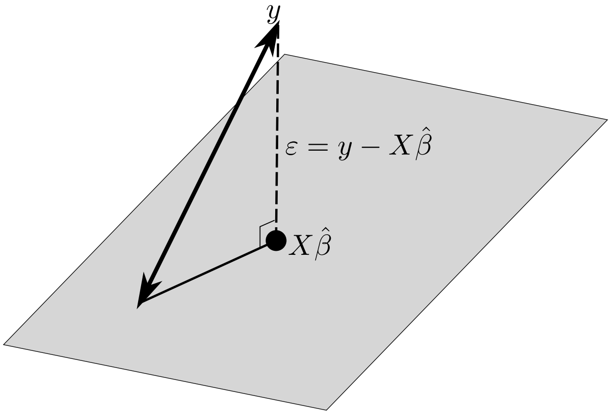
Suppose we have $y = X\beta + \varepsilon$ and the random variables ε_i are such that

- they have mean zero: $E(\varepsilon_i) = 0$;
- they are homoscedastic, that is all have the same finite variance: $var(\varepsilon_i) = \sigma^2 < +\infty$ for all i ;
- distinct error terms are uncorrelated: $(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$.

Thus, the Gauss–Markov theorem tells us that the estimator $\hat{\beta}$ obtained by Weighted Least Squares is BLUE (Best Linear Unbiased Estimator). In the case of a linear model $y = X\beta + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0; \sigma V)$, the Weighted Least Squares is also the Maximum Likelihood estimation.

From a geometric point of view, we consider an expectation plane formed by all vectors in the form of Xa, and we search for the vector $\hat{\beta}$ that minimizes the distance (in this case, the weighted Euclidean distance or the Mahalanobis distance) between this plane and y. Thus, $y - X\hat{\beta}$ must be orthogonal to the expectation plane.

In other words, $\hat{\beta}$ is the projection of y onto the expectation plane.



2.2.4 Matrix decomposition

In order to determine the approximate value of β in the IRLS method, we need to calculate the inverse of $(X^{\mathrm{T}}W^{(t)}X)$. To prevent computational difficulties, it can be advantageous to substitute the inverse calculation with a decomposition of the matrix $(X^{\mathrm{T}}W^{(t)}X)$. This can be accomplished through the use of either Singular Value Decomposition (SVD) or, as explained here, QR.

Within the weighted least square approach, we encounter

$$(X^{\mathrm{t}}WX)\beta = X^{\mathrm{t}}W\mathbf{y} \Leftrightarrow \left(W^{\frac{1}{2}}X\right)^t \left(W^{\frac{1}{2}}X\right)\beta = \left(W^{\frac{1}{2}}X\right)^t W^{\frac{1}{2}}\mathbf{y}$$

$W^{\frac{1}{2}}X$ can be expressed as the product of two matrices, Q and R. Matrix Q is orthogonal, meaning that its transpose multiplied by itself equals the identity matrix. Matrix R is an upper-triangular matrix, specifically of the form $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$.

Thus

$$W^{\frac{1}{2}}X = QR = (Q_1, Q_2)\begin{pmatrix} R_1 \\ 0 \end{pmatrix} = Q_1R_1$$

Consequently

$$\left(W^{\frac{1}{2}}X\right)^t \left(W^{\frac{1}{2}}X\right)\beta = \left(W^{\frac{1}{2}}X\right)^t W^{\frac{1}{2}}\mathbf{y} \Leftrightarrow \beta = \left[\left(W^{\frac{1}{2}}X\right)^t \left(W^{\frac{1}{2}}X\right)\right]^{-1} \left(W^{\frac{1}{2}}X\right)^t W^{\frac{1}{2}}\mathbf{y}$$
$$\Leftrightarrow W^{\frac{1}{2}}X\beta = W^{\frac{1}{2}}X \left[\left(W^{\frac{1}{2}}X\right)^t \left(W^{\frac{1}{2}}X\right)\right]^{-1} \left(W^{\frac{1}{2}}X\right)^t W^{\frac{1}{2}}\mathbf{y}$$

By using the QR decomposition

$$\begin{aligned} W^{\frac{1}{2}}X\beta &= Q_1R_1\left[R_1^tQ_1^tQ_1R_1\right]^{-1} R_1^tQ_1^t\mathbf{y} \\ &= Q_1R_1R_1^{-1}\left(R_1^t\right)^{-1} R_1^tQ_1^t\mathbf{y} \end{aligned}$$

$$
\begin{aligned}
&= Q_1 Q_1^t \mathbf{y} \\
&= Q \begin{pmatrix} Z_1 \\ 0 \end{pmatrix}
\end{aligned}
$$

where $Z_1 = Q_1^t \mathbf{y}$.

We can thus find $\hat{\beta}$ without computing a matrix inverse.

$$
\begin{aligned}
W^{\frac{1}{2}} X\beta &= Q \begin{pmatrix} Z_1 \\ 0 \end{pmatrix} \\
\Rightarrow \quad QR\beta &= Q \begin{pmatrix} Z_1 \\ 0 \end{pmatrix} \\
\Rightarrow \quad Q^t QR\beta &= Q^t Q \begin{pmatrix} Z_1 \\ 0 \end{pmatrix} \\
\Rightarrow \quad R\beta &= \begin{pmatrix} Z_1 \\ 0 \end{pmatrix} \\
\Rightarrow \quad R_1\beta &= Z_1
\end{aligned}
$$

It is a simple task to solve for β by recalling that R_1 is an upper-triangular matrix.

2.3 Expectation maximization algorithm

2.3.1 General framework

To increase the chances of success, we can employ the EM algorithm (EM algorithm)Dempster et al. (1977) Foulley (n.d.[a])Roche (n.d.)). This algorithm is particularly effective for addressing issues related to missing data.

Let's consider a dataset, denoted as $U = (Y, Z)$, where Y represents the known portion of U and Z represents any missing information. U is referred to as the complete data. We can define f as the density function of Y and h as the density function of U. Additionally, let g be the conditional density of Z given Y, and let ψ represent a set of parameters. Using these definitions, we can express $h(U;\psi)$ as$f(Y;\psi)g(Z|Y;\psi)$. Finally, we can calculate the log-likelihood as l.

$$l(U;\psi) = l(Y;\psi) + \log(g(Z|Y;\psi)) \tag{2.29}$$

which implies

$$l(Y;\psi) = l(U;\psi) - \log(g(Z|Y;\psi)) \tag{2.30}$$

The EM algorithm involves maximizing the log likelihood by iterating a sequence $\psi^{(t)}$, beginning with an initial value $\psi^{(0)}$, until it converges to ψ. During each step t, if we calculate the conditional expectation of equation (2.30) given Y, we obtain the following.

$$\begin{aligned} l(\psi;Y) &= E_{\psi^{(t)}}\left(l(U;\psi)|Y\right) - E_{\psi^{(t)}}\left(\log(g(Z|Y;\psi))|Y\right) && (2.31)\\ &= Q\left(\psi;\psi^{(t)}\right) - H\left(\psi;\psi^{(t)}\right) && (2.32)\end{aligned}$$

Thus, to optimize the log likelihood, our goal is to maximize the function Q.
When we reach iteration $(t+1)$, $\psi^{(t+1)}$ represents the value that maximizes the function $\psi \mapsto Q\left(\psi;\psi^{(t)}\right)$. In other words, $\psi^{(t+1)}$ is the argument that gives the maximum value for $Q\left(\psi;\psi^{(t)}\right)$.
Consequently, we can conclude that $Q\left(\psi^{(t+1)};\psi^{(t)}\right) \geq Q\left(\psi^{(t)};\psi^{(t)}\right)$ and by applying the Jensen inequality, we find that $H\left(\psi^{(t+1)};\psi^{(t)}\right) \leq H\left(\psi^{(t)};\psi^{(t)}\right)$. Ultimately, this leads us to the conclusion that

$$l(\psi^{(t+1)};Y) \geq l(\psi^{(t)};Y). \tag{2.33}$$

Thus, the EM algorithm consists in 2 steps:

- E step : computation of $Q\left(\psi;\psi^{(t)}\right)$;
- M step : computation of $\psi^{(t+1)} = \arg\max_{\psi} Q\left(\psi;\psi^{(t)}\right)$.

We iterate these 2 steps until $\left|\psi^{(t+1)} - \psi^{(t)}\right|$ is below a given threshold.

2.3.2 GBTM framework

It is important to note that the index (t) used in the EM algorithm is not the same as the index t representing time in GBTM.
Let $Z_i = (Z_{ik})_{k=1...,K}, i = 1,...,n$ denote a random variable that indicates the membership of subjects in a particular group (which is typically unknown in practice).

$$Z_{ik} = \begin{cases} 1 \text{ if } C_i = k \\ 0 \text{ otherwise.} \end{cases} \tag{2.34}$$

we can represent Z_i as a vector with K coordinates, where only one coordinate is 1 and the rest are 0. The values of Z_i are determined by a multinomial distribution with parameters

$(1, \pi_1, \cdots \pi_K)$, where π_K is calculated as $1 - \sum_{k=1}^{K-1} \pi_k$.

$$P(Z_i) = P(Z_i = z_i) = P\left(Z_{i1} = z_{i1}, \cdots, Z_{iK} = z_{iK}\right) = \prod_{k=1}^{K} \pi_k^{z_{ik}}$$

The vector Z_i allows us merely to write the likelihood of the complete sample in a simplified way.

$$\begin{aligned} P(Y_1, \cdots, Y_n, Z_1, \cdots, Z_n) &= P(Y_1, \cdots, Y_n | Z_1, \cdots, Z_n) P(Z_1, \cdots, Z_n) && (2.35) \\ &= \prod_{i=1}^{n} P\left(Y_i | Z_1, \cdots, Z_n\right) \prod_{i=1}^{n} P(Z_i) && (2.36) \end{aligned}$$

Since Y_i only dependent on Z_i,

$$P\left(Y_i | Z_1, \cdots, Z_n\right) = P\left(Y_i | Z_i\right) = \prod_{k=1}^{K} P\left(Y_i | Z_{ik} = z_{ik}\right)^{z_{ik}} = \prod_{k=1}^{K} g_k(y_i; \Theta_k)^{z_{ik}}$$

By replacing in (2.36), we finally get,

$$L(\psi; y) = \prod_{i=1}^{n} \prod_{k=1}^{K} g_k(y_i; \Theta_k)^{z_{ik}} \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}} = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\pi_k g_k(y_i; \Theta_k)\right)^{z_{ik}} \tag{2.37}$$

Thus, the log likelihood function l is given by

$$l(\psi; y) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log\left(\pi_k g_k(y_i; \Theta_k)\right) \tag{2.38}$$

The function Q become

$$Q\left(\psi; \psi^{(t)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} E_{\psi^{(t)}}(z_{ik} | Y_i = y_i) \log\left(\pi_k g_k(y_i; \Theta_k)\right) \tag{2.39}$$

But,

$$\begin{aligned} E(Z_{ik} | Y_i = y_i) &= P(Z_{ik} = 1 | Y_i = y_i) && (2.40) \\ &= \frac{P(Y_i = y_i | Z_{ik} = 1) P(Z_{ik} = 1)}{\sum_{k=1}^{K} P(Y_i = y_i | Z_{ik} = 1) P(Z_{ik} = 1)} && (2.41) \\ &= \frac{\pi_k g_k(y_i, \Theta_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i, \Theta_k)} && (2.42) \\ &= \tau_{ik} && (2.43) \end{aligned}$$

Hence,

$$E_{\psi^{(t)}}(z_{ik}|Y_i = y_i) = \tau_{ik}^{(t)} = \frac{\pi_k^{(t)} g_k(y_i, \Theta_k^{(t)})}{\sum\limits_{k=1}^{K} \pi_k^{(t)} g_k(y_i, \Theta_k^{(t)})} \tag{2.44}$$

and

$$Q\left(\psi; \psi^{(t)}\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \log\left(\pi_k g_k\left(y_i; \Theta_k\right)\right) \tag{2.45}$$

In the M step, our objective is to maximize the function $\psi \mapsto Q\left(\psi; \psi^{(t)}\right)$ in order to obtain the updated value $\psi^{(t+1)}$.

Initially, we determine the updated value of $\pi_k^{(t+1)}$. In an intuitive sense, if we knew all the z_{ik} values, π_k would be equivalent to the sum of all indicators indicating which individuals belong to group k, $\pi_k = \frac{\sum_{i=1}^{n} z_{ik}}{n}$. Therefore, during iteration $(t+1)$,

$$\pi_k^{(t+1)} = \frac{\sum\limits_{k=i}^{n} \tau_{ik}^{(t)}}{n} \tag{2.46}$$

To address this situation, our approach involves optimizing $Q\left(\psi; \psi^{(t)}\right)$ with respect to π_k, while ensuring that the constraint $\sum\limits_{k=1}^{K} \pi_k = 1$ is satisfied. To achieve this, we will employ the method of Lagrange multipliers Solomon (2015). Let us introduce λ as the Lagrange multiplier, and proceed with finding a solution.

$$\frac{\partial}{\partial \lambda}\left[Q\left(\psi; \psi^{(t)}\right) - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] = 0 \tag{2.47}$$

$$\frac{\partial}{\partial \pi_k}\left[Q\left(\psi; \psi^{(t)}\right) - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] = 0 \tag{2.48}$$

2.47 give us the constraint relation instead 2.48 become

$$\sum_{i=1}^{n} \frac{\tau_{ik}^{(t)}}{\pi_k} - \lambda = 0 \Leftrightarrow \sum_{i=1}^{n} \tau_{ik}^{(t)} - \lambda \pi_k = 0 \Leftrightarrow \sum_{i=1}^{n} \tau_{ik}^{(t)} = \lambda \pi_k \tag{2.49}$$

Summing equation 2.49 over all k values, we have

$$\sum_{k=1}^{K} \sum_{i=1}^{n} \tau_{ik}^{(t)} = \lambda \sum_{k=1}^{K} \pi_k \tag{2.50}$$

$$\lambda = n \tag{2.51}$$

and by isolating π_k in equation 2.49

$$\pi_k = \frac{\sum_{k=i}^{n} \tau_{ik}^{(t)}}{n} \tag{2.52}$$

Thus, at iteration $(t+1)$

$$\pi_k^{(t+1)} = \frac{\sum_{k=i}^{n} \tau_{ik}^{(t)}}{n} \tag{2.53}$$

In order to determine the value of Θ_k in the next iteration $(t+1)$, our objective is to maximize $Q\left(\psi; \psi^{(t)}\right)$ by optimizing θ_l. In the upcoming chapter 3, we will tackle this equation for every distribution.

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \frac{\partial \log\left(g_k\left(y_i; \Theta_k\right)\right)}{\partial \Theta_l} = 0 \tag{2.54}$$

2.3.3 Probability descriptor

When the likelihood of being part of a group is contingent on a covariate X, we represent this likelihood through the use of a generalized logit function.

$$\pi_{ik} = \pi_{i,\theta_k} = \frac{e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}}. \tag{2.55}$$

In order to establish distinguishability, we must assign a value of zero to θ_k for a specific group k. For the purpose of establishing a baseline, we designate group 1 and set the values of θ_1 to be equal to zero.

The M-step remains unchanged in its calculations, but the E-step requires alteration. Rather than using a Lagrange multiplier to find π_k, we determine θ_k by maximizing $\log\left(\pi_{k;\theta_k}\right)$.

In equation 2.49 or 2.52, we replace the calculation of $\pi_k^{(t+1)}$ with the calculation of $\pi_{ik}^{(t+1)} = \pi_{i,\theta_k^{(t+1)}} = \frac{e^{\theta_k^{(t+1)} x_i}}{\sum_{k=1}^{K} e^{\theta_k^{(t+1)} x_i}}$. As the calculation of a specific $\theta_{k'}$ depends on all θ_k values, we must find all the θ_k values simultaneously. Let $\theta^{(t+1)} = (\theta_1^{(t+1)}, \cdots, \theta_K^{(t+1)})$ be the set of these values.

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \log\left(\pi_{i,\theta_k^{(t)}}\right) \tag{2.56}$$

$$= \arg\max_{\theta} \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \left[\theta_k^{(t)} x_i - \log\left(\sum_{k=1}^{K} e^{\theta_k^{(t)} x_i}\right)\right] \tag{2.57}$$

For $1 \leq k \leq K$ and $1 \leq l \leq n_\theta$ by derivating by θ_{kl} this amounts to find the root of the function

$$\sum_{i=1}^{n} x_{il}\left[z_{ik}-\pi_{ik}\right] \tag{2.58}$$

Indeed, let be

$$\begin{aligned} f(\theta) &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\theta_k x_i - \log\left(\sum_{k=1}^{K} e^{\theta_k x_i}\right)\right] && (2.59)\\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\theta_k x_i - \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left(\sum_{k=1}^{K} e^{\theta_k x_i}\right) && (2.60)\end{aligned}$$

We find the derivative of the two terms. The right one is $\frac{\partial(\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\theta_k x_i)}{\partial\theta_{kl}} = z_{ik}x_{il}$ and the left one is

$$\begin{aligned} \frac{\partial\left(\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left(\sum_{k=1}^{K} e^{\theta_k x_i}\right)\right)}{\partial\theta_{kl}} &= \sum_{i=1}^{n}\sum_{k'=1}^{K} z_{ik'}\frac{x_{il}e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}} && (2.61)\\ &= \sum_{i=1}^{n}\left(\frac{x_{il}e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}}\sum_{k'=1}^{K} z_{ik'}\right) && (2.62)\\ &= \sum_{i=1}^{n}\frac{x_{il}e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}} && (2.63)\end{aligned}$$

Finally we have

$$\frac{\partial f(\theta)}{\partial\theta_{kl}} = \sum_{i=1}^{n} z_{ik}x_{il} - \sum_{i=1}^{n}\frac{x_{il}e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}} = \sum_{i=1}^{n} x_{il}\left[z_{ik} - \frac{e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}}\right] \tag{2.64}$$

We lack an explicit formula for solving equation 2.58, necessitating the utilization of quasi Newton method or IRLS method. We present a revised edition of IRLS to address the aforementioned equations.

Through the implementation of the Newton Raphson's algorithm, we obtain

$$\theta^{(t+1)} = \theta^{(t)} - \left(S'|_{\theta^{(t)}}\right)^{-1} S\left(\theta^{(t)}\right) \tag{2.65}$$

where S is the score function.

More precisely, by equation 2.58, $S(\theta)$ is a vector which elements l is $\sum_{i=1}^{n} x_{il}\left(z_{ik}-\pi_{ik}\right)$ and $\frac{\partial S(\theta)}{\partial\theta_k}$ is the vector which elements l is, for $k=k'$,

$$\frac{\partial S(\theta)}{\partial\theta_{kl'}} = \frac{\partial^2 f(\theta)}{\partial\theta_{kl'}\partial\theta_{kl}} = \sum_{i=1}^{n} -x_{il}x_{il'}\pi_{ik}(1-\pi_{ik})$$

and for $k \neq k'$

$$\frac{\partial S(\theta)}{\partial \theta_{k'l'}} = \frac{\partial^2 f(\theta)}{\partial \theta_{k'l'} \partial \theta_{kl}} = \sum_{i=1}^{n} x_{il} x_{il'} \pi_{ik} \pi_{ik'}$$

Let $Z = (Z_{11}, \cdots, Z_{n1}, \cdots, Z_{1K}, \cdots, Z_{nK})^t$, $\Pi = (\pi_{11}, \cdots, \pi_{n1}, \cdots \pi_{1K}, \cdots, \pi_{nK})^t$, Π_{kl} a matrix with diagonal's elements are $\big(\pi_{1k}(\mathbb{1}_{(k=l)} - \pi_{1l}), \cdots, \pi_{nk}(\mathbb{1}_{(k=l)} - \pi_{nl})\big)$, $\tilde{X}$ the $Kn_x \times Kn_x$ matrix with block diagonal element X and Π_W a block matrix $\Pi_W = (\Pi_{kl})_{kl}$ for $1 \leq k, l \leq K$.

$$\tilde{X} = \begin{pmatrix} X & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & X \end{pmatrix}, \quad \Pi_W = \begin{pmatrix} \Pi_{11} & \Pi_{12} & \cdots & \Pi_{1K} \\ \Pi_{12} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Pi_{(K-1)K} \\ \Pi_{1K} & \cdots & \Pi_{(K-1)K} & \Pi_{KK} \end{pmatrix}$$

where, for $k \neq l$

$$\Pi_{kk} = \begin{pmatrix} \pi_{1k}(1-\pi_{1k}) & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \pi_{nk}(1-\pi_{nk}) \end{pmatrix} \text{and } \Pi_{kl} = \begin{pmatrix} -\pi_{1k}\pi_{1l} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & -\pi_{nk}\pi_{nl} \end{pmatrix}$$

Thus we can write

$$S(\theta) = \tilde{X}^t (Z - \Pi) \tag{2.66}$$

$$\frac{\partial S(\theta)}{\partial \theta} = -\tilde{X}^t \Pi_W \tilde{X} \tag{2.67}$$

We replace this quantities in the equation (2.65) to obtain

$$\theta^{(t+1)} = \theta^{(t)} + \left(\tilde{X}^t \Pi_W \tilde{X}\right)^{-1} \tilde{X}^t (Z - \Pi) \tag{2.68}$$

$$= \left(\tilde{X}^t \Pi_W \tilde{X}\right)^{-1} \tilde{X}^t \Pi_W \tilde{X} \theta_k^{(t)} + \left(\tilde{X}^t \Pi_W X\right)^{-1} \tilde{X}^t (Z - \Pi) \tag{2.69}$$

$$= \left(\tilde{X}^t \Pi_W \tilde{X}\right)^{-1} \tilde{X}^t \left(\Pi_W \tilde{X} \theta_k^{(t)} + (Z - \Pi)\right) \tag{2.70}$$

2.3.4 Estimation of the standard errors

Estimating the standard errors of the parameters involves inverting the Fisher information matrix obtained from the observed data. However, since the incomplete data log-likelihood function is not directly utilized in the EM algorithm, we are unable to directly obtain it for the observed data. To address this, we follow Louis' procedure (Louis (1982)) to determine the observed information matrix using the complete log likelihood in the EM algorithm.

Let us define the score function for the sample data as $S(\psi; y) = \frac{\partial l(\psi;y)}{\partial \psi}$. This function represents the gradient vectors of the log-likelihood for the observed data. Similarly, we denote the score function for the complete data as $S_c(\psi; u) = \frac{\partial l(\psi;u)}{\partial \psi}$. This function represents the gradient vectors of the log-likelihood for the complete data. The negatives of the associated second derivative matrices are denoted as $I(\hat{\psi}; y)$ and $I_c\left(\hat{\psi}; u\right)$. Louis (1982) has provided proof for the validity of the following statements.

$$S(\psi; y) = E\left(S_c(\psi; u) | U = u\right) \tag{2.71}$$

$$I(\hat{\psi}; y) = E\left(I_c\left(\hat{\psi}; u\right) | U = u\right) - E\left(S_c(\psi; u) S_c^t(\psi; u) | U = u\right)_{|\psi=\hat{\psi}} \tag{2.72}$$

The information matrix for the complete data is represented by the first term in equation 2.72, while the second term provides the information for the conditional distribution of U given u. Essentially, the first part of the equation represents complete information, while the second part accounts for the information that is missing due to incomplete data. To make things simpler, we can simplify this expression by

$$I(\hat{\psi}; y) = I_c\left(\hat{\psi}; u\right) - I_{y/u}\left(\hat{\psi}; y\right) \tag{2.73}$$

We can write this equation as

$$I(\hat{\psi}; y) = E\left(-\frac{\partial^2}{\partial \psi^2} l(U; \psi) | U = u\right) - cov\left(\frac{\partial}{\partial \psi} l(U; \psi) | U = u\right)_{|\psi=\hat{\psi}} \tag{2.74}$$

In the upcoming chapter 3, we will perform explicit computations for both parts of each distribution.

CHAPTER 3

Underlying distributions

3.1 Censored normal distribution

Assuming that the variable Y_{it} is a censored variable, meaning it is a part of another variable that can be observed, we can infer that its values fall between two numbers, y_{min} and y_{max}.

To analyze this, we introduce a normally distributed variable, Y^*_{it}, which can be represented as:

$$y^*_{it} = f(a_{it}; \beta_k, \delta_k, \sigma_k) + \epsilon_{it} = \beta_k A_{it} + \delta_k W_{it} + \epsilon_{itk} \tag{3.1}$$

Here, $\epsilon_{itk} \sim \mathcal{N}(0;\ \sigma_k)$, $A_{it} = (1, a_{it}, a^2_{it}, \cdots, a^{n_\beta - 1}_{it})^t$, $W_{it} = (w^1_{it}, \cdots, w^{n_\delta}_{it})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ and $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$.

Referring to the matrix W on page 6, W_{it} represents a vector with elements corresponding to the value in row i and column t for each covariate matrix, $W^1, \cdots, W^{n_\delta}$.

Additionally, we can establish a connection between y^*_{it} and the observed and censored data, y_{it}.

$$y_{it} = \begin{cases} y_{min} \text{ if } y^*_{it} < y_{min} \\ y^*_{it} \text{ if } y_{min} \leq y^*_{it} \leq y_{max} \\ y_{max} \text{ if } y^*_{it} > y_{max} \end{cases} \tag{3.2}$$

Assuming a normal distribution with a mean of zero and a standard deviation of σ_k, the variable ϵ_{itk} can be represented as $\epsilon_{itk} \sim \mathcal{N}(0, \sigma_k)$. Consequently, the variable y^*_{it} is also normally distributed, with a mean of $\beta_k A_{it} + \delta_k W_{it}$, and is conditional on A and W with a standard deviation of σ_k. Let ϕ represent the density function of a standard normal distribution with a mean of zero and a standard deviation of one, and Φ represent its cumulative distribution function. Therefore, for a specific group k, we can express the equation as follows:

$$P\left(Y_{it} = y_{min}\right) = P\left(Y^*_{it} < y_{min}\right) = \Phi\left(\frac{y_{min} - \beta_k A_{it} + \delta_k W_{it}}{\sigma_k}\right) \tag{3.3}$$

$$P\left(Y_{it} = y_{max}\right) = P\left(Y^*_{it} > y_{max}\right) = 1 - \Phi\left(\frac{y_{max} - \beta_k A_{it} + \delta_k W_{it}}{\sigma_k}\right) \tag{3.4}$$

$$P\left(Y_{it} = y^*_{it}\right) = \frac{1}{\sigma_k}\phi\left(\frac{y_{it} - \beta_k A_{it} + \delta_k W_{it}}{\sigma_k}\right) \tag{3.5}$$

In conclusion, if we set $\mu_{ikt} = \beta_k A_{it} + \delta_k W_{it}$, we can express it as follows:

$$P(Y_{it} = y_{it}|W_i = w_i, C_i = k) = \begin{cases} \Phi\left(\frac{y_{min} - \mu_{ikt}}{\sigma_k}\right) \text{ if } y^*_{it} < y_{min} \\ \frac{1}{\sigma_k}\phi\left(\frac{y_{it} - \mu_{ikt}}{\sigma_k}\right) \text{ if } y_{min} \leq y^*_{it} \leq y_{max} \\ 1 - \Phi\left(\frac{y_{max} - \mu_{ikt}}{\sigma_k}\right) \text{ if } y^*_{it} > y_{max} \end{cases} \tag{3.6}$$

Thus the log-likelihood 2.10 becomes

$$l(\psi; y) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \sigma_k)\right) \tag{3.7}$$

where

$$g_k(y_i; \beta_k, \delta_k, \sigma_k) = \prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min} - \mu_{ikt}}{\sigma_k}\right) \prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it} - \mu_{ikt}}{\sigma_k}\right) \prod_{y_{it}=y_{max}} \left(1 - \Phi\left(\frac{y_{max} - \mu_{ikt}}{\sigma_k}\right)\right) \tag{3.8}$$

3.1.1 Likelihood

In order to fit the parameters, we employ quasi-Newton methods and must address the equations 2.16 and 2.17 in this specific scenario. These equations can be expressed as follows:

$$\frac{\partial l(\psi; y)}{\partial \theta_{kl}} = \sum_{i=1}^{n} \frac{\frac{\partial \pi_k}{\partial \theta_{kl}} g_k(y_i; \beta_k, \delta_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \sigma_k)} \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_\theta \tag{3.9}$$

$$\frac{\partial l(\psi; y)}{\partial \beta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \beta_{kl}} g_k(y_i; \beta_k, \delta_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \sigma_k)} \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_{\beta_k} \tag{3.10}$$

$$\frac{\partial l(\psi; y)}{\partial \delta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \delta_{kl}} g_k(y_i; \beta_k, \delta_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \sigma_k)} \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_\delta \tag{3.11}$$

$$\frac{\partial l(\psi; y)}{\partial \sigma_k} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \delta_{kl}} g_k(y_i; \beta_k, \delta_k, \sigma_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \sigma_k)} \quad 1 \leq k \leq K \tag{3.12}$$

When we employ likelihood to match the model, the probability membership takes the shape of $\pi_k = \dfrac{e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}}$. To compute the above equation, we will break it down into multiple steps, utilizing the properties that $(\Phi(u))' = u'\phi(u)$ and $(\phi(u))' = -u'u\phi(u)$ for a specific function u.

3.1.1.1 Differential by θ_k

Same as section 2.2.

3.1.1.2 Differential by β_{kl}

For $1 \leq l \leq n_{\beta_k}$, our initial step involves computing the differential of 3 blocks:

$$
\begin{aligned}
d_{min}^{\beta_{kl}} &= \frac{\partial}{\partial \beta_{kl}} \left(\prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right)\right) \\
d_{max}^{\beta_{kl}} &= \frac{\partial}{\partial \beta_{kl}} \left(\prod_{y_{it}=y_{max}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right)\right) \\
d^{\beta_{kl}} &= \frac{\partial}{\partial \beta_{kl}} \left(\prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right)\right)
\end{aligned}
$$

we have

$$
d_{min}^{\beta_{kl}} = \sum_{y_{it}=y_{min}} \frac{-a_{it}^{l-1}}{\sigma_k}\phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right) \prod_{\substack{y_{it'}=y_{min},\\ y_{it'}\neq y_{it}}} \Phi\left(\frac{y_{min}-\mu_{ikt'}}{\sigma_k}\right) \tag{3.13}
$$

$$
d_{max}^{\beta_{kl}} = \sum_{y_{it}=y_{max}} \frac{a_{it}^{l-1}}{\sigma_k}\phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right) \prod_{\substack{y_{it'}=y_{max},\\ y_{it'}\neq y_{it}}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt'}}{\sigma_k}\right)\right) \tag{3.14}
$$

$$
d^{\beta_{kl}} = \sum_{y_{min}<y_{it}<y_{max}} \frac{a_{it}^{l-1}}{\sigma_k^3}\left(y_{it}-\mu_{ikt}\right)\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right) \prod_{\substack{y_{min}<y_{it'}<y_{max},\\ y_{it'}\neq y_{it}}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it'}-\mu_{ikt'}}{\sigma_k}\right) \tag{3.15}
$$

Next, we proceed to compute the derivative of g_k.

$$
\begin{aligned}
\frac{\partial}{\partial \beta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k) =& d_{min}^{\beta_{kl}} \prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right) \prod_{y_{it}=y_{max}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right) \\
&+ d_{max}^{\beta_{kl}} \prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right) \prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right) \\
&+ d^{\beta_{kl}} \prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right) \prod_{y_{it}=y_{max}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right)
\end{aligned}
$$

3.1.1.3 Differential by δ_{kl}

For $1 \leq l \leq n_\delta$, our initial step involves computing the differential of 3 blocks:

$$\begin{aligned}
d_{min}^{\delta_{kl}} &= \frac{\partial}{\partial \delta_{kl}} \left(\prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right)\right) \\
d_{max}^{\delta_{kl}} &= \frac{\partial}{\partial \delta_{kl}} \left(\prod_{y_{it}=y_{max}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right)\right) \\
d^{\delta_{kl}} &= \frac{\partial}{\partial \delta_{kl}} \left(\prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right)\right)
\end{aligned}$$

we have

$$d_{min}^{\delta_{kl}} = \sum_{y_{it}=y_{min}} \frac{-w_{it}^l}{\sigma_k}\phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right) \prod_{\substack{y_{it'}=y_{min},\\ y_{it'}\neq y_{it}}} \Phi\left(\frac{y_{min}-\mu_{ikt'}}{\sigma_k}\right) \tag{3.16}$$

$$d_{max}^{\delta_{kl}} = \sum_{y_{it}=y_{max}} \frac{w_{it}^l}{\sigma_k}\phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right) \prod_{\substack{y_{it'}=y_{max},\\ y_{it'}\neq y_{it}}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt'}}{\sigma_k}\right)\right) \tag{3.17}$$

$$d^{\delta_{kl}} = \sum_{y_{min}<y_{it}<y_{max}} \frac{w_{it}^l}{\sigma_k^3}\left(y_{it}-\mu_{ikt}\right)\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right) \prod_{\substack{y_{min}<y_{it'}<y_{max},\\ y_{it'}\neq y_{it}}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it'}-\mu_{ikt'}}{\sigma_k}\right) \tag{3.18}$$

Next, we proceed to compute the derivative of g_k.

$$\begin{aligned}
\frac{\partial}{\partial \delta_{kl}} g_k(y_i;\delta_k,\delta_k) = & d_{min}^{\delta_{kl}} \prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right) \prod_{y_{it}=y_{max}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right) \\
& + d_{max}^{\delta_{kl}} \prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right) \prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right) \\
& + d^{\delta_{kl}} \prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right) \prod_{y_{it}=y_{max}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right)
\end{aligned}$$

3.1.1.4 Differential by σ_k

Similar to the aforementioned, we compute the variation of three blocks:

$$\begin{aligned}
d_{min}^{\sigma_k} &= \frac{\partial}{\partial \sigma_k} \left(\prod_{y_{it}=y_{min}} \Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right)\right) \\
d_{max}^{\sigma_k} &= \frac{\partial}{\partial \sigma_k} \left(\prod_{y_{it}=y_{max}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right)\right) \\
d^{\sigma_k} &= \frac{\partial}{\partial \sigma_k} \left(\prod_{y_{min}<y_{it}<y_{max}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right)\right)
\end{aligned}$$

and we have

$$d^{\sigma_k}_{min} = \sum_{y_{it}=y_{min}} -\frac{y_{min}-\mu_{ikt}}{\sigma_k^2}\phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right)\prod_{\substack{y_{it'}=y_{min},\\ y_{it'}\neq y_{it}}} \Phi\left(\frac{y_{min}-\mu_{ikt'}}{\sigma_k}\right) \tag{3.19}$$

$$d^{\sigma_k}_{max} = \sum_{y_{it}=y_{max}} \frac{y_{max}-\mu_{ikt}}{\sigma_k^2}\phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\prod_{\substack{y_{it'}=y_{max},\\ y_{it'}\neq y_{it}}} \left(1-\Phi\left(\frac{y_{max}-\mu_{ikt'}}{\sigma_k}\right)\right) \tag{3.20}$$

$$d^{\sigma_k} = \sum_{y_{min}<y_{it}<y_{max}} \frac{(y_{it}-\mu_{ikt})^2-\sigma_k^2}{\sigma_k^4}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right)\prod_{\substack{y_{min}<y_{it'}<y_{max},\\ y_{it'}\neq y_{it}}} \frac{1}{\sigma_k}\phi\left(\frac{y_{it'}-\mu_{ikt'}}{\sigma_k}\right) \tag{3.21}$$

Next, we proceed to compute the derivative of g_k.

$$\begin{aligned}\frac{\partial}{\partial\sigma_k}g_k(y_i;\beta_k,\delta_k,\sigma_k) =& d^{\sigma_k}_{min}\prod_{y_{min}<y_{it}<y_{max}}\frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right)\prod_{y_{it}=y_{max}}\left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right)\\ &+d^{\sigma_k}_{max}\prod_{y_{it}=y_{min}}\Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right)\prod_{y_{min}<y_{it}<y_{max}}\frac{1}{\sigma_k}\phi\left(\frac{y_{it}-\mu_{ikt}}{\sigma_k}\right)\\ &+d^{\sigma_k}\prod_{y_{it}=y_{min}}\Phi\left(\frac{y_{min}-\mu_{ikt}}{\sigma_k}\right)\prod_{y_{it}=y_{max}}\left(1-\Phi\left(\frac{y_{max}-\mu_{ikt}}{\sigma_k}\right)\right)\end{aligned}$$

3.1.2 Numerical transformation

In order to prevent negative values of σ_k from occurring during the quasi Newton method, a solution is to define $\sigma_k = e^{\alpha_k}$ since σ_k should always be positive.
Therefore, we will rephrase the score function.
If we define

$$\begin{aligned}d^{\beta_{kl}}_{min} &= \frac{\partial}{\partial\beta_k}\left(\prod_{y_{it}=y_{min}}\Phi\left(e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right)\right)\right)\\ d^{\beta_{kl}}_{max} &= \frac{\partial}{\partial\beta_k}\left(\prod_{y_{it}=y_{max}}\left(1-\Phi\left(e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right)\right)\right)\right)\\ d^{\beta_{kl}} &= \frac{\partial}{\partial\beta_k}\left(\prod_{y_{min}<y_{it}<y_{max}}e^{-\alpha_k}\phi\left(e^{-\alpha_k}\left(y_{it}-\mu_{ikt}\right)\right)\right)\end{aligned}$$

$$\begin{aligned}d^{\beta_{kl}}_{min} &= \frac{\partial}{\partial\beta_k}\left(\prod_{y_{it}=y_{min}}\Phi\left(e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right)\right)\right)\\ d^{\beta_{kl}}_{max} &= \frac{\partial}{\partial\beta_k}\left(\prod_{y_{it}=y_{max}}\left(1-\Phi\left(e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right)\right)\right)\right)\\ d^{\beta_{kl}} &= \frac{\partial}{\partial\beta_k}\left(\prod_{y_{min}<y_{it}<y_{max}}e^{-\alpha_k}\phi\left(e^{-\alpha_k}\left(y_{it}-\mu_{ikt}\right)\right)\right)\end{aligned}$$

$$d_{min}^{\beta_{kl}} = \sum_{y_{it}=y_{min}} -a_{it}^{l-1} e^{-\alpha_k} \phi\left(e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right)\right) \prod_{\substack{y_{it'}=y_{min},\\ y_{it'}\neq y_{it}}} \Phi\left(e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right)\right) \tag{3.22}$$

$$d_{max}^{\beta_{kl}} = \sum_{y_{it}=y_{max}} a_{it}^{l-1} e^{-\alpha_k} \phi\left(e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right)\right) \prod_{\substack{y_{it'}=y_{max},\\ y_{it'}\neq y_{it}}} \left(1-\Phi\left(e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right)\right)\right) \tag{3.23}$$

$$d^{\beta_{kl}} = \sum_{y_{min}<y_{it}<y_{max}} a_{it}^{l-1} e^{-3\sigma_k}\left(y_{it}-\mu_{ikt}\right) \phi\left(e^{-\alpha_k}\left(y_{it}-\mu_{ikt}\right)\right) \prod_{\substack{y_{min}<y_{it'}<y_{max},\\ y_{it'}\neq y_{it}}} e^{-\alpha_k} \phi\left(e^{-\alpha_k}\left(y_{it}-\mu_{ikt}\right)\right) \tag{3.24}$$

and with:

$$\begin{aligned}
d_{min}^{\alpha_k} &= \frac{\partial}{\partial\alpha_k}\left(\prod_{y_{it}=y_{min}} \Phi\left(e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right)\right)\right)\\
d_{max}^{\alpha_k} &= \frac{\partial}{\partial\alpha_k}\left(\prod_{y_{it}=y_{max}} \left(1-\Phi\left(e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right)\right)\right)\right)\\
d^{\alpha_k} &= \frac{\partial}{\partial\alpha_k}\left(\prod_{y_{min}<y_{it}<y_{max}} e^{-\alpha_k}\phi\left(e^{-\alpha_k}\left(y_{it}-\mu_{ikt}\right)\right)\right)
\end{aligned}$$

$$d_{min}^{\alpha_k} = \sum_{y_{it}=y_{min}} -e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right) \phi\left(e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right)\right) \prod_{\substack{y_{it'}=y_{min},\\ y_{it'}\neq y_{it}}} \Phi\left(e^{-\alpha_k}\left(y_{min}-\mu_{ikt}\right)\right) \tag{3.25}$$

$$d_{max}^{\alpha_k} = \sum_{y_{it}=y_{max}} e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right) \phi\left(e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right)\right) \prod_{\substack{y_{it'}=y_{max},\\ y_{it'}\neq y_{it}}} \left(1-\Phi\left(e^{-\alpha_k}\left(y_{max}-\mu_{ikt}\right)\right)\right) \tag{3.26}$$

$$d^{\alpha_k} = \sum_{y_{min}<y_{it}<y_{max}} e^{-\alpha_k}\left(-1+e^{-2\alpha}(y_{it}-\mu_{ikt})^2\right) \phi\left(e^{-\alpha_k}\left(y_{it}-\mu_{ikt}\right)\right) \tag{3.27}$$

$$\times \prod_{\substack{y_{min}<y_{it'}<y_{max},\\ y_{it'}\neq y_{it}}} e^{-\alpha_k}\phi\left(e^{-\alpha_k}\left(y_{it}-\mu_{ikt}\right)\right) \tag{3.28}$$

In order to achieve a scenario where all the σ_k are constant and equal to σ, meaning that all the α_k are equal to α, the derivative of the score function with respect to α can be expressed as follows:

$$\frac{\partial l(\psi;y)}{\partial\alpha} = \sum_{i=1}^{n} \frac{\sum_{k=1}^{K}\pi_k \frac{\partial}{\partial\alpha} g_k(y_i;\mu_k,\delta_k,\alpha)}{\sum_{k=1}^{K}\pi_k g_k(y_i;\mu_k,\delta_k,\alpha)}$$

Here, the term $\frac{\partial}{\partial\alpha} g_k(y_i;\mu_k,\delta_k,\alpha)$ remains the same as previously mentioned, with the removal of the index k from α_k.

3.1.3 EM algorithm

3.1.3.1 Non censored

If we make the assumption that Y is not a censored variable, the complete likelihood 2.38 transforms.

$$l(\psi; y) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k) \tag{3.29}$$

$$- \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \sum_{t=1}^{T} \log(\sigma_k) + \log(\sqrt{2\pi}) + \frac{1}{2} \left(\frac{y_{it} - (\beta_k A_{it} + \delta_k W_{it})}{\sigma_k} \right)^2 \tag{3.30}$$

To express the same concept using matrix notation,

$$l(\psi; y) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k) \tag{3.31}$$

$$- \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(T \log(\sigma_k) + T \log(\sqrt{2\pi}) + \frac{1}{2\sigma_k^2} (Y_i - (\beta_k A_i + \delta_k W_i))^t (Y_i - (\beta_k A_i + \delta_k W_i)) \right) \tag{3.32}$$

where

$$A_i = (A_{i1}, \cdots, A_{iT}) = \begin{pmatrix} 1 & \cdots & 1 \\ a_{i1} & \cdots & a_{iT} \\ \vdots & & \vdots \\ a_{i1}^{(n_\beta - 1)} & \cdots & a_{iT}^{(n_\beta - 1)} \end{pmatrix} \tag{3.33}$$

$$W_i = (W_{i1}, \cdots, W_{iT}) = \begin{pmatrix} w_{i1}^1 & \cdots & w_{iT}^1 \\ \vdots & & \vdots \\ w_{i1}^{n_\delta} & \cdots & w_{iT}^{n_\delta} \end{pmatrix} \tag{3.34}$$

Following the EM methods developed in section 2.3 we compute the two steps E and M:

- E-step

  Calculation of $E_{\psi^{(t)}}(z_{ik} | Y_i = y_i) = \tau_{ik}^{(t)} = \dfrac{\pi_k^{(t)} g_k \left(y_i, \beta_k^{(t)}, \delta_k^{(t)}, \sigma_k^{(t)} \right)}{\sum_{k=1}^{K} \pi_k^{(t)} g_k \left(y_i, \beta_k^{(t)}, \delta_k^{(t)}, \sigma_k^{(t)} \right)}$

- M-step

  Calculate of $\psi^{(t+1)} = \underset{\psi^{(t)}}{\arg\max} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \log(\pi_k g_k(y_i, \beta_k, \delta_k, \sigma_k))$

Thus, with matrix notation and since the parameters are independent, we have

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^n \tau_{ik}^{(t)}}{n} \tag{3.35}$$

$$\beta_k^{(t+1)} = \arg\min_{\beta_k} \sum_{i=1}^n \tau_{ik}^{(t)} \frac{1}{2\sigma_k^2} \left(Y_i - (\beta_k A_i + \delta_k W_i)\right)^t \left(Y_i - (\beta_k A_i + \delta_k W_i)\right) \tag{3.36}$$

$$\delta_k^{(t+1)} = \arg\min_{\delta_k} \sum_{i=1}^n \tau_{ik}^{(t)} \frac{1}{2\sigma_k^2} \left(Y_i - (\beta_k A_i + \delta_k W_i)\right)^t \left(Y_i - (\beta_k A_i + \delta_k W_i)\right) \tag{3.37}$$

$$\sigma_k^{(t+1)} = \sqrt{\frac{\sum_{i=1}^n \tau_{ik}^{(t)} \left(Y_i - \left(\beta_k^{(t+1)} A_i + \delta_k^{(t+1)} W_i\right)\right)^t \left(Y_i - \left(\beta_k^{(t+1)} A_i + \delta_k^{(t+1)} W_i\right)\right)}{T \sum_{i=1}^n \tau_{ik}^{(t)}}} \tag{3.38}$$

To find $\beta_k^{(t+1)}$, we calculate the differential by β_k

$$\frac{\partial}{\partial \beta_k} \sum_{i=1}^n \tau_{ik}^{(t)} \frac{1}{2\sigma_k^2} \left(Y_i - (\beta_k A_i + \delta_k W_i)\right)^t \left(Y_i - (\beta_k A_i + \delta_k W_i)\right) \tag{3.39}$$

$$= \frac{1}{2\sigma_k^2} \sum_{i=1}^n \tau_{ik}^{(t)} \frac{\partial}{\partial \beta_k} \left(Y_i^t Y_i - Y_i^t \beta_k A_i - Y_i^t \delta_k W_i - A_i^t \beta_k^t Y_i\right. \tag{3.40}$$

$$\left. + A_i^t \beta_k^t \beta_k A_i + A_i^t \beta_k^t \delta_k W_i - W_i^t \delta_k^t Y_i + W_i^t \delta_k^t \beta_k A_i + W_i^t \delta_k^t \delta_k W_i\right) \tag{3.41}$$

$$= \frac{1}{2\sigma_k^2} \sum_{i=1}^n \tau_{ik}^{(t)} \left(-Y_i A_i^t - Y_i A_i^t + \beta_k \left(A_i A_i^t + A_i A_i^t\right) + \delta_k W_i A_i^t + \delta_k W_i A_i^t\right) \tag{3.42}$$

$$= \frac{1}{\sigma_k^2} \sum_{i=1}^n \tau_{ik}^{(t)} \left(-Y_i A_i^t + \beta_k A_i A_i^t + \delta_k W_i A_i^t\right) \tag{3.43}$$

Assuming that $1 \leq t \leq T$, if the values of a_{it} and $a_{i't}$ are equal for all $1 \leq i, i' \leq n$, then it follows that $A_i = A_{i'}$ for all $1 \leq i, i' \leq n$. This is often the case in practical situations, where the time measurements are consistent. Now, let $1 \leq t, t' \leq T$ and $1 \leq i \leq n$. Since the evaluation of Y occurs at different times, it can be observed that $a_{it} \neq a_{it'}$. Consequently, $A_i A_i^t$ is invertible (A VOIR). To simplify the situation, we can factorize by A_1, using it as an example. Thus, we reach the same conclusion.

$$\beta_k^{(t+1)} = \frac{\left[\sum_{i=1}^n \tau_{ik}^{(t)} \left(Y_i A_i^t - \delta_k^{(t)} W_i A_i^t\right)\right] \left(A_1 A_1^t\right)^{-1}}{\sum_{i=1}^n \tau_{ik}^{(t)}}$$

If any one of the values a_{it} is not the same as the others, we will have

$$\beta_k^{(t+1)} = \left[\sum_{i=1}^n \tau_{ik}^{(t)} \left(Y_i A_i^t - \delta_k^{(t)} W_i A_i^t\right)\right] \left(\sum_{i=1}^n \tau_{ik}^{(t)} \left(A_i A_i^t\right)\right)^{-1}$$

Similar to the previous explanation, we can establish that the root of $\delta_k^{(t+1)}$ is

$$\frac{1}{\sigma_k^2}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(-Y_iW_i^t+\delta_k^{(t)}W_iW_i^t+\beta_k^{(t)}A_iW_i^t\right)$$

We find

$$\delta_k^{(t+1)}=\left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(Y_iW_i^t-\beta_k^{(t)}A_iW_i^t\right)\right]\left(\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(W_iW_i^t\right)\right)^{-1}$$

Finally, the M-step becomes

$$\pi_k^{(t+1)}=\frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}}{n} \tag{3.44}$$

$$\beta_k^{(t+1)}=\frac{\left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(Y_iA_i^t-\delta_k^{(t)}W_iA_i^t\right)\right]\left(A_1A_1^t\right)^{-1}}{\sum_{i=1}^{n}\tau_{ik}^{(t)}} \tag{3.45}$$

$$\delta_k^{(t+1)}=\left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(Y_iW_i^t-\beta_k^{(t)}A_iW_i^t\right)\right]\left(\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(W_iW_i^t\right)\right)^{-1} \tag{3.46}$$

$$\sigma_k^{(t+1)}=\sqrt{\frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(Y_i-\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right)^t\left(Y_i-\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right)}{T\sum_{i=1}^{n}\tau_{ik}^{(t)}}} \tag{3.47}$$

If our aim is to obtain a distinct σ, we employ

$$\sigma^{(t+1)}=\sqrt{\frac{\sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(Y_i-\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right)^t\left(Y_i-\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right)}{T\sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(t)}}} \tag{3.48}$$

In the event that π_k is dependent on a covariate X, we can replace the calculation of $\pi_k^{(t+1)}$ with the calculation of $\pi_{ik}^{(t+1)}=\pi_{i,\theta_k^{(t+1)}}=\dfrac{e^{\theta_k^{(t+1)}x_i}}{\sum\limits_{k=1}^{K}e^{\theta_k^{(t+1)}x_i}}$. Equation 2.70, using the same notation as this section, provides us with the following expression:

$$\theta^{(t+1)}=\left(\tilde{X}^t\Pi_W\tilde{X}\right)^{-1}\tilde{X}^t\left(\Pi_W\tilde{X}\theta_k^{(t)}+(Z-\Pi)\right) \tag{3.49}$$

or a vector with coordinates $\theta_{kl}^{(t+1)}$ for $1\leq k\leq K$ and $1\leq l\leq n_\theta$ is the solution to the following equation:

$$\sum_{i=1}^{n}x_{il}\left[z_{ik}-\pi_{ik}\right]=0 \tag{3.50}$$

3.1.3.2 Censored

When applying the EM algorithm to censored data, there are two types of missing data to consider: the membership group of Y_i and the censored data of Y_i. The likelihood function 2.38 can be redefined as follows:

$$\begin{aligned} l(\psi; y) =& \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left(\pi_k\right) \qquad (3.51)\\ &-\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\sum_{t=1}^{T}\log\left(\sigma_k\right)+\log(\sqrt{2\pi})+\frac{1}{2}\left(\frac{y_{it}^{*}-\mu_{ikt}}{\sigma_k}\right)^{2} \qquad (3.52)\end{aligned}$$

Thus, the EM algorithm becomes

- E-step

  Calculation of $Q\left(\psi;\psi^{(t)}\right)$ with the coefficient $E_{\psi^{(t)}}(z_{ik}|Y_i^{*}=y_i^{*})=\tau_{ik}^{(t)}$

- M-step

  Calculation of

  $$\psi^{(t+1)}=\arg\max_{\psi} Q\left(\psi;\psi^{(t)}\right)=\arg\max_{\psi} E\left(l_C(\psi;y)|Y_i^{*}=y_i^{*}\right)$$

  To carry out this calculus, we are required to assess

  $$E_{\psi^{(t)}}\left(\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left(\pi_k\right)-\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\sum_{t=1}^{T}\log\left(\sigma_k\right)+\log(\sqrt{2\pi})+\frac{1}{2}\left(\frac{y_{it}^{*}-\mu_{ikt}}{\sigma_k}\right)^{2}|Y_i^{*}=y_i^{*}\right)$$

  and therefore to calculate the first and second moments of y_i^{*} , given $y_i^{*}\geq y_{max}$ or $y_i^{*}\leq y_{min}$, i.e. additional to compute $E_{\psi^{(t)}}(z_{ik}|Y_i^{*}=y_i^{*})$ like above we have to compute too $E_{\psi^{(t)}}\left(z_{ik}y_{it}^{*}|Y_{it}^{*}=y_{it}^{*}\right)$ and $E_{\psi^{(t)}}\left(z_{ik}y_{it}^{*^2}|Y_{it}^{*}=y_{it}^{*}\right)$.

Proposition 1.

$$E_{\psi^{(t)}}\left(y_{it}^{*}|y_{it}^{*}\geq y_{max}\right)=\mu_{ikt}^{(t)}+\sigma_k^{(t)}q_{max,ikt}^{(t)} \qquad (3.53)$$

$$E_{\psi^{(t)}}\left(y_{it}^{*}|y_{it}^{*}\leq y_{min}\right)=\mu_{ikt}^{(t)}-\sigma_k^{(t)}q_{min,ikt}^{(t)} \qquad (3.54)$$

$$E_{\psi^{(t)}}\left(y_{it}^{*^2}|y_{it}^{*}\geq y_{max}\right)=\sigma_k^{(t)^2}\left(1+\alpha_{max,ikt}^{(t)}q_{max,ikt}^{(t)}\right)+\mu_{ikt}^{(t)^2}+2\mu_{ikt}^{(t)}\sigma_k^{(t)}q_{max,ikt}^{(t)} \qquad (3.55)$$

$$E_{\psi^{(t)}}\left(y_{it}^{*^2}|y_{it}^{*}\leq y_{min}\right)=\sigma_k^{(t)^2}\left(1+\alpha_{min,ikt}^{(t)}q_{min,ikt}^{(t)}\right)+\mu_{ikt}^{(t)^2}-2\mu_{ikt}^{(t)}\sigma_k^{(t)}q_{min,ikt}^{(t)} \qquad (3.56)$$

where

$$\alpha_{max,ikt}^{(t)}=\frac{y_{max}-\mu_{ikt}^{(t)}}{\sigma_k^{(t)}} \qquad (3.57)$$

$$\alpha_{min,ikt}^{(t)} = \frac{y_{min} - \mu_{ikt}^{(t)}}{\sigma_k^{(t)}} \tag{3.58}$$

$$q_{max,ikt}^{(t)} = \frac{\phi\left(\alpha_{max,ikt}^{(t)}\right)}{1 - \Phi\left(\alpha_{max,ikt}^{(t)}\right)} \tag{3.59}$$

$$q_{min,ikt}^{(t)} = \frac{\phi\left(\alpha_{min,ikt}^{(t)}\right)}{\Phi\left(\alpha_{min,ikt}^{(t)}\right)} \tag{3.60}$$

Proof. Let $Y \sim \mathcal{N}\left(\mu, \sigma^2\right)$.

ϕ and Φ denotes the pdf and cdf of a Standard Normal distribution.

We assume that we have $Y \in [a; b]$, where $-\infty < a < b < \infty$. Thus, the conditional density of Y is

$$f(y|y \in [a;b]) = \frac{\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

We consider the moment generating function for $l \in \mathbb{R}$

$$m(t) = E\left(e^{tY}|Y \in [a;b]\right) = \int_{-\infty}^{+\infty} e^{ty} f(y|y \in [a;b])dy = \int_a^b e^{ty} \frac{\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} dy \tag{3.61}$$

However,

$$\begin{aligned}
\int_a^b e^{ty}\frac{1}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)dy &= \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{ty}e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}dy && (3.62)\\
&= \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{ty-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}dy && (3.63)\\
&= \frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{1}{2\sigma^2}\left(y-(\sigma^2t+\mu)\right)^2+\mu^2-(\sigma^2t+\mu)^2}dy && (3.64)\\
&= e^{-\frac{1}{2\sigma^2}(\mu^2-(\sigma^2t+\mu)^2)}\frac{1}{\sigma\sqrt{2\pi}}\int_a^b e^{-\frac{1}{2\sigma^2}\left(y-(\sigma^2t+\mu)\right)^2}dy && (3.65)\\
&= e^{\mu t+\frac{\sigma^2t^2}{2}}\int_a^b\frac{1}{\sigma}\phi\left(\frac{y-(\sigma^2t+\mu)}{\sigma}\right)dy && (3.66)\\
&= e^{\mu t+\frac{\sigma^2t^2}{2}}\left(\Phi\left(\frac{b-(\sigma^2t+\mu)}{\sigma}\right)-\Phi\left(\frac{a-(\sigma^2t+\mu)}{\sigma}\right)\right) && (3.67)
\end{aligned}$$

We replace in (3.61) and we finally have

$$m(t) = \frac{e^{\mu t+\frac{\sigma^2t^2}{2}}\left(\Phi\left(\frac{b-(\sigma^2t+\mu)}{\sigma}\right)-\Phi\left(\frac{a-(\sigma^2t+\mu)}{\sigma}\right)\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{3.68}$$

$$= \frac{e^{\mu t+\frac{\sigma^2 t^2}{2}}\left(\Phi\left(\frac{b-\mu}{\sigma}-\sigma t\right)-\Phi\left(\frac{a-\mu}{\sigma}-\sigma t\right)\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{3.69}$$

When dealing with a given function u, it is important to bear in mind that the first derivative of $\Phi(u)$ can be expressed as $(\Phi(u))' = u'\phi(u)$, while the second derivative of $\phi(u)$ can be obtained as $(\phi(u))' = -u'u\phi(u)$.

$$m'(t) = \frac{e^{\mu t+\frac{\sigma^2 t^2}{2}}}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}\left[(\mu+\sigma^2 t)\left(\Phi_b(t)-\Phi_a(t)\right)-\sigma\left(\phi_b(t)-\phi_a(t)\right)\right] \tag{3.70}$$

and

$$m''(t) = \frac{e^{\mu t+\frac{\sigma^2 t^2}{2}}}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}\left[(\mu+\sigma^2 t)^2\left(\Phi_b(t)-\Phi_a(t)\right)-\sigma(\mu+\sigma^2 t)\left(\phi_b(t)-\phi_a(t)\right)\right. \tag{3.71}$$

$$+\ \sigma^2\left(\Phi_b(t)-\Phi_a(t)\right)-\sigma(\mu-\sigma^2 t)\left(\phi_b(t)-\phi_a(t)\right) \tag{3.72}$$

$$\left.-\sigma^2\left(\left(\frac{b-\mu}{\sigma}-\sigma t\right)\phi_b(t)-\left(\frac{a-\mu}{\sigma}-\sigma t\right)\phi_a(t)\right)\right] \tag{3.73}$$

Where $\Phi_a(t) = \Phi\left(\frac{a-\mu}{\sigma}-\sigma t\right)$, $\Phi_b(t) = \Phi\left(\frac{b-\mu}{\sigma}-\sigma t\right)$, $\phi_a(t) = \phi\left(\frac{a-\mu}{\sigma}-\sigma t\right)$ and $\phi_b(t) = \phi\left(\frac{b-\mu}{\sigma}-\sigma t\right)$.

Now, we can use this function to find the first and second moment of $Y|Y \in [a;b]$.

$$E\left(Y|Y \in [a;b]\right) = m'(t)|_{t=0} = \mu - \sigma\frac{\phi_b-\phi_a}{\Phi_b-\Phi_a}$$

$$E\left(Y^2|Y \in [a;b]\right) = m''(t)|_{t=0} = \mu^2+\sigma^2-2\sigma\mu\frac{\phi_b-\phi_a}{\Phi_b-\Phi_a}-\frac{\sigma\left((b-\mu)\phi_b-(a-\mu)\phi_a\right)}{\Phi_b-\Phi_a}$$

where $\phi_a = \phi\left(\frac{a-\mu}{\sigma}\right)$, $\phi_b = \phi\left(\frac{b-\mu}{\sigma}\right)$, $\Phi_a = \Phi\left(\frac{a-\mu}{\sigma}\right)$ and $\Phi_b = \Phi\left(\frac{b-\mu}{\sigma}\right)$

Remembering that

$$\lim_{a\mapsto\pm\infty}\phi_a = 0 \tag{3.74}$$

$$\lim_{a\mapsto-\infty}\Phi_a = 0 \tag{3.75}$$

$$\lim_{a\mapsto+\infty}\Phi_a = 1 \tag{3.76}$$

Letting b tend to infinity,

$$E\left(Y|Y \geq a\right) = \mu+\sigma\frac{\phi_a}{1-\Phi_a}$$

and

$$E\left(Y^2|Y\geq a\right)=\mu^2+\sigma^2+2\sigma\mu\frac{\phi_a}{1-\Phi_a}+\frac{\sigma(a-\mu)\phi_a}{1-\Phi_a} \tag{3.77}$$

$$=\mu^2+\sigma^2\left(1+\frac{(a-\mu)}{\sigma}\frac{\phi_a}{1-\Phi_a}\right)+2\sigma\mu\frac{\phi_a}{1-\Phi_a} \tag{3.78}$$

Letting a tend to minus infinity,

$$E\left(Y|Y\leq b\right)=\mu-\sigma\frac{\phi_b}{\Phi_b}$$

and

$$E\left(Y^2|Y\leq b\right)=\mu^2+\sigma^2-2\sigma\mu\frac{\phi_b}{\Phi_b}-\frac{\sigma(b-\mu)\phi_b}{\Phi_b} \tag{3.79}$$

$$=\mu^2+\sigma^2\left(1-\frac{(b-\mu)}{\sigma}\frac{\phi_b}{\Phi_b}\right)-2\sigma\mu\frac{\phi_b}{\Phi_b} \tag{3.80}$$

□

The value of $Q\left(\psi;\psi^{(t)}\right)$ is determined by the aforementioned calculations and the fact that $E_{\psi^{(t)}}\left(z_{ik}y^*_{it}|y^*_{it}=y^*_{it}\right)=P\left(z_{ik}=1|y^*_{it}\right)E_{\psi^{(t)}}\left(y^*_{it}|y^*_{it},z_{ik}=1\right)=\tau^{(t)}_{ik}E_{\psi^{(t)}}\left(y^*_{it}|y^*_{it},z_{ik}=1\right)$ and $E_{\psi^{(t)}}\left(z_{ik}y^{*^2}_{it}|y^*_{it}=y^*_{it}\right)=P\left(z_{ik}=1|y^*_{it}\right)E_{\psi^{(t)}}\left(y^{*^2}_{it}|y^*_{it},z_{ik}=1\right)=\tau^{(t)}_{ik}E_{\psi^{(t)}}\left(y^{*^2}_{it}|y^*_{it},z_{ik}=1\right)$ is

$$Q\left(\psi;\psi^{(t)}\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau^{(t)}_{ik}\log\left(\pi_k\right) \tag{3.81}$$

$$-\sum_{i=1}^{n}\sum_{k=1}^{K}\tau^{(t)}_{ik}\left[T\left(\log\left(\sigma_k\right)+\log(\sqrt{2\pi})\right)\right. \tag{3.82}$$

$$\left.+\frac{1}{2\sigma^2_k}\sum_{t=1}^{T}E_{\psi^{(t)}}\left(y^{*^2}_{it}|Y^*_{it}=y^*_{it}\right)-2E_{\psi^{(t)}}\left(y^*_{it}|Y^*_{it}=y^*_{it}\right)\mu_{ikt}+\mu^2_{ikt}\right] \tag{3.83}$$

where

$$E_{\psi^{(t)}}\left(y^{*^2}_{it}|Y^*_{it}=y^*_{it}\right)=\begin{cases}E_{\psi^{(t)}}\left(y^{*^2}_{it}|y^*_{it}\leq y_{min}\right) & \text{if } y^*_{it}\leq y_{min}\\ y^2_{it}\quad \text{if } y_{min}<y^*_{it}<y_{max}\\ E_{\psi^{(t)}}\left(y^{*^2}_{it}|y^*_{it}\geq y_{max}\right) & \text{if } y^*_{it}\geq y_{max}\end{cases} \tag{3.84}$$

$$E_{\psi^{(t)}}\left(y_{it}|Y^*_{it}=y^*_{it}\right)=\begin{cases}E_{\psi^{(t)}}\left(y^*_{it}|y^*_{it}\leq y_{min}\right) & \text{if } y^*_{it}\leq y_{min}\\ y_{it}\quad \text{if } y_{min}<y^*_{it}<y_{max}\\ E_{\psi^{(t)}}\left(y^*_{it}|y^*_{it}\geq y_{max}\right) & \text{if } y^*_{it}\geq y_{max}\end{cases} \tag{3.85}$$

In order to determine the value of β_k, we compute the root of the derivative of Q with respect to β_k.

$$\frac{\partial Q\left(\psi;\psi^{(t)}\right)}{\partial \beta_k} = \frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}\frac{\tau_{ik}^{(t)}}{2\sigma_k^2}\sum_{t=1}^{T}\left(-2E_{\psi^{(t)}}\left(y_{it}^*|Y_{it}^*=y_{it}^*\right)\mu_{ikt}+\mu_{ikt}^2\right) \tag{3.86}$$

$$= \frac{\partial}{\partial \beta_k}\sum_{i=1}^{n}\frac{\tau_{ik}^{(t)}}{2\sigma_k^2}\left(-2\tilde{Y}_i\left(\beta_k A_i+\delta_k W_i\right)^t+\left(\beta_k A_i+\delta_k W_i\right)^t\left(\beta_k A_i+\delta_k W_i\right)\right) \tag{3.87}$$

In the event that certain values in the vector Y_i fall below y_{min} or exceed y_{max}, the vector $\tilde{Y}_i$ becomes censored. In this case, the order of the elements is altered, and $\tilde{Y}_i$ is defined as $\left(y_{it}, E_{\psi^{(t)}}\left(y_{it}^*|y_{it}^*\le y_{min}\right), E_{\psi^{(t)}}\left(y_{it}^*|y_{it}^*\ge y_{max}\right)\right)$, leaves to change the order of the elements..

$$\frac{\partial Q\left(\psi;\psi^{(t)}\right)}{\partial \beta_k} = \frac{1}{2\sigma_k^2}\sum_{i=1}^{n}\tau_{ik}^{(t)}\frac{\partial}{\partial \beta_k}\left(-2\tilde{Y}_i\left(\beta_k A_i+\delta_k W_i\right)^t+\left(\beta_k A_i+\delta_k W_i\right)^t\left(\beta_k A_i+\delta_k W_i\right)\right) \tag{3.88}$$

$$= \frac{1}{2\sigma_k^2}\sum_{i=1}^{n}\tau_{ik}^{(t)}\frac{\partial}{\partial \beta_k}\left(-2\tilde{Y}_i\left(\beta_k A_i+\delta_k W_i\right)^t+A_i^t\beta_k^t\beta_k A_i+A_i^t\beta_k^t\delta_k W_i\right. \tag{3.89}$$

$$\left.+W_i^t\delta_k\beta_k A_i+W_i^t\delta_k^t\delta_k W_i\right) \tag{3.90}$$

$$= \frac{1}{2\sigma_k^2}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(-2\tilde{Y}_iA_i^t+2\beta_kA_iA_i^t+2\delta_kW_iA_i^t\right) \tag{3.91}$$

$$= \frac{1}{\sigma_k^2}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(-\tilde{Y}_iA_i^t+\beta_kA_iA_i^t+\delta_kW_iA_i^t\right) \tag{3.92}$$

In a similar manner to the equation on page 32, if all A_i are equal, we can derive the following expression:

$$\beta_k^{(t+1)} = \frac{\left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\tilde{Y}_iA_i^t-\delta_k^{(t)}W_iA_i^t\right)\right]\left(A_1A_1^t\right)^{-1}}{\sum_{i=1}^{n}\tau_{ik}^{(t)}}$$

However, if there exists at least one a_{it} that differs from the others, the equation becomes:

$$\beta_k^{(t+1)} = \left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\tilde{Y}_iA_i^t-\delta_k^{(t)}W_iA_i^t\right)\right]\left(\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(A_iA_i^t\right)\right)^{-1}$$

We can apply the same procedure to calculate $\delta_k^{(t+1)}$.

$$\frac{\partial Q\left(\psi;\psi^{(t)}\right)}{\partial \delta_k} = \frac{1}{\sigma_k^2}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(-\tilde{Y}_iW_i^t+\delta_k^{(t)}W_iW_i^t+\beta_k^{(t)}A_iW_i^t\right) \tag{3.93}$$

and finally

$$\delta_k^{(t+1)} = \left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\tilde{Y}_iW_i^t-\beta_k^{(t)}A_iW_i^t\right)\right]\left(\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(W_iW_i^t\right)\right)^{-1}$$

To find σ_k we calculate the differential of Q by σ_k.

$$\frac{\partial Q\left(\psi;\psi^{(t)}\right)}{\partial\sigma_k} = \frac{\partial}{\partial\sigma_k}\left(-\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\left[T\left(\log\left(\sigma_k\right)+\log(\sqrt{2\pi})\right)\right.\right. \tag{3.94}$$

$$\left.\left.+\frac{1}{2\sigma_k^2}\sum_{t=1}^{T}E_{\psi^{(t)}}\left(y_{it}^{*^2}|Y_{it}^*=y_{it}^*\right)-2E_{\psi^{(t)}}\left(y_{it}^*|Y_{it}^*=y_{it}^*\right)\mu_{ikt}+\mu_{ikt}^2\right]\right) \tag{3.95}$$

$$=-\sum_{i=1}^{n}\tau_{ik}^{(t)}\left[T\sigma_k^2+\sum_{t=1}^{T}E_{\psi^{(t)}}\left(y_{it}^{*^2}|Y_{it}^*=y_{it}^*\right)-2E_{\psi^{(t)}}\left(y_{it}^*|Y_{it}^*=y_{it}^*\right)\mu_{ikt}+\mu_{ikt}^2\right] \tag{3.96}$$

$$=-\sum_{i=1}^{n}\tau_{ik}^{(t)}\left[T\sigma_k^2+\tilde{Y}_{i,2}^t\mathbb{1}_T-2\tilde{Y}_i^t\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right. \tag{3.97}$$

$$\left.+\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)^t\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right] \tag{3.98}$$

where $\mathbb{1}_T$ represent a vector of length T consisting solely of the value 1, and let $\tilde{Y}_{i,2}$ be a vector that contains the squared values of Y_i. If any of these squared values fall below the threshold y_{min} or exceed the threshold y_{max}, the vector $\tilde{Y}_{i,2}$ is considered censored. In this case, $\tilde{Y}_{i,2}$ is redefined as $\left(y_{it}^2, E_{\psi^{(t)}}\left(y_{it}^{*^2}|y_{it}^*\le y_{min}\right), E_{\psi^{(t)}}\left(y_{it}^{*^2}|y_{it}^*\ge y_{max}\right)\right)$, with a change in the order of its elements.

The M-step become

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}}{n} \tag{3.99}$$

$$\beta_k^{(t+1)} = \frac{\left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\tilde{Y}_iA_i^t-\delta_k^{(t)}W_iA_i^t\right)\right]\left(A_1A_1^t\right)^{-1}}{\sum_{i=1}^{n}\tau_{ik}^{(t)}} \tag{3.100}$$

$$\delta_k^{(t+1)} = \left[\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\tilde{Y}_iW_i^t-\beta_k^{(t)}A_iW_i^t\right)\right]\left(\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(W_iW_i^t\right)\right)^{-1} \tag{3.101}$$

$$\sigma_k^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\tilde{Y}_{i,2}^t\mathbb{1}_T-2\tilde{Y}_i^t\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)+\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)^t\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right)}{T\sum_{i=1}^{n}\tau_{ik}^{(t)}}} \tag{3.102}$$

If we want unique σ we use

$$\sigma^{(t+1)} = \sqrt{\frac{\sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\tilde{Y}_{i,2}^t\mathbb{1}_T-2\tilde{Y}_i^t\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)+\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)^t\left(\beta_k^{(t)}A_i+\delta_k^{(t)}W_i\right)\right)}{T\sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{ik}^{(t)}}} \tag{3.103}$$

3.1.4 Estimation of standard error

We implement the technique outlined in subsection 2.3.4.

We calculate the entire score function.

$$S_C(\psi; y) = \left(\frac{\partial l_C(\psi; y)}{\partial \pi}, \frac{\partial l_C(\psi; y)}{\partial \beta}, \frac{\partial l_C(\psi; y)}{\partial \delta}, \frac{\partial l_C(\psi; y)}{\partial \sigma}\right)^t$$

where $\frac{\partial l_C(\psi;y)}{\partial \pi} = \left(\frac{\partial l_C(\psi;y)}{\partial \pi_1}, \dots, \frac{\partial l_C(\psi;y)}{\partial \pi_K}\right)$, $\frac{\partial l_C(\psi;y)}{\partial \beta} = \left(\frac{\partial l_C(\psi;y)}{\partial \beta_{11}}, \dots, \frac{\partial l_C(\psi;y)}{\partial \beta_{Kn_\beta}}\right)$, $\frac{\partial l_C(\psi;y)}{\partial \delta} = \left(\frac{\partial l_C(\psi;y)}{\partial \delta_{11}}, \dots, \frac{\partial l_C(\psi;y)}{\partial \delta_{Kn_\delta}}\right)$ and $\frac{\partial l_C(\psi;y)}{\partial \sigma} = \frac{\partial l_C(\psi;y)}{\partial \sigma_1}, \cdots, \frac{\partial l_C(\psi;y)}{\partial \sigma_K}$.

It is important to recall that the sum of all the values in the sequence π is equal to 1. Using this fact, we can express π_K as $\sum_{k=1}^{K} \pi_k = 1$. Additionally, when j is not equal to K, the derivative of π_K with respect to π_j is equal to -1. Consequently, we can conclude that for all values of k from 1 to $K-1$, the following equation holds true:

$$\frac{\partial l_C(\psi; y)}{\partial \pi_k} = \sum_{i=1}^{n} \left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K}\right)$$

If we utilize equation 2.64 to forecast the likelihood of membership, with $1 \leq k \leq K$ and $1 \leq l \leq n_\theta$,

$$\frac{\partial l_C\left(\psi; y\right)}{\partial \theta_{kl}} = \sum_{i=1}^{n} x_{il}\left(Z_{ik} - \pi_{ik}\right)$$

For $1 \leq k \leq K$ and $1 \leq l \leq n_{\beta_k}$

$$\frac{\partial l_C(\psi; y)}{\partial \beta_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} \frac{z_{ik} a_{it}^{l-1}\left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)}{\sigma_k^2}$$

For $1 \leq k \leq K$ and $1 \leq l \leq n_\delta$

$$\frac{\partial l_C(\psi; y)}{\partial \delta_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} \frac{z_{ik} w_{it}^{l}\left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)}{\sigma_k^2}$$

For $1 \leq k \leq K$

$$\frac{\partial l_C(\psi; y)}{\partial \sigma_k} = -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{z_{ik}\left[\sigma_k^2 - \left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)^2\right]}{\sigma_k^3}$$

3.1.4.1 Computation of the negative second derivative matrix

The negative of the second derivative matrix of the complete likelihood is,

$$-B_C(y;\hat{\psi}) = -\begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\pi^2} & \frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\beta} & \frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\delta} & \frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\sigma} \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\pi} & \frac{\partial^2 l_C(\psi;y)}{\partial\beta^2} & \frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta} & \frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma} \\ \frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\pi} & \frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\beta} & \frac{\partial^2 l_C(\psi;y)}{\partial\delta^2} & \frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\sigma} \\ \frac{\partial^2 l_C(\psi;y)}{\partial\sigma\partial\pi} & \frac{\partial^2 l_C(\psi;y)}{\partial\sigma\partial\beta} & \frac{\partial^2 l_C(\psi;y)}{\partial\sigma\partial\delta} & \frac{\partial^2 l_C(\psi;y)}{\partial\sigma^2} \end{pmatrix}$$

The dimensions of this matrix can be determined by evaluating the expression: $\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+2K-1\right)\times\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+2K-1\right)$.

If we are using a predictor for probability, we need to make a modification: we replace all instances of the derivative π with θ. Consequently, the dimensions of the matrix will be given by $\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+(K-1)n_\theta)+K\right)\times\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+(K-1)n_\theta)+K\right)$.

3.1.4.1.1 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi^2}$ or $\frac{\partial^2 l_C(\psi;y)}{\partial\theta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\pi^2}$ has for size $(K-1)\times(K-1)$.

For $1\leq k,l\leq K-1$ is

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi^2} = \left(\frac{\partial^2 l_C(\psi;y)}{\partial\pi_k\partial\pi_l}\right)_{kl}$$

which elements are for $1\leq k,l\leq K-1$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi_k\partial\pi_l} = \begin{cases} \sum_{i=1}^{n} -\left(\frac{z_{ik}}{\pi_k^2}+\frac{z_{iK}}{\pi_K^2}\right),\ k=l \\ \sum_{i=1}^{n} -\frac{z_{iK}}{\pi_K^2},\ k\neq l \end{cases}$$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\theta^2}$ has for size $Kn_\theta\times Kn_\theta$ and is composed by

$$\frac{\partial^2 l_C(\psi;y)}{\partial\theta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\theta_1\partial\theta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\theta_K\partial\theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\theta_K\partial\theta_K} \end{pmatrix}$$

which elements, for $1\leq l,l'\leq n_\theta$, are

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\theta_k\partial\theta_{k'}}\right)_{ll'} = \frac{\partial^2 l_C(\psi;y)}{\partial\theta_{kl}\partial\theta_{k'l'}} = \begin{cases} \sum_{i=1}^{n} -x_{il}x_{il'}\left(\frac{e^{\theta_k x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}}\right)\left(1-\frac{e^{\theta_k x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}}\right),\ k=k' \\ \sum_{i=1}^{n} x_{il}x_{il'}\frac{e^{\theta_k x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}}\frac{e^{\theta_{k'} x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}},\ k\neq k' \end{cases}$$

The outcomes remain the same for both π and θ in the subsequent sections.

3.1.4.1.2 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\beta}$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\beta}=0$$

3.1.4.1.3 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\delta}$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\delta}=0$$

3.1.4.1.4 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\sigma}$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\sigma}=0$$

3.1.4.1.5 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\beta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\beta^2}$ has for size $\sum_{k=1}^K n_{\beta_k}\times\sum_{k=1}^K n_{\beta_k}$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta^2}=\begin{pmatrix}\frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\beta_K}\\ \vdots & & \vdots\\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\beta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\beta_K}\end{pmatrix}$$

which elements are, for $1\le k,k'\le K$, $1\le l\le n_{\beta_k}$ and $1\le l'\le n_{\beta_{k'}}$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'}\partial\beta_k}\right)_{l'l}=\frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'l'}\partial\beta_{kl}}=\begin{cases}\sum_{i=1}^n\sum_{t=1}^T -z_{ik}\dfrac{a_{it}^{l-1}a_{it}^{l'-1}}{\sigma_k^2},\ k=k'\\ 0,\ k\neq k'\end{cases}$$

3.1.4.1.6 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta}$ has for size $\sum_{k=1}^K n_{\beta_k}\times Kn_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$, $1 \leq l \leq n_{\beta_k}$ and $1 \leq l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'}\partial\delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'l'}\partial\delta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T \dfrac{z_{ik} w_{it}^l a_{it}^{l'-1}}{\sigma_k^2}, \ k = k' \\ 0, \ k \neq k' \end{cases}$$

3.1.4.1.7 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma}$ has for size $\sum_{k=1}^K n_{\beta_k} \times K$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\sigma_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\sigma_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\sigma_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\sigma_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l \leq n_{\beta_k}$,

$$\left(\frac{\partial l_C(\psi;y)}{\partial\beta_{k'}\partial\sigma_k}\right)_{l'k} = \frac{\partial l_C(\psi;y)}{\partial\beta_{k'l}\partial\sigma_k} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T \dfrac{-2 z_{ik} a_{it}^{l-1} \left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)}{\sigma_k^3}, \ k = k' \\ 0, \ k \neq k' \end{cases}$$

3.1.4.1.8 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2}$ has for size $Kn_\delta \times Kn_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l, l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi; y)}{\partial \delta_{k'} \partial \delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi; y)}{\partial \delta_{k'l'} \partial \delta_{kl}} = \begin{cases} \sum_{i=1}^{n} \sum_{t=1}^{T} -z_{ik} \frac{w_{it}^{l} w_{it}^{l'}}{\sigma_k^2},\ k = k' \\ 0,\ k \neq k' \end{cases}$$

3.1.4.1.9 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \delta \partial \sigma}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \delta \partial \sigma}$ has for size $Kn_\delta \times K$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi; y)}{\partial \delta \partial \sigma} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \delta_1 \partial \sigma_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \delta_1 \partial \sigma_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial \delta_K \partial \sigma_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \delta_K \partial \sigma_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l \leq n_\delta$,

$$\left(\frac{\partial l_C(\psi; y)}{\partial \delta_{k'} \partial \sigma_k}\right)_{lk} = \frac{\partial l_C(\psi; y)}{\partial \delta_{k'l} \partial \sigma_k} = \begin{cases} \sum_{i=1}^{n} \sum_{t=1}^{T} \frac{-2 z_{ik} w_{it}^{l} \left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)}{\sigma_k^3},\ k = k' \\ 0,\ k \neq k' \end{cases}$$

3.1.4.1.10 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \sigma^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \sigma^2}$ has for size $K \times K$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi; y)}{\partial \sigma^2} = \left(\frac{\partial^2 l_C(\psi; y)}{\partial \sigma_k \partial \sigma_l}\right)_{kl}$$

where

$$\frac{\partial^2 l_C(\psi; y)}{\partial \sigma_k \partial \sigma_l} = \begin{cases} -\sum_{i=1}^{n} \sum_{t=1}^{T} \frac{z_{ik} \left(-\sigma_k^2 + 3 \left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)^2\right)}{\sigma_k^4},\ k = l \\ 0,\ k \neq l \end{cases}$$

To calculate the conditional expectation of the negative of the second derivative matrix $E\left(-B_C(x; \hat{\Psi}) | X = x\right)$, simply substitute τ_{ik} for z_{ik} in the equations given, where $E_\psi(Z_{ik} | Y_i = y_i) = \dfrac{\pi_k g_k(y_i, \theta)}{\sum_{k=1}^{K} \pi_k g_k(y_i, \theta_k)} = \tau_{ik}$.

3.1.5 Computation of $cov\left(S_C(\hat{\psi};u)|U=u\right)$

The conditional matrix of the score vector is given by

$$I_{y/u}\left(\hat{\psi};y\right)=\begin{pmatrix} cov\left(S_c(\pi)\right) & cov\left(S_c(\pi),S_c(\beta)\right) & cov\left(S_c(\pi),S_c(\delta)\right) & cov\left(S_c(\pi),S_c(\sigma)\right)\\ cov\left(S_c(\beta),S_c(\pi)\right) & cov\left(S_c(\beta)\right) & cov\left(S_c(\beta),S_c(\delta)\right) & cov\left(S_c(\beta),S_c(\sigma)\right)\\ cov\left(S_c(\delta),S_c(\pi)\right) & cov\left(S_c(\delta),S_c(\beta)\right) & cov\left(S_c(\delta)\right) & cov\left(S_c(\delta),S_c(\sigma)\right)\\ cov\left(S_c(\sigma),S_c(\pi)\right) & cov\left(S_c(\sigma),S_c(\beta)\right) & cov\left(S_c(\sigma),S_c(\delta)\right) & cov\left(S_c(\sigma)\right)\end{pmatrix}$$

The size of this matrix is $(K(n_\beta+n_\delta+2)-1)\times(K(n_\beta+n_\delta+2)-1)$ or $(K(n_\beta+n_\delta+n_\theta+1))\times(K(n_\beta+n_\delta+n_\theta+1))$ in the case where we are using predictors for the probability.

In order to assess the conditional covariance of the score vector based on the data, we must consider the auxiliary information required by the EM algorithm, which is the membership group information denoted as Z_{ik} (refer to equation 2.34).

Proposition 2.

- $E_\psi(Z_{ik}|Y_i=y_i)=\dfrac{\pi_k g_k(y_i,\theta)}{\sum_{k=1}^{K}\pi_k g_k(y_i,\theta_k)}=\tau_{ik}$
- $var\left(Z_{ik}\right)=E\left(Z_{ik}^2\right)-E^2\left(Z_{ik}\right)=\tau_{ik}(1-\tau_{ik})$
- $cov\left(Z_{ik},Z_{il}\right)=E\left(Z_{ik}Z_{il}\right)-E\left(Z_{ik}\right)E\left(Z_{il}\right)=-\tau_{ik}\tau_{il}$ *for* $k\neq l$
- $cov\left(Z_{ik},Z_{jl}\right)=0$

Proof.

First: by definition.

Second: we remark that $Z_{ik}^2=Z_{ik}$.

Third: for $k\neq l$, $Z_{ik}Z_{il}=0$.

Fourth: the membership of individual i and j are independent. □

3.1.5.1 Matrix $cov\left(S_c(\pi)\right)$

The matrix as for dimension $(K-1)\times(K-1)$.

For a diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write for $1\leq k\leq K-1$

$$cov\left(S_c(\pi)\right)_{kk}=\sum_{i=1}^{n}\left(\frac{\tau_{ik}(1-\tau_{ik})}{\pi_k^2}+\frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2}-2\frac{\tau_{ik}\tau_{iK}}{\pi_k\pi_K}\right)$$

For a non-diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write for $1 \leq k, l \leq K-1$

$$cov\left(S_c(\pi)\right)_{kl} = \sum_{i=1}^{n}\left(-\frac{\tau_{ik}\tau_{il}}{\pi_k\pi_l} + \frac{\tau_{ik}\tau_{iK}}{\pi_k\pi_K} + \frac{\tau_{iK}\tau_{il}}{\pi_K\pi_l} + \frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2}\right)$$

Proof. For a diagonal element

$$\begin{aligned} cov\left(S_c(\pi)\right)_{kk} &= cov\left(\sum_{i=1}^{n}\left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K}\right), \sum_{i=1}^{n}\left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K}\right)\right) && (3.104)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K}, \frac{z_{jk}}{\pi_k} - \frac{z_{jK}}{\pi_K}\right) && (3.105)\\ &= \sum_{i=1}^{n}\left(var\left(\frac{z_{ik}}{\pi_k}\right) + var\left(\frac{z_{iK}}{\pi_K}\right) + 2cov\left(\frac{z_{ik}}{\pi_k}, \frac{z_{iK}}{\pi_K}\right)\right) && (3.106) \end{aligned}$$

Furthermore,
$var\left(\frac{z_{ik}}{\pi_k}\right) = \frac{\tau_{ik}(1-\tau_{ik})}{\pi_k^2}$ and $var\left(\frac{z_{iK}}{\pi_K}\right) = \frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2}$.
$cov\left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K}, \frac{z_{jk}}{\pi_k} - \frac{z_{jK}}{\pi_K}\right) = cov\left(\frac{z_{ik}}{\pi_k}, \frac{z_{jk}}{\pi_k}\right) + cov\left(\frac{z_{iK}}{\pi_K}, \frac{z_{jK}}{\pi_K}\right) - cov\left(\frac{z_{ik}}{\pi_k}, \frac{z_{jK}}{\pi_K}\right) - cov\left(\frac{z_{iK}}{\pi_K}, \frac{z_{jk}}{\pi_k}\right)$
It is important to note that this model operates under the assumption that there is no correlation between individuals within the same group or across different time points. As a result, all covariances in these scenarios are considered to be zero.

For a non-diagonal element

$$\begin{aligned} &cov\left(\sum_{i=1}^{n}\left(\frac{Z_{ik}}{\pi_k} - \frac{Z_{iK}}{\pi_K}\right), \sum_{i=1}^{n}\left(\frac{Z_{il}}{\pi_l} - \frac{Z_{iK}}{\pi_K}\right)\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(cov\left(\frac{Z_{ik}}{\pi_k}, \frac{Z_{jl}}{\pi_l}\right) - cov\left(\frac{Z_{ik}}{\pi_k}, \frac{Z_{jK}}{\pi_K}\right) - cov\left(\frac{Z_{iK}}{\pi_K}, \frac{Z_{jl}}{\pi_l}\right) + cov\left(\frac{Z_{iK}}{\pi_K}, \frac{Z_{jK}}{\pi_K}\right)\right) \end{aligned}$$

As mentioned earlier, we leverage the independence between individuals within the same group and across time points. As a result, in the sum, most terms vanish, particularly the terms where i is not equal to j. Thus,

$$cov\left(S_c(\pi)\right)_{kl} = \sum_{i=1}^{n}\left(cov\left(\frac{Z_{ik}}{\pi_k}, \frac{Z_{il}}{\pi_l}\right) - cov\left(\frac{Z_{ik}}{\pi_k}, \frac{Z_{iK}}{\pi_K}\right) - cov\left(\frac{Z_{iK}}{\pi_K}, \frac{Z_{il}}{\pi_l}\right) + cov\left(\frac{Z_{iK}}{\pi_K}, \frac{Z_{iK}}{\pi_K}\right)\right) \tag{3.107}$$

$$= \sum_{i=1}^{n}\left(-\frac{\tau_{ik}\tau_{il}}{\pi_k\pi_l} + \frac{\tau_{ik}\tau_{iK}}{\pi_k\pi_K} + \frac{\tau_{iK}\tau_{il}}{\pi_K\pi_l} + \frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2}\right) \tag{3.108}$$

□

3.1.5.2 Matrix $cov\left(S_c(\pi), S_c(\beta)\right)$

$cov\left(S_c(\pi), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\beta_k)\right)$ for all groups k which dimension is $(K-1)\times n_{\beta_k}$. Thus, the dimension of the first matrix is $(K-1)\times Kn_{\beta_k}$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\beta_k)\right)$ that is a matrix with elements, for $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^{n}\left(B_{ikl}\tau_{ik}\left(\frac{1-\tau_{ik}}{\pi_k}+\frac{\tau_{iK}}{\pi_K}\right)\right), k'=k \\ \sum_{i=1}^{n}\left(B_{ikl}\tau_{ik}\left(-\frac{\tau_{ik'}}{\pi_{k'}}+\frac{\tau_{iK}}{\pi_K}\right)\right), k'\neq k \end{cases}$$

and for $k=K$,

$$cov\left(S_c(\pi), S_c(\beta_K)\right)_{k'l} = \sum_{i=1}^{n}\left(B_{iKl}\tau_{iK}\left(\frac{1-\tau_{iK}}{\pi_K}+\frac{\tau_{ik'}}{\pi'_k}\right)\right)$$

where $B_{ikl} = \sum_{t=1}^{T}\dfrac{a_{it}^{l-1}\left(y_{it}-(\beta_k A_{it}+\delta_k W_{it})\right)}{\sigma_k^2}$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$.

Proof. For $1 \leq k \leq K-1$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{k'l} = cov\left(\sum_{i=1}^{n}\left(\frac{Z_{ik'}}{\pi_{k'}}-\frac{Z_{iK}}{\pi_K}\right), \sum_{i=1}^{n} Z_{ik}B_{ikl}\right) \tag{3.109}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jk}B_{jkl}\right) - cov\left(\frac{Z_{iK}}{\pi_K}, Z_{jk}B_{jkl}\right)\right) \tag{3.110}$$

$$= \begin{cases} \sum_{i=1}^{n}\left(B_{ikl}\tau_{ik}\left(\frac{1-\tau_{ik}}{\pi_k}+\frac{\tau_{iK}}{\pi_K}\right)\right), k'=k \\ \sum_{i=1}^{n}\left(B_{ikl}\tau_{ik}\left(-\frac{\tau_{ik'}}{\pi'_k}+\frac{\tau_{iK}}{\pi_K}\right)\right), k'\neq k \end{cases} \tag{3.111}$$

For $k=K$,

$$cov\left(S_c(\pi), S_c(\beta_K)\right)_{k'l} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jK}B_{jKl}\right) - cov\left(\frac{Z_{iK}}{\pi_K}, Z_{jK}B_{jKl}\right)\right) \tag{3.112}$$

$$= \sum_{i=1}^{n}\left(B_{iKl}\tau_{iK}\left(\frac{1-\tau_{iK}}{\pi_K}+\frac{\tau_{ik'}}{\pi'_k}\right)\right) \tag{3.113}$$

□

3.1.5.3 Matrix $cov\left(S_c(\pi), S_c(\delta)\right)$

$cov\left(S_c(\pi), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\delta)\right)$ for all groups k which dimension is $(K-1)\times n_\delta$. Thus, the dimension of the first matrix is $(K-1)\times Kn_\delta$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\delta_k)\right)$ that is a matrix with elements, for

$1 \leq k' \leq K-1$ and $1 \leq l \leq n_\delta$

$$cov\left(S_c(\pi), S_c(\delta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^n \left(D_{ikl}\tau_{ik}\left(\frac{1-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right)\right), k' = k \\ \sum_{i=1}^n \left(D_{ikl}\tau_{ik}\left(-\frac{\tau_{ik'}}{\pi_{k'}} + \frac{\tau_{iK}}{\pi_K}\right)\right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\delta_K)\right)_{k'l} = \sum_{i=1}^n \left(D_{iKl}\tau_{iK}\left(\frac{1-\tau_{iK}}{\pi_K} + \frac{\tau_{ik'}}{\pi'_k}\right)\right)$$

where $D_{ikl} = \sum_{t=1}^T \dfrac{w_{it}^l\left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)}{\sigma_k^2}$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\delta_k}$.

3.1.5.4 Matrix $cov\left(S_c(\pi), S_c(\sigma)\right)$

$cov\left(S_c(\pi), S_c(\sigma)\right)$ as for dimension $(K-1) \times K$.

For $1 \leq k \leq K-1$ and $1 \leq l \leq K-1$

$$\left(cov\left(S_c(\pi), S_c(\sigma)\right)\right)_{kl} = \begin{cases} \sum_{i=1}^n \left(\tau_{ik}S_{ik}\left(\frac{(1-\tau_{ik})}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right)\right), \ k = l \\ \sum_{i=1}^n \left(\frac{\tau_{il}S_{il}}{\pi_l}\left(\frac{-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right)\right), \ k \neq l \end{cases}$$

and for $l = K$

$$\left(cov\left(S_c(\pi), S_c(\sigma)\right)\right)_{kK} = \sum_{i=1}^n \left(\tau_{iK}S_{iK}\left(\frac{(1-\tau_{iK})}{\pi_K} - \frac{\tau_{ik}}{\pi_k}\right)\right)$$

where $S_{ik} = -\sum_{t=1}^T \dfrac{\left[\sigma_k^2 - \left(y_{it} - (\beta_k A_{it} + \delta_k W_{it})\right)^2\right]}{\sigma_k^3}$

Proof. If $k = l$

$$\begin{aligned} \left(cov\left(S_c(\pi), S_c(\sigma)\right)\right)_{kk} &= cov\left(\sum_{i=1}^n \left(\frac{Z_{ik}}{\pi_k} - \frac{Z_{iK}}{\pi_K}\right), \sum_{i=1}^n Z_{ik}S_{ik}\right) && (3.114) \\ &= \sum_{i=1}^n \sum_{j=1}^n \left(\frac{S_{ik}}{\pi_k} cov\left(Z_{ik}, Z_{jk}\right) - \frac{S_{ik}}{\pi_K} cov\left(Z_{iK}, Z_{jk}\right)\right) && (3.115) \\ &= \sum_{i=1}^n \left(\tau_{ik}S_{ik}\left(\frac{(1-\tau_{ik})}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right)\right) && (3.116) \end{aligned}$$

If $k \neq l$

$$\begin{aligned} \left(cov\left(S_c(\pi), S_c(\sigma)\right)\right)_{kl} &= cov\left(\sum_{i=1}^n \left(\frac{Z_{ik}}{\pi_k} - \frac{Z_{iK}}{\pi_K}\right), \sum_{i=1}^n Z_{il}S_{il}\right) && (3.117) \\ &= \sum_{i=1}^n \sum_{j=1}^n \left(\frac{S_{il}}{\pi_k} cov\left(Z_{ik}, Z_{jl}\right) - \frac{S_{il}}{\pi_K} cov\left(Z_{iK}, Z_{jl}\right)\right) && (3.118) \end{aligned}$$

$$= \sum_{i=1}^{n} \left(\frac{\tau_{il} S_{il}}{\pi_l} \left(\frac{-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K} \right) \right) \tag{3.119}$$

For $l = K$

$$(cov\left(S_c(\pi), S_c(\sigma)\right))_{kK} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\frac{S_{iK}}{\pi_k} cov\left(Z_{ik}, Z_{jK}\right) - \frac{S_{iK}}{\pi_K} cov\left(Z_{iK}, Z_{jK}\right) \right) \tag{3.120}$$

$$= \sum_{i=1}^{n} \left(\tau_{iK} S_{iK} \left(\frac{(1-\tau_{iK})}{\pi_K} - \frac{\tau_{ik}}{\pi_k} \right) \right) \tag{3.121}$$

□

3.1.5.5 Matrix $cov\left(S_c(\theta)\right)$

si on choist comme référence le gr k on enlève la mtrice cov thetak et on met thetak à 0

If we use predictors for the membership probability we have to calculate the matrix with θ parameters.

$cov\left(S_c(\theta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\theta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_\theta$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\theta$.

A diagonal matrix, for $1 \le k \le K$ and $1 \le p, q \le n_\theta$ is done by

$$(cov\left(S_c(\theta_k), S_c(\theta_k)\right))_{pq} = \sum_{i=1}^{n} x_{ip} x_{iq} \tau_{ik} (1 - \tau_{ik})$$

A non diagonal matrix, for $1 \le k, l \le K$ $1 \le p \le n_\theta$ and $1 \le q \le n_\theta$, is done by

$$(cov\left(S_c(\theta_k), S_c(\theta_l)\right))_{pq} = -\sum_{i=1}^{n} x_{ip} x_{iq} \tau_{ik} \tau_{il}$$

Proof. For a diagonal matrix, for $1 \le p, q \le n_\theta$

$$(cov\left(S_c(\theta_k), S_c(\theta_k)\right))_{pq} = cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} x_{iq}\left(Z_{ik} - \pi_{ik}\right) \right) \tag{3.122}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ip} x_{jq} cov\left(Z_{ik}, Z_{jk}\right) \tag{3.123}$$

$$= \sum_{i=1}^{n} x_{ip} x_{iq} \tau_{ik} (1 - \tau_{ik}) \tag{3.124}$$

For a non diagonal matrix, $1 \le p, q \le n_\theta$

$$(cov\left(S_c(\theta_k), S_c(\theta_l)\right))_{pq} = cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} x_{iq}\left(Z_{il} - \pi_{il}\right) \right) \tag{3.125}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ip}x_{jq}cov\left(Z_{ik}, Z_{jl}\right) \tag{3.126}$$

$$= -\sum_{i=1}^{n} x_{ip}x_{iq}\tau_{ik}\tau_{il} \tag{3.127}$$

□

3.1.5.6 Matrix $cov\left(S_c(\theta), S_c(\beta)\right)$

$cov\left(S_c(\theta), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\beta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_{\beta_l}$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_{\beta_l}$.
A diagonal matrix, for $1 \le k \le K$, $1 \le p \le n_\theta$ and $1 \le q \le n_{\beta_k}$ is done by

$$\left(cov\left(S_c(\theta_k), S_c(\beta_k)\right)\right)_{pq} = \sum_{i=1}^{n} x_{ip}B_{ikq}\tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \le k, l \le K$, $1 \le p \le n_\theta$ and $1 \le q \le n_{\beta_l}$, is done by

$$\left(cov\left(S_c(\theta_k), S_c(\beta_l)\right)\right)_{pq} = \sum_{i=1}^{n} -x_{ip}B_{ilq}\tau_{ik}\tau_{il}$$

Proof. For a diagonal matrix, $1 \le q \le n_{\beta_k}$ and $1 \le p \le n_\theta$

$$\left(cov\left(S_c(\theta_k), S_c(\beta_k)\right)\right)_{pq} = cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} Z_{ik}B_{ikq}\right) \tag{3.128}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(x_{ip}Z_{ik}, Z_{jk}B_{jkq}\right) \tag{3.129}$$

$$= \sum_{i=1}^{n} x_{ip}B_{ikq}\tau_{ik}(1-\tau_{ik}) \tag{3.130}$$

For a non diagonal matrix, $1 \le q \le n_{\beta_l}$ and $1 \le p \le n_\theta$

$$\left(cov\left(S_c(\theta_k), S_c(\beta_l)\right)\right)_{pq} = cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} Z_{il}B_{ilq}\right) \tag{3.131}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(x_{ip}Z_{ik}, Z_{jl}B_{jlq}\right) \tag{3.132}$$

$$= \sum_{i=1}^{n} -x_{ip}B_{ilq}\tau_{ik}\tau_{il} \tag{3.133}$$

□

3.1.5.7 Matrix $cov\left(S_c(\theta), S_c(\delta)\right)$

$cov\left(S_c(\theta), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\delta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_\delta$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\delta$.

A diagonal matrix, for $1 \le k \le K$, $1 \le p \le n_\theta$ and $1 \le q \le n_\delta$ is done by

$$\left(cov\left(S_c(\theta_k), S_c(\delta_k)\right)\right)_{pq} = \sum_{i=1}^{n} x_{ip} D_{ikq} \tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \le k, l \le K$, $1 \le p \le n_\theta$ and $1 \le q \le n_\delta$, is done by

$$\left(cov\left(S_c(\theta_k), S_c(\delta_l)\right)\right)_{pq} = \sum_{i=1}^{n} -x_{ip} D_{ilq} \tau_{ik} \tau_{il}$$

Proof. For a diagonal matrix, $1 \le q \le n_\delta$ and $1 \le p \le n_\theta$

$$\begin{aligned}\left(cov\left(S_c(\theta_k), S_c(\delta_k)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} Z_{ik} D_{ikq}\right) && (3.134)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(x_{ip} Z_{ik}, Z_{jk} D_{jkq}\right) && (3.135)\\ &= \sum_{i=1}^{n} x_{ip} D_{ikq} \tau_{ik}(1-\tau_{ik}) && (3.136)\end{aligned}$$

For a non diagonal matrix, $1 \le q \le n_\delta$ and $1 \le p \le n_\theta$

$$\begin{aligned}\left(cov\left(S_c(\theta_k), S_c(\delta_l)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} Z_{il} D_{ilq}\right) && (3.137)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(x_{ip} Z_{ik}, Z_{jl} D_{jlq}\right) && (3.138)\\ &= \sum_{i=1}^{n} -x_{ip} D_{ilq} \tau_{ik} \tau_{il} && (3.139)\end{aligned}$$

□

3.1.5.8 Matrix $cov\left(S_c(\theta), S_c(\sigma)\right)$

$cov\left(S_c(\theta), S_c(\sigma)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\sigma_l)\right)$ for all groups k, l which dimension is $(n_{\theta_k} \times K)$. Thus, the dimension of the first matrix is $Kn_{\theta_k} \times K$.

Given $1 \le k \le K$ we compute $cov\left(S_c(\theta_k), S_c(\sigma_l)\right)$ that is a matrix with elements $1 \le p \le N_\theta$ and

$$cov\left(S_c(\theta_k), S_c(\sigma_l)\right)_p = \begin{cases} \sum_{i=1}^{n} x_{ip} S_{ik} \tau_{ik}(1-\tau_{ik}),\ k = l \\ \sum_{i=1}^{n} -x_{ip} S_{il} \tau_{ik} \tau_{il},\ k \ne l \end{cases}$$

Proof.

$$\begin{aligned}
cov\left(S_c(\beta_k), S_c(\sigma_l)\right)_p &= cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik}-\pi_{ik}\right), \sum_{i=1}^{n} Z_{il}S_{il}\right) && (3.140)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(cov\left(x_{ip}Z_{ik}, Z_{jl}S_{il}\right)\right) && (3.141)\\
&= \begin{cases} \sum_{i=1}^{n} x_{ip}S_{il}\tau_{ik}(1-\tau_{ik}),\ k=l \\ \sum_{i=1}^{n} -x_{ip}S_{il}\tau_{ik}\tau_{il},\ k\neq l \end{cases} && (3.142)
\end{aligned}$$

□

3.1.5.9 Matrix $cov\left(S_c(\beta)\right)$

$cov\left(S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\beta_l)\right)$ for all groups k and l which dimension is $n_{\beta_k}\times n_{\beta_l}$. Thus, the dimension of the first matrix is $Kn_{\beta_k}\times Kn_{\beta_l}$.
A diagonal matrix, for $1\le k\le K$ and $1\le p,q\le n_{\beta_k}$ is done by

$$\left(cov\left(S_c(\beta_k), S_c(\beta_k)\right)\right)_{pq} = \sum_{i=1}^{n} B_{ikp}B_{ikq}\tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1\le k,l\le K$, $1\le p\le n_{\beta_k}$ and $1\le q\le n_{\beta_l}$, is done by

$$\left(cov\left(S_c(\beta_k), S_c(\beta_l)\right)\right)_{pq} = \sum_{i=1}^{n} -B_{ikp}B_{ilq}\left(\tau_{ik}\tau_{il}\right)$$

Proof. For a diagonal matrix, for $1\le p,q\le n_\beta$

$$\begin{aligned}
\left(cov\left(S_c(\beta_k), S_c(\beta_k)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} Z_{ik}B_{ikp}, \sum_{i=1}^{n} Z_{ik}B_{ikq}\right) && (3.143)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z_{ik}B_{ikp}, Z_{jk}B_{jkq}\right) && (3.144)\\
&= \sum_{i=1}^{n} B_{ikp}B_{ikq}\tau_{ik}(1-\tau_{ik}) && (3.145)
\end{aligned}$$

For a non diagonal matrix, $1\le p\le n_{\beta_k}$ and $1\le q\le n_{\beta_l}$

$$\begin{aligned}
\left(cov\left(S_c(\beta_k), S_c(\beta_l)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} Z_{ik}B_{ikp}, \sum_{i=1}^{n} Z_{il}B_{ilq}\right) && (3.146)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z_{ik}B_{ikp}, Z_{jl}B_{jlq}\right) && (3.147)
\end{aligned}$$

$$= \sum_{i=1}^{n} -B_{ikp}B_{ilq}\left(\tau_{ik}\tau_{il}\right) \tag{3.148}$$

□

3.1.5.10 Matrix $cov\left(S_c(\beta), S_c(\delta)\right)$

$cov\left(S_c(\beta), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\delta_l)\right)$ for $1 \leq k \leq n_{\beta_k}$ and $1 \leq l \leq n_\delta$ which dimension is $n_{\beta_k} \times n_\delta$. Thus, the dimension of the first matrix is $Kn_{\beta_k} \times kn_\delta$.
An element of the matrix $cov\left(S_c(\beta_k), S_c(\delta_l)\right)$ for $1 \leq k, l \leq K$ is, for $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_\delta$

$$cov\left(S_c(\beta_k), S_c(\delta_l)\right)_{pq} = \begin{cases} \sum_{i=1}^{n} B_{ikp}D_{ikq}\tau_{ik}(1-\tau_{ik}),\ k = l \\ \sum_{i=1}^{n} -B_{ikp}D_{ilq}\tau_{ik}\tau_{il},\ k \neq l \end{cases}$$

Proof.

$$cov\left(S_c(\beta_k), S_c(\delta_l)\right)_{pq} = cov\left(\sum_{i=1}^{n} Z_{ik}B_{ikp}, \sum_{i=1}^{n} Z_{il}D_{ilq}\right) \tag{3.149}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z_{ik}B_{ikp}, Z_{jl}D_{jlq}\right) \tag{3.150}$$

$$= \begin{cases} \sum_{i=1}^{n} B_{ikp}D_{ikq}\tau_{ik}(1-\tau_{ik}),\ k = l \\ \sum_{i=1}^{n} -B_{ikp}D_{ilq}\tau_{ik}\tau_{il},\ k \neq l \end{cases} \tag{3.151}$$

□

3.1.5.11 Matrix $cov\left(S_c(\beta), S_c(\sigma)\right)$

$cov\left(S_c(\beta), S_c(\sigma)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\sigma_l)\right)$ for all groups k, l which dimension is $(n_{\beta_k} \times K)$. Thus, the dimension of the first matrix is $Kn_{\beta_k} \times K$.
Given $1 \leq k, l \leq K$ we compute $cov\left(S_c(\beta_k), S_c(\sigma_l)\right)$ that is a matrix with elements $1 \leq p \leq n_{\beta_k}$

$$cov\left(S_c(\beta_k), S_c(\sigma_l)\right)_p = \begin{cases} \sum_{i=1}^{n} B_{ikp}S_{ik}\tau_{ik}(1-\tau_{ik}),\ k = l \\ \sum_{i=1}^{n} -B_{ikp}S_{il}\tau_{ik}\tau_{il},\ k \neq l \end{cases}$$

Proof.

$$cov\left(S_c(\beta_k), S_c(\sigma_l)\right)_p = cov\left(\sum_{i=1}^{n} Z_{ik}B_{ikp}, \sum_{i=1}^{n} Z_{il}S_{il}\right) \tag{3.152}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(cov\left(Z_{ik}B_{ikp}, Z_{jl}S_{il}\right)\right) \tag{3.153}$$

$$= \begin{cases} \sum_{i=1}^{n} B_{ikp} S_{ik} \tau_{ik}(1-\tau_{ik}),\ k=l \\ \sum_{i=1}^{n} -B_{ikp} S_{il} \tau_{ik}\tau_{il},\ k\neq l \end{cases} \tag{3.154}$$

□

3.1.5.12 Matrix $cov\,(S_c(\delta))$

$cov\,(S_c(\delta))$ is composed by the matrix $cov\,(S_c(\delta_k), S_c(\delta_l))$ for all groups k and l which dimension is $n_\delta \times n_\delta$. Thus, the dimension of the first matrix is $Kn_\delta \times Kn_\delta$.
A diagonal matrix, for $1 \leq p, q \leq n_\delta$ is done by

$$(cov\,(S_c(\delta_k), S_c(\delta_k)))_{pq} = \sum_{i=1}^{n} D_{ikp} D_{ikq} \tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, $1 \leq p \leq n_\delta$ and $1 \leq q \leq n_\delta$, is done by

$$(cov\,(S_c(\delta_k), S_c(\delta_l)))_{pq} = \sum_{i=1}^{n} -D_{ikp} D_{ilq}\,(\tau_{ik}\tau_{il})$$

3.1.5.13 Matrix $cov\,(S_c(\delta), S_c(\sigma))$

$cov\,(S_c(\delta), S_c(\sigma))$ is composed by the matrix $cov\,(S_c(\delta_k), S_c(\sigma_l))$ for all groups k, l which dimension is $(n_\delta \times K)$. Thus, the dimension of the first matrix is $Kn_\delta \times K$.
Given $1 \leq k, l \leq K$ we compute $cov\,(S_c(\delta_k), S_c(\sigma_l))$ that is a matrix with elements for $1 \leq p \leq n_\delta$

$$cov\,(S_c(\delta_k), S_c(\sigma_l))_p = \begin{cases} \sum_{i=1}^{n} D_{ikp} S_{ik} \tau_{ik}(1-\tau_{ik}),\ k=l \\ \sum_{i=1}^{n} -D_{ikp} S_{il} \tau_{ik}\tau_{il},\ k\neq l \end{cases}$$

3.1.5.14 Matrix $cov\,(S_c(\sigma))$

The matrix as for dimension $K \times K$.
For a diagonal element of the matrix $cov\,(S_c(\sigma))$, we can write for $1 \leq k \leq K$ and $1 \leq l \leq K$,

$$cov\,(S_c(\sigma))_{kl} = \begin{cases} \sum_{i=1}^{n} S_{ik}^2 \tau_{ik}(1-\tau_{ik}),\ k=l \\ \sum_{i=1}^{n} -S_{ik} S_{il} \tau_{ik}\tau_{il},\ k\neq l \end{cases}$$

Proof. For a diagonal element we have

$$cov\,(S_c(\sigma))_{kk} = cov\left(\sum_{i=1}^{n} Z_{ik} S_k, \sum_{i=1}^{n} Z_{ik} S_k\right) \tag{3.155}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\,(Z_{ik} S_k, Z_{jk} S_k) \tag{3.156}$$

$$= \sum_{i=1}^{n} S_k^2 \tau_{ik}(1-\tau_{ik}) \tag{3.157}$$

For a non diagonal element we have

$$\begin{aligned} cov\left(S_c(\sigma)\right)_{kl} &= cov\left(\sum_{i=1}^{n} Z_{ik}S_k, \sum_{i=1}^{n} Z_{il}S_l\right) && (3.158)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z_{ik}S_k, Z_{jl}S_l\right) && (3.159)\\ &= \sum_{i=1}^{n} -S_kS_l\tau_{ik}\tau_{il} && (3.160) \end{aligned}$$

□

3.1.5.15 Censored data and estimation of standard error

In the case of censored data, if Y_i is censored, we can apply the same technique discussed in section 3.1.3.2. By leveraging the relationship $E_{\psi^{(t)}}\left(z_{ik}y_{it}^*|y_{it}^* = y_{it}^*\right) = \tau_{ik}^{(t)} E_{\psi^{(t)}}\left(y_{it}^*|y_{it}^*, z_{ik}=1\right)$, we can replace the censored vector Y_i with a modified vector $\tilde{Y}_i$. This new vector, denoted as $\tilde{Y}_i$, includes Y_i as well as additional elements to account for values below y_{min} and above y_{max}. Specifically, $\tilde{Y}_i$ is defined as $\left(Y_i, E_{\psi^{(t)}}\left(y_{it}^*|y_{it}^* \leq y_{min}\right), E_{\psi^{(t)}}\left(y_{it}^*|y_{it}^* \geq y_{max}\right)\right)$. It is important to note that the order of the elements may be altered in this modified vector without consequence.

3.1.6 Numerical application

We will now evaluate the various equations by comparing them to the output of the SAS procedure traj. In each example, we follow the same procedure: we create a sample with a structure of K clusters consisting of 500 values of variable Y, which are distributed normally. The trajectories of these values adhere to a group-controlled pattern.

Consequently, we can compare the theoretical parameter values with those obtained from the SAS procedure traj, as well as the EM algorithm mentioned earlier and the Likelihood method mentioned earlier.

3.1.6.1 Two groups, not censored, same sigma

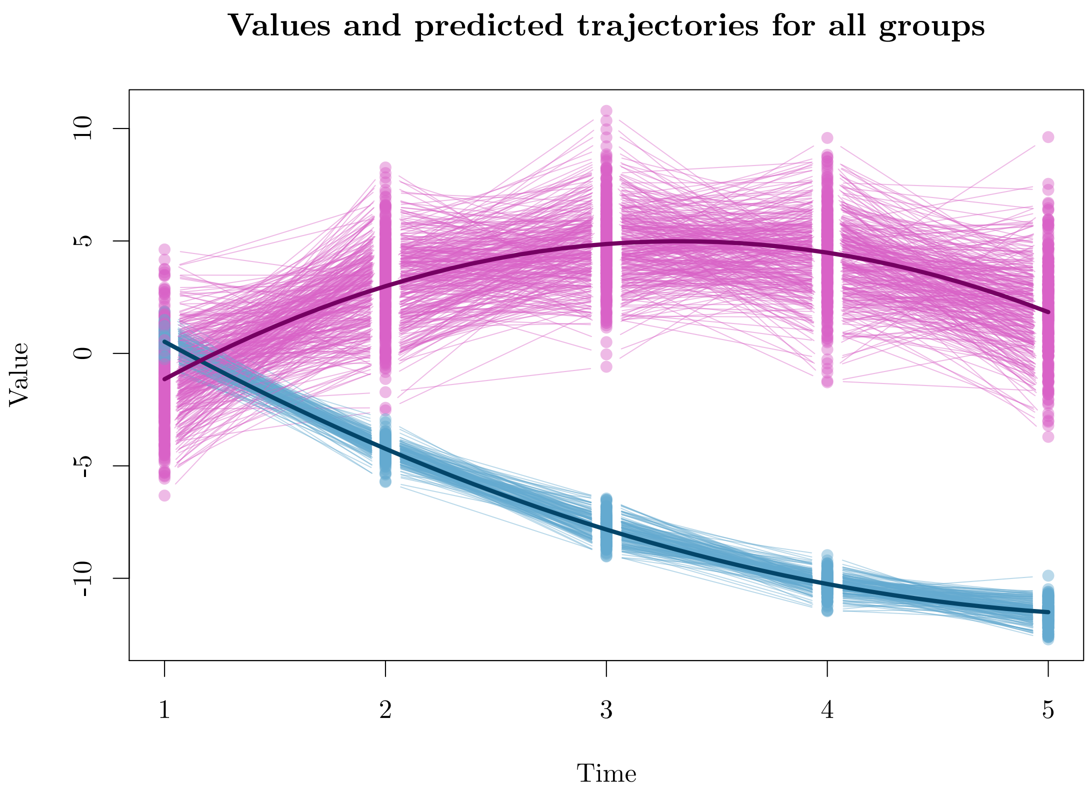
We set theoretical values as

	Polynomial			
Cluster	Degree	Shape	σ_k	Probability π_k
1	2	$\beta_1 = (6.32, -6.4, 0.567)$	2	0.32
2	2	$\beta_2 = (-7.31, 7.41, -1.12)$	2	0.68

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms. The default evaluation of Traj yields the following values: $\beta_1 = (-4.23817, 0, 0)$, $\beta_2 = (3.22114, 0, 0)$, $\sigma = 5.59449$, $\pi_1 = 0.5$. Consequently, we obtain the following results:

	Theoritical	Likelihood Traj		Likelihood trajeR		EM	
		param.	SE	param.	SE	param.	SE
β_{11}	6.32	6.45088	0.26880	6.45089	0.25153	6.45089	0.26842
β_{21}	-6.40	-6.51492	0.20484	-6.51492	0.19239	-6.51492	0.20455
β_{12}	0.57	0.58473	0.03349	0.58473	0.03158	0.58473	0.03345
β_{22}	-7.31	-7.53980	0.19121	-7.5398	0.19086	-7.5398	0.19094
β_{13}	7.41	7.52380	0.14571	7.52381	0.14563	7.52381	0.14551
β_{23}	-1.12	-1.12967	0.02383	-1.12967	0.02379	-1.12967	0.02379
σ	2	1.62214	0.02297	1.62214	0.03239	1.62214	0.02166
π_1	0.32	0.33600	0.02115	0.336	0.02128	0.336	0.02112
π_2	0.68	0.66400	0.02115	0.664	0.02128	0.664	0.02112

We draw the graph of the values and the shape of the trajectory for all the groups.



3.1.6.2 Two groups, censored, same sigma

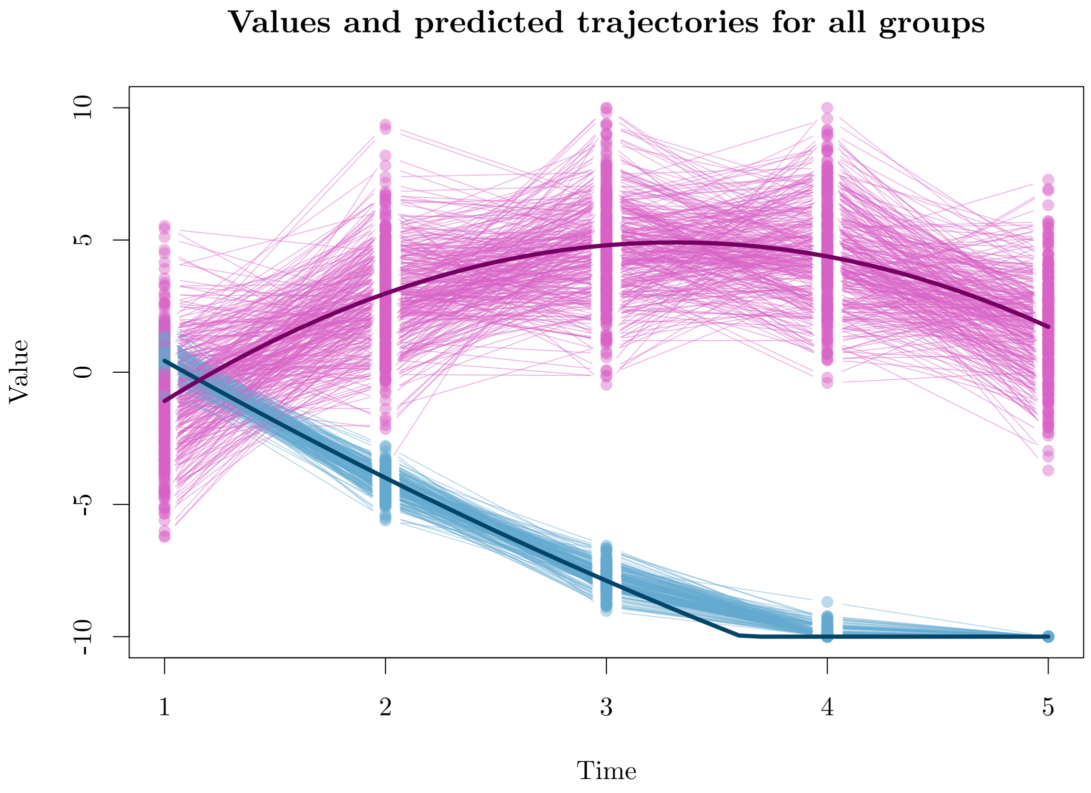
We set theoretical values as

	Polynomial			
Cluster	Degree	Shape	σ_k	Probability π_k
1	2	$\beta_1 = (6.32, -6.4, 0.567)$	2	0.32
2	2	$\beta_2 = (-7.31, 7.41, -1.12)$	2	0.68

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms, with censorship conditions of $y_{min} = -10$ and $y_{max} = 10$. The default evaluation of Traj yields the following values: $\beta_1 = (-4.33011, 0, 0)$, $\beta_2 = (2.91028, 0, 0)$, $\sigma = 5.43029$, $\pi_1 = 0.5$. Consequently, we obtain the following results:

	Theoritical	Likelihood Traj		Likelihood trajeR		EM	
			SE		SE		SE
β_{11}	6.32	5.43826	0.32481	5.43825	0.32332	5.43825	0.25963
β_{21}	-6.40	-5.28141	0.28597	-5.2814	0.28463	-5.2814	0.18483
β_{12}	0.57	0.28047	0.05539	0.28047	0.05524	0.28047	0.02518
β_{22}	-7.31	-7.39798	0.20786	-7.39798	0.21009	-7.39798	0.20756
β_{13}	7.41	7.43075	0.15840	7.43075	0.15998	7.43075	0.15818
β_{23}	-1.12	-1.12154	0.02590	-1.12154	0.02611	-1.12154	0.02586
σ	2	1.71761	0.02599	1.71761	0.03759	1.71761	0.02468
π_1	0.32	0.36999	0.21622	0.37	0.02182	0.37	0.02159
π_2	0.68	0.63001	0.21622	0.63	0.02182	0.63	0.02159

We draw the graph of the values and the shape of the trajectory for all the groups.



3.1.6.3 Three groups, censored, same sigma

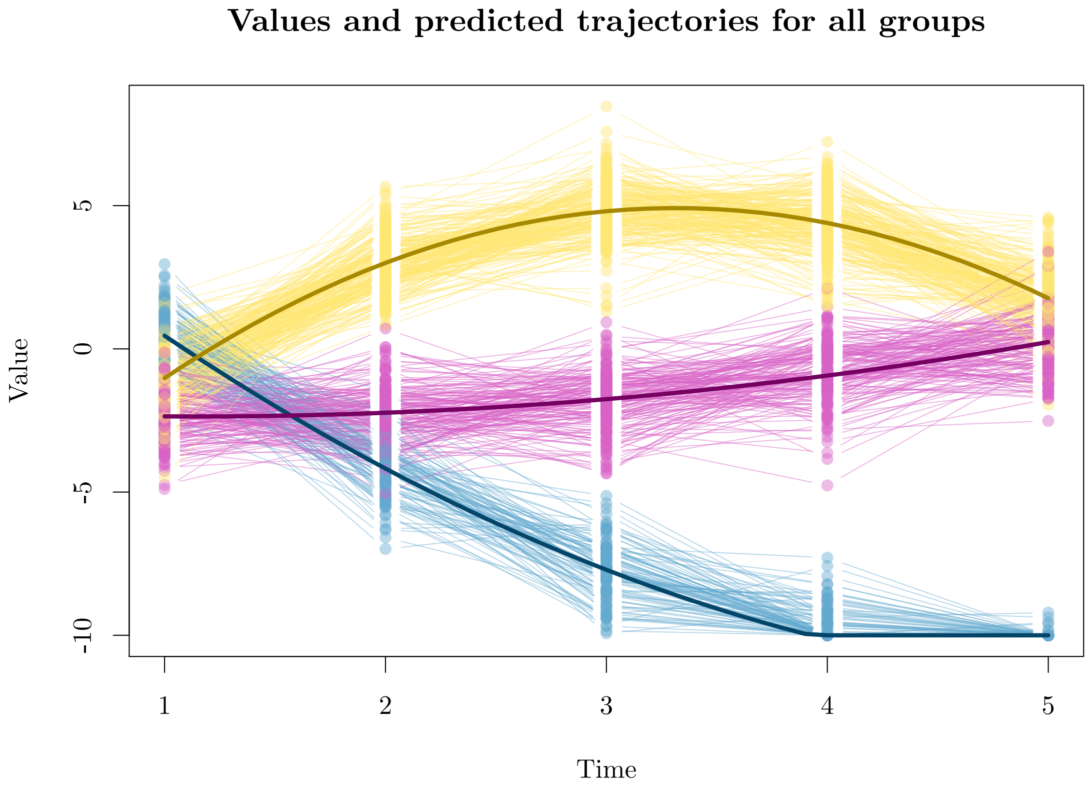
We set theoretical values as

	Polynomial			
Cluster	Degree	Shape	σ_k	Probability π_k
1	2	$\beta_1 = (6.32, -6.4, 0.567)$	1	0.32
2	2	$\beta_2 = (-7.31, 7.41, -1.12)$	1	0.54
3	2	$\beta_3 = (-2.34, -0.2, 0.14)$	1	0.14

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms, with censorship conditions of $y_{min} = -10$. The default evaluation of Traj yields the following values: $\beta_1 = (-4.65719, 0, 0)$, $\beta_2 = (-0.45513, 0, 0)$,$\beta_3 = (3.74692, 0, 0)$, $\sigma = 4.20206$, $\pi_1 = \pi_2 = \pi_3 = 0.33333$. Consequently, we obtain the following results:

	Theoritical	Likelihood Traj		Likelihood trajeR		EM	
			SE		SE		SE
β_{11}	6.32	6.20585	0.23489	6.20585	0.24855	6.20585	0.23004
β_{21}	-5.80	-6.29939	0.19108	-6.29938	0.19506	-6.29938	0.18783
β_{31}	1.00	0.55338	0.03397	0.55338	0.0343	0.55338	0.03462
β_{12}	-6.69	-2.13727	0.17021	-2.13728	0.17869	-2.13728	0.16984
β_{22}	6.92	-0.39408	0.12971	-0.39408	0.13539	-0.39408	0.12943
β_{32}	-1.23	0.17383	0.02121	0.17383	0.02228	0.17383	0.02116
β_{13}	-2.34	-7.25379	0.14086	-7.25378	0.14041	-7.25378	0.14055
β_{23}	-0.20	7.34223	0.10735	7.34223	0.10772	7.34223	0.10711
β_{33}	0.14	-1.10750	0.01755	-1.1075	0.01774	-1.1075	0.01751
σ	1	1.01100	0.01481	1.011	0.02609	1.011	0.02137
π_1	0.32	0.198	0.1786	0.198	0.03061	0.198	0.01782
π_2	0.54	0.326	0.21009	0.326	0.02029	0.326	0.02096
π_3	0.14	0.476	0.22384	0.476	0.02292	0.476	0.02751

We draw the graph of the values and the shape of the trajectory for all the groups.



3.1.6.4 Three groups, censored, same sigma but bigger

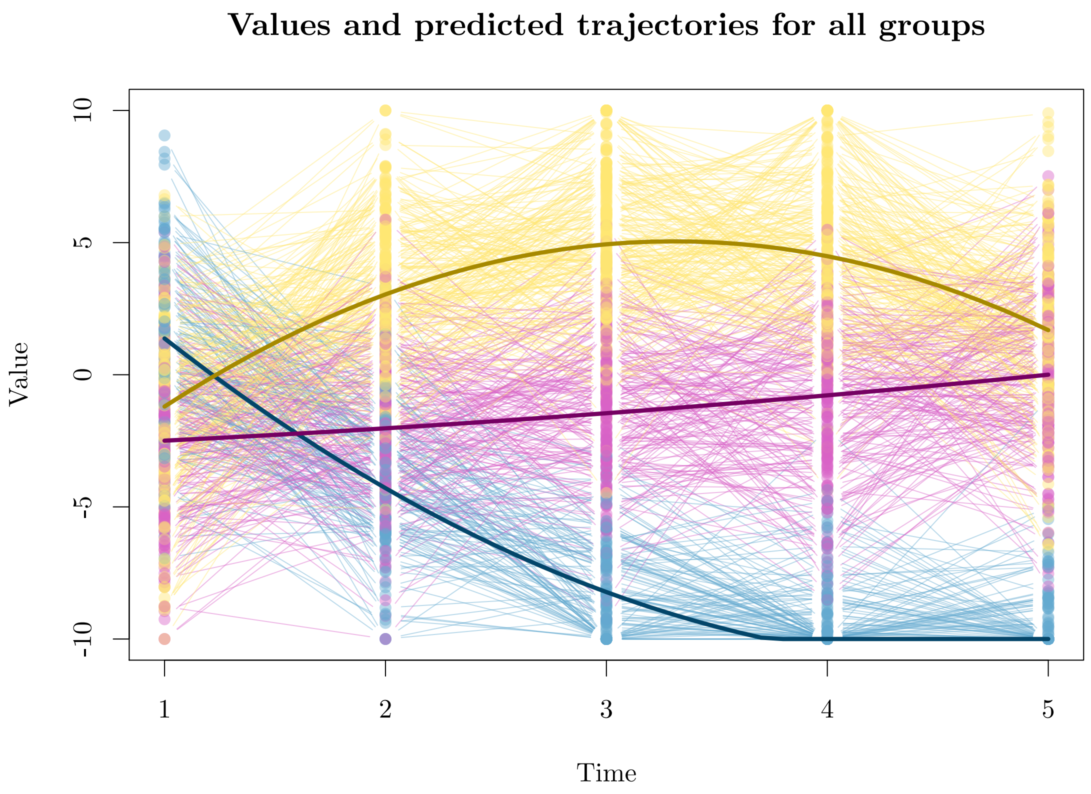
We set theoretical values as

	Polynomial			
Cluster	Degree	Shape	σ_k	Probability π_k
1	2	$\beta_1 = (6.32, -6.4, 0.567)$	3	0.32
2	2	$\beta_2 = (-7.31, 7.41, -1.12)$	3	0.54
3	2	$\beta_3 = (-2.34, -0.2, 0.14)$	3	0.14

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms, with censorship conditions of $y_{min} = -10$. The default evaluation of Traj yields the following values: $\beta_1 = (-4.65719, 0, 0)$, $\beta_2 = (-0.45513, 0, 0)$,$\beta_3 = (3.74692, 0, 0)$, $\sigma = 4.20206$, $\pi_1 = \pi_2 = \pi_3 = 1/3$. Consequently, we obtain the following results:

	Theoritical	Likelihood Traj		Likelihood trajeR		EM	
			SE		SE		SE
β_{11}	6.32	8.77009	0.62884	8.77008	0.63866	8.77008	0.61173
β_{21}	-6.40	-8.25825	0.48912	-8.25825	0.49685	-8.25825	0.46623
β_{31}	0.57	0.86305	0.08129	0.86305	0.0824	0.86305	0.07677
β_{12}	-7.31	-2.86283	0.55533	-2.86283	0.56286	-2.86283	0.55327
β_{22}	7.41	0.31732	0.43404	0.31732	0.43876	0.31732	0.432
β_{32}	-1.12	0.05114	0.07073	0.05114	0.07154	0.05114	0.0704
β_{13}	-2.34	-7.78899	0.46525	-7.789	0.48551	-7.789	0.46318
β_{23}	-0.20	7.75605	0.36216	7.75606	0.37741	7.75606	0.35949
β_{33}	0.14	-1.17223	0.05918	-1.17223	0.06166	-1.17223	0.05876
σ	3	3.04703	0.04681	3.04703	0.08364	3.04703	0.06954
π_1	0.32	0.2294	0.01889	0.2294	0.03464	0.2294	0.01868
π_2	0.54	0.32679	0.0238	0.32679	0.02399	0.32679	0.02128
π_3	0.14	0.44381	0.0249	0.44381	0.02499	0.44381	0.02832

We draw the graph of the values and the shape of the trajectory for all the groups.



3.2 Logistic distribution

3.2.1 Generality

If we are dealing with dichotomous data represented by Y, we can employ the LOGIT model to estimate it. We introduce a latent variable y_{it}^* in this case.

$$y_{it}^* = f(a_{it}; \beta_k, \delta_k) + \epsilon_{itk} = \beta_k A_{it} + \delta_k W_{it} + \epsilon_{ikt} \tag{3.161}$$

where $\epsilon_{itk} \sim \mathcal{N}(0; \sigma_k)$, $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1})^t$, $W_{it} = (w_{it}^1, \cdots, w_{it}^{n_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ and $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$.

In the traditional sense, it is commonly believed that the binary variable $y_{it} = 1$ when $y_{it}^* > 0$ and $y_{it} = 0$ when $y_{it}^* \leq 0$.

If we assume that ϵ_{ikt} follows a normal distribution, we can define the probit function. Let $\rho_{ikt} = P(Y_{it} = 1 | W_i = w_i, C_i = k)$ be the probability of $y_{it} = 1$ given membership in group k. We make the assumption:

$$\rho_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \tag{3.162}$$

Thus the log-likelihood 2.10 becomes

$$l(\psi; y) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k) \right) \tag{3.163}$$

where

$$g_k(y_i; \beta_k, \delta_k) = \prod_{t=1}^{T} \left(\frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \right)^{y_{it}} \left(\frac{1}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \right)^{1 - y_{it}} \tag{3.164}$$

3.2.2 Likelihood

To fit the parameters, we employ quasi-Newton techniques and need to resolve equations 2.16 and 2.17. This becomes necessary in this specific scenario:

$$\frac{\partial l(\psi; y)}{\partial \theta_{kl}} = \sum_{i=1}^{n} \frac{\frac{\partial \pi_k}{\partial \theta_{kl}} g_k(y_i; \beta_k, \delta_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k)} = 0 \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_\theta \tag{3.165}$$

$$\frac{\partial l(\psi; y)}{\partial \beta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \beta_{kl}} g_k(y_i; \beta_k, \delta_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k)} = 0 \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_{\beta_k} \tag{3.166}$$

$$\frac{\partial l(\psi;y)}{\partial \delta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \delta_{kl}} g_k(y_i;\beta_k,\delta_k)}{\sum\limits_{k=1}^{K} \pi_k g_k(y_i;\beta_k,\delta_k)} = 0 \quad 1 \le k \le K, \text{ and } 1 \le l \le n_\delta \tag{3.167}$$

$$\tag{3.168}$$

When employing likelihood to fit the model, the probability membership takes the shape of $\pi_k = \frac{e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}}$. The aforementioned equation will be computed through multiple stages, yet it is important to acknowledge that there are no solutions available in closed form.

3.2.2.1 Differential by θ_k

Same as section 2.2.

3.2.2.2 Differential by β_{kl}

Let $1 \le l \le n_{\beta_k}$, the derivative $\frac{\partial}{\partial \beta_{kl}} g_k(y_i;\beta_k,\delta_k)$ is

$$\sum_{t'=1}^{T} \frac{a_{it'}^{l-1} e^{y_{it'}(\beta_k A_{it'}+\delta_k W_{it'})}}{\left(1+e^{\beta_k A_{it'}+\delta_k W_{it'}}\right)^2} \left(y_{it'} - (1-y_{it'})\, e^{\beta_k A_{it'}+\delta_k W_{it'}}\right) \prod_{\substack{t=1,\\ t\neq t'}}^{T} \left(\frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{y_{it}} \left(\frac{1}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{1-y_{it}} \tag{3.169}$$

Proof. In our calculations, we employ the Leibniz formula, which provides us with the necessary formula to find the derivative of a product:

$$\frac{\mathrm{d}}{\mathrm{d}x} \prod_{t=1}^{T} f_t(x) = \sum_{t=1}^{T} \left(\frac{\mathrm{d}}{\mathrm{d}x} f_t(x) \prod_{j\neq t} f_j(x) \right)$$

However,

$$\frac{\partial}{\partial \beta_{kl}} \left(\frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{y_{it}} \left(\frac{1}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{1-y_{it}} \tag{3.170}$$

$$= y_{it} \left(\frac{a_{it}^{l-1} e^{\beta_k A_{it}+\delta_k W_{it}} \left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right) - a_{it}^{l-1} e^{\beta_k A_{it}+\delta_k W_{it}} e^{\beta_k A_{it}+\delta_k W_{it}}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2} \right) \left(\frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{y_{it}-1} \tag{3.171}$$

$$\times \left(\frac{1}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{1-y_{it}} + \tag{3.172}$$

$$\left(\frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{y_{it}} \times (1-y_{it}) \left(\frac{-a_{it}^{l-1} e^{\beta_k A_{it}+\delta_k W_{it}}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2}\right) \left(\frac{1}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{-y_{it}} \tag{3.173}$$

$$= \frac{y_{it} a_{it}^{l-1} \left(e^{\beta_k A_{it}+\delta_k W_{it}}\right)^{y_{it}}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2} - \frac{(1-y_{it}) a_{it}^{l-1} \left(e^{\beta_k A_{it}+\delta_k W_{it}}\right)^{1+y_{it}}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2} \tag{3.174}$$

$$= \frac{a_{it}^{l-1}\left(e^{\beta_k A_{it}+\delta_k W_{it}}\right)^{y_{it}}\left(y_{it}-(1-y_{it})e^{\beta_k A_{it}+\delta_k W_{it}}\right)}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2} \tag{3.175}$$

We deduce the result. □

3.2.2.3 Differential by δ_{kl}

Let $1 \leq l \leq n_\delta$, similarly the derivative $\frac{\partial}{\partial \delta_{kl}} g_k(y_i; \beta_k, \delta_k)$ is

$$\sum_{t=1}^{T} \frac{w_{it}^{l} e^{y_{it}(\beta_k A_{it}+\delta_k W_{it})}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2}\left(y_{it}-(1-y_{it})\, e^{\beta_k A_{it}+\delta_k W_{it}}\right)\prod_{\substack{t=1,\\ t\neq l}}^{T}\left(\frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{y_{it}}\left(\frac{1}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}\right)^{1-y_{it}} \tag{3.176}$$

3.2.3 EM algorithm

The complete likelihood 2.38

$$l_C(\psi; y) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left(\sum_{t=1}^{T} y_{it}\left(\beta_k A_{it}+\delta_k W_{it}\right) - \log\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)\right) \tag{3.177}$$

Following the EM methods developed in section 2.3 we compute the two steps E and M:

- E step :

  Calculation of $E_{\psi^{(t)}}(z_{ik}|Y_i = y_i) = \tau_{ik}^{(t)} = \dfrac{\pi_k^{(t)} g_k\left(y_i, \beta_k^{(t)}, \delta_k^{(t)}\right)}{\sum_{k=1}^{K}\pi_k^{(t)} g_k\left(y_i, \beta_k^{(t)}, \delta_k^{(t)}\right)}$

- M step :

  Calculate of $\psi^{(t+1)} = \arg\max_{\psi} \sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\log\left(\pi_k g_k\left(y_i, \beta_k, \delta_k\right)\right)$ which is done by

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}}{n} \tag{3.178}$$

$$\beta_k^{(t+1)} = \arg\max_{\beta_k}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\sum_{t=1}^{T} y_{it}\left(\beta_k A_{it}+\delta_k W_{it}\right) - \log\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)\right) \tag{3.179}$$

$$\delta_k^{(t+1)} = \arg\max_{\delta_k}\sum_{i=1}^{n}\tau_{ik}^{(t)}\left(\sum_{t=1}^{T} y_{it}\left(\beta_k A_{it}+\delta_k W_{it}\right) - \log\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)\right) \tag{3.180}$$

Or

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}}{n} \tag{3.181}$$

$$\beta_{kl}^{(t+1)} \text{ root of } \sum_{i=1}^{n}\sum_{t=1}^{T}\tau_{ik}^{(t)} a_{it}^{l-1}\left(y_{it} - \frac{e^{\beta_k^{(t)} A_{it}+\delta_k^{(t)} W_{it}}}{1+e^{\beta_k^{(t)} A_{it}+\delta_k^{(t)} W_{it}}}\right) \quad 1 \leq l \leq n_\beta \tag{3.182}$$

$$\delta_{kl}^{(t+1)} \text{ root of } \sum_{i=1}^{n} \sum_{t=1}^{T} \tau_{ik}^{(t)} w_{it}^{l} \left(y_{it} - \frac{e^{\beta_k^{(t)} A_{it} + \delta_k^{(t)} W_{it}}}{1 + e^{\beta_k^{(t)} A_{it} + \delta_k^{(t)} W_{it}}} \right) \quad 1 \leq l \leq n_\delta \tag{3.183}$$

Neither β_k nor δ_k have a closed form. Thus, we use quasi-newton methods such as the BFGS approach to seek for their approximation.

3.2.4 Iteratively Reweighted Least Squares

A modified version of IRLS is suggested to solve the aforementioned equations.

According to equation 2.28, which is the Newton Raphson's algorithm, we have

$$\beta_k^{(t+1)} = \beta_k^{(t)} - \left(S'|_{\beta_k^{(t)}} \right)^{-1} S(\beta_k^{(t)}) \tag{3.184}$$

where S is the score function.
More precisely, $S(\beta_k)$ is a vector which elements l is $\sum_{i=1}^{n} \sum_{t=1}^{T} \tau_{ik} a_{it}^{l-1} \left(y_{it} - \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \right)$ for $1 \leq l \leq n_\beta$. We write it as $\sum_{i=1}^{n} \sum_{t=1}^{T} \tau_{ik} a_{it}^{l-1} (y_{it} - \rho_{ikt})$ and $\frac{\partial S(\beta_k)}{\partial \beta_{kl'}}$ is the vector which elements l is

$$\left(\frac{\partial S(\beta_k)}{\partial \beta_{kl'}} \right)_l = \sum_{i=1}^{n} \sum_{t=1}^{T} -\tau_{ik} a_{it}^{l-1} a_{it}^{l'-1} \rho_{ikt} (1 - \rho_{ikt})$$

for $1 \leq l, l' \leq n_\beta$.
Let be the matrix

$$A_\beta = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ a_{11}^{1} & \cdots & a_{1T}^{1} & \cdots & a_{n1}^{1} & \cdots & a_{nT}^{1} \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{11}^{n_{\beta_k}-1} & \cdots & a_{1T}^{n_{\beta_k}-1} & \cdots & a_{n1}^{n_{\beta_k}-1} & \cdots & a_{nT}^{n_{\beta_k}-1} \end{pmatrix} \tag{3.185}$$

$$W = \begin{pmatrix} w_{11}^{1} & \cdots & w_{1T}^{1} & \cdots & w_{11}^{n_\delta} & \cdots & w_{1T}^{n_\delta} \\ \vdots & & \vdots & & \vdots & & \vdots \\ w_{n1}^{1} & \cdots & w_{nT}^{1} & \cdots & w_{n1}^{n_{\delta_k}} & \cdots & w_{nT}^{n_{\delta_k}} \end{pmatrix} \tag{3.186}$$

$$W_\rho = \begin{pmatrix} \rho_{1k1}(1-\rho_{1k1}) & 0 & & \cdots & & & 0 \\ 0 & \ddots & & & & & \\ & & \rho_{1kT}(1-\rho_{1kT}) & & & & \\ \vdots & & & \ddots & & & \vdots \\ & & & & \rho_{nk1}(1-\rho_{nk1}) & & \\ & & & & & \ddots & 0 \\ 0 & & & \cdots & & & \rho_{nkT}(1-\rho_{nkT}) \end{pmatrix} \tag{3.187}$$

Z is a diagonal matrix with diagonal elements $\left(\underbrace{\tau_{1k},\cdots,\tau_{1k}}_{T},\cdots,\underbrace{\tau_{nk},\cdots,\tau_{nk}}_{T}\right)$, Y is the matrix with all values Y_i, $Y=(Y_{11},\cdots Y_{1T},\cdots,Y_{n1},\cdots,Y_{nT})^t$ and $P=(\rho_{1k1},\cdots\rho_{1kT},\cdots,\rho_{nk1},\cdots,\rho_{nkT})^t$.

Thus we can write

$$S(\beta_k) = A_\beta Z\left(Y-P\right) \tag{3.188}$$

$$\frac{\partial S(\beta_k)}{\partial \beta_k} = -A_\beta Z W_\rho A_\beta^t \tag{3.189}$$

We replace this quantities in the equation (3.184) to obtain

$$\beta_k^{(t+1)} = \beta_k^{(t)} + \left(A_\beta Z^{(t)} W_\rho^{(t)} A_\beta^t\right)^{-1} A_\beta Z^{(t)}\left(Y-P^{(t)}\right) \tag{3.190}$$

$$= \left(A_\beta Z^{(t)} W_\rho^{(t)} A_\beta^t\right)^{-1} A_\beta Z^{(t)} W_\rho^{(t)} A_\beta^t \beta_k^{(t)} + \left(A_\beta Z^{(t)} W_\rho^{(t)} A_\beta^t\right)^{-1} A_\beta Z^{(t)}\left(Y-P^{(t)}\right) \tag{3.191}$$

$$= \left(A_\beta Z^{(t)} W_\rho^{(t)} A_\beta^t\right)^{-1} A_\beta Z^{(t)} W_\rho^{(t)}\left(A_\beta^t \beta_k^{(t)} + W_\rho^{(t)^{-1}}\left(Y-P^{(t)}\right)\right) \tag{3.192}$$

We can observe that $A_\beta^t \beta_k^{(t)} + W_\rho^{(t)^{-1}}\left(Y-P^{(t)}\right)$ is a vector whose elements are, with respect to the notations and indices provided earlier, the following:

$$\beta_k^{(t)} A_{it} + \frac{Y_{it}-\rho_{ikt}^{(t)}}{\rho_{ikt}^{(t)}\left(1-\rho_{ikt}^{(t)}\right)} \tag{3.193}$$

In the same way we write for δ_k,

$$S(\delta_k) = WZ\left(Y-P\right) \tag{3.194}$$

$$\frac{\partial S(\delta_k)}{\partial \delta_k} = -WZW_\rho W^t \tag{3.195}$$

and

$$\delta_k^{(t+1)} = \left(WZ^{(t)}W_\rho^{(t)}W\right)^{-1} WZ^{(t)}W_\rho^{(t)} \left(W^t\beta_k^{(t)} + W_\rho^{(t)^{-1}} \left(Y - P^{(t)}\right)\right) \tag{3.196}$$

Furthermore, we have the following information: $W^t\beta_k^{(t)} + W_\rho^{(t)^{-1}} \left(Y - P^{(t)}\right)$ is a vector whose elements, based on the provided notations and indices, can be expressed as:

$$\delta_k^{(t)}W_{it} + \frac{Y_{it} - \rho_{ikt}^{(t)}}{\rho_{ikt}^{(t)}(1 - \rho_{ikt}^{(t)})} \tag{3.197}$$

3.2.5 Numerical method

To avoid problems with singular matrices when attempting to find $\left(A_\beta Z^{(t)}W_\rho^{(t)}A_\beta^t\right)^{-1}$, we can employ the QR method to determine $\beta_k^{(t+1)}$. Using the results presented in Section 2.2.3, $\beta_k^{(t+1)}$ should satisfy the following condition:

$$\left(A_\beta(Z^{(t)}W_\rho^{(t)})^{1/2}\left((Z^{(t)}W_\rho^{(t)})^{1/2}\right)^t A_\beta^t\right)\beta_k^{(t+1)} \tag{3.198}$$

$$= A_\beta(Z^{(t)}W\rho^{(t)})^{1/2}\left[(Z^{(t)}W_\rho^{(t)})^{1/2}\left(A_\beta^t\beta_k^{(t)} + W\rho^{(t)^{-1}}\left(S^{(t)} - P^{(t)}\right)\right)\right] \tag{3.199}$$

We note that $Z^{(t)}W_\rho^{(t)}$ is a diagonal matrix with positive terms. Hence, $(Z^{(t)}W_\rho^{(t)})^{1/2}$ is a diagonal matrix with the square root of $Z^{(t)}W_\rho^{(t)}$. Thus, we compute the following expression:

$$E^{(t)} \leftarrow QR\left(\left(A_\beta(Z^{(t)}W\rho^{(t)})^{1/2}\right)^t\right)$$

$$Q^{(t)} \leftarrow QR.Q(E^{(t)})$$

$$R^{(t)} \leftarrow QR.R(E^{(t)})$$

$$\text{backsolve}\left(R^{(t)}, Q^{(t)^t}(Z^{(t)}W_\rho^{(t)})^{1/2}\left(A_\beta^t\beta_k^{(t)} + W\rho^{(t)^{-1}}\left(S^{(t)} - P^{(t)}\right)\right)\right)$$

3.2.6 Estimation of standard error

We use the method describe in subsection 2.3.4.

We compute the complete score function

$$S_C(\psi; y) = \left(\frac{\partial l_C(\psi; y)}{\partial \pi}, \frac{\partial l_C(\psi; y)}{\partial \beta}, \frac{\partial l_C(\psi; y)}{\partial \delta}\right)^t$$

where $\frac{\partial l_C(\psi;y)}{\partial \pi} = \left(\frac{\partial l_C(\psi;y)}{\partial \pi_1}, \dots, \frac{\partial l_C(\psi;y)}{\partial \pi_K}\right)$, $\frac{\partial l_C(\psi;y)}{\partial \beta} = \left(\frac{\partial l_C(\psi;y)}{\partial \beta_{11}}, \dots, \frac{\partial l_C(\psi;y)}{\partial \beta_{Kn_\beta}}\right)$ and $\frac{\partial l_C(\psi;y)}{\partial \delta} = \left(\frac{\partial l_C(\psi;y)}{\partial \delta_{11}}, \dots, \frac{\partial l_C(\psi;y)}{\partial \delta_{Kn_\delta}}\right)$.

It is important to recall that the sum of all the values in the sequence π is equal to 1. Using this fact, we can express π_K as $\sum_{k=1}^{K} \pi_k = 1$. Additionally, when j is not equal to K, the derivative of π_K with respect to π_j is equal to -1. Consequently, we can conclude that for all values of k from 1 to $K-1$, the following equation holds true:

$$\frac{\partial l_C(\psi; y)}{\partial \pi_k} = \sum_{i=1}^{n} \left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K} \right)$$

If we utilize equation 2.64 to forecast the likelihood of membership, with $1 \leq k \leq K$ and $1 \leq l \leq n_\theta$,

$$\frac{\partial l_C(\psi; y)}{\partial \theta_{kl}} = \sum_{i=1}^{n} x_{il} \left(Z_{ik} - \pi_{ik} \right)$$

For $1 \leq k \leq K$ and $1 \leq l \leq n_{\beta_k}$

$$\frac{\partial l_C(\psi; y)}{\partial \beta_{kl}} = \sum_{i=1}^{n} \sum_{t=1}^{T} z_{ik} a_{it}^{l-1} \left(y_{it} - \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \right)$$

For $1 \leq k \leq K$ and $1 \leq l \leq n_\delta$

$$\frac{\partial l_C(\psi; y)}{\partial \delta_{kl}} = \sum_{i=1}^{n} \sum_{t=1}^{T} z_{ik} w_{it}^{l} \left(y_{it} - \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \right)$$

3.2.6.1 Computation of the negative second derivative matrix

The negative of the second derivative matrix of the complete likelihood is,

$$-B_C(y; \hat{\psi}) = - \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \pi^2} & \frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \beta} & \frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \delta} \\ \frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \pi} & \frac{\partial^2 l_C(\psi;y)}{\partial \beta^2} & \frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \delta} \\ \frac{\partial^2 l_C(\psi;y)}{\partial \delta \partial \pi} & \frac{\partial^2 l_C(\psi;y)}{\partial \delta \partial \beta} & \frac{\partial^2 l_C(\psi;y)}{\partial \delta^2} \end{pmatrix}$$

The dimensions of this matrix can be determined by evaluating the expression:
$\left(\sum_{k=1}^{K} (n_{\beta_k} + n_\delta) + K - 1 \right) \times \left(\sum_{k=1}^{K} (n_{\beta_k} + n_\delta) + K - 1 \right)$.
If we are using a predictor for probability, we need to make a modification: we replace all instances of the derivative π with θ. Consequently, the dimensions of the matrix will be given by $\left(\sum_{k=1}^{K} (n_{\beta_k} + n_\delta) + (K-1)n_\theta \right) \times \left(\sum_{k=1}^{K} (n_{\beta_k} + n_\delta) + (K-1)n_\theta \right)$.

3.2.6.1.1 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \pi^2}$ or $\frac{\partial^2 l_C(\psi;y)}{\partial \theta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \pi^2}$ has for size $(K-1) \times (K-1)$.

For $1 \leq k, l \leq K-1$ is

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi^2} = \left(\frac{\partial^2 l_C(\psi; y)}{\partial \pi_k \partial \pi_l} \right)_{kl}$$

which elements are for $1 \leq k, l \leq K-1$

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi_k \partial \pi_l} = \begin{cases} \sum_{i=1}^n -\left(\frac{z_{ik}}{\pi_k^2} + \frac{z_{iK}}{\pi_K^2} \right), \; k = l \\ \sum_{i=1}^n -\frac{z_{iK}}{\pi_K^2}, \; k \neq l \end{cases}$$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \theta^2}$ has for size $Kn_\theta \times Kn_\theta$ and is composed by

$$\frac{\partial^2 l_C(\psi; y)}{\partial \theta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \theta_1 \partial \theta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial \theta_K \partial \theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \theta_K \partial \theta_K} \end{pmatrix}$$

which elements, for $1 \leq l, l' \leq n_\theta$, are

$$\left(\frac{\partial^2 l_C(\psi; y)}{\partial \theta_k \partial \theta_{k'}} \right)_{ll'} = \frac{\partial^2 l_C(\psi; y)}{\partial \theta_{kl} \partial \theta_{k'l'}} = \begin{cases} \sum_{i=1}^n -x_{il} x_{il'} \left(\frac{e^{\theta_k x_i}}{\sum_{k=1}^K e^{\theta_k x_i}} \right) \left(1 - \frac{e^{\theta_k x_i}}{\sum_{k=1}^K e^{\theta_k x_i}} \right), \; k = k' \\ \sum_{i=1}^n x_{il} x_{il'} \frac{e^{\theta_k x_i}}{\sum_{k=1}^K e^{\theta_k x_i}} \frac{e^{\theta_{k'} x_i}}{\sum_{k=1}^K e^{\theta_k x_i}}, \; k \neq k' \end{cases}$$

The outcomes remain the same for both π and θ in the subsequent sections.

3.2.6.1.2 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \beta}$

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi \partial \beta} = 0$$

3.2.6.1.3 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \delta}$

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi \partial \delta} = 0$$

3.2.6.1.4 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \beta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \beta^2}$ has for size $\sum_{k=1}^K n_{\beta_k} \times \sum_{k=1}^K n_{\beta_k}$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\beta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\beta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\beta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\beta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l, l' \leq n_\beta$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'}\partial\beta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'l'}\partial\beta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{-z_{ik} a_{it}^{l+l'-2} e^{\beta_k A_{it}+\delta_k W_{it}}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2}, \ k = k' \\ 0, \ k \neq k' \end{cases}$$

3.2.6.1.5 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta}$ has for size $\sum_{k=1}^K n_{\beta_k} \times K n_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$, $1 \leq l \leq n_\beta$ and $1 \leq l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'}\partial\delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'l'}\partial\delta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{-z_{ik} a_{it}^{l-1} w_{it}^{l'} e^{\beta_k A_{it}+\delta_k W_{it}}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2}, \ k = k' \\ 0, \ k \neq k' \end{cases}$$

3.2.6.1.6 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2}$ has for size $K n_\delta \times K n_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq, l, l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi; y)}{\partial \delta_{k'} \partial \delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi; y)}{\partial \delta_{k'l'} \partial \delta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{-z_{ik} w_{it}^{l+l'} e^{\beta_k A_{it} + \delta_k W_{it}}}{\left(1 + e^{\beta_k A_{it} + \delta_k W_{it}}\right)^2}, \ k = k' \\ 0, \ k \neq k' \end{cases}$$

3.2.7 Computation of $cov\left(S_C(\hat{\psi}; u) | U = u\right)$

The conditional matrix of the score vector is given by

$$I_{y/u}\left(y; \hat{\psi}\right) = \begin{pmatrix} cov\left(S_c(\pi)\right) & cov\left(S_c(\pi), S_c(\beta)\right) & cov\left(S_c(\pi), S_c(\delta)\right) \\ cov\left(S_c(\beta), S_c(\pi)\right) & cov\left(S_c(\beta)\right) & cov\left(S_c(\beta), S_c(\delta)\right) \\ cov\left(S_c(\delta), S_c(\pi)\right) & cov\left(S_c(\delta), S_c(\beta)\right) & cov\left(S_c(\delta)\right) \end{pmatrix}$$

The size of this matrix is $(K(n_\beta + n_\delta + 1) - 1) \times (K(n_\beta + n_\delta + 1) - 1)$ or $(K(n_\beta + n_\delta + n_\theta + 1)) \times (K(n_\beta + n_\delta + n_\theta + 1))$ in the case where we are using predictors for the probability.

As a reminder (see proposition 2 page 45),

- $E(Z_{ik}) = \tau_{ik}$
- $var\left(Z_{ik}\right) = \tau_{ik}(1 - \tau_{ik})$
- $cov\left(Z_{ik}, Z_{il}\right) = -\tau_{ik}\tau_{il}$ for $k \neq l$
- $cov\left(Z_{ik}, Z_{jl}\right) = 0$

3.2.7.1 Matrix $cov\left(S_c(\pi)\right)$

The matrix as for dimension $(K - 1) \times (K - 1)$.

For a diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write for $1 \leq k \leq K - 1$

$$cov\left(S_c(\pi)\right)_{kk} = \sum_{i=1}^n \left(\frac{\tau_{ik}(1 - \tau_{ik})}{\pi_k^2} + \frac{\tau_{iK}(1 - \tau_{iK})}{\pi_K^2} + 2\frac{\tau_{ik}\tau_{iK}}{\pi_k \pi_K}\right)$$

For a non-diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write for $1 \leq k, l \leq K - 1$.

$$cov\left(S_c(\pi)\right)_{kl} = \sum_{i=1}^n \left(-\frac{\tau_{ik}\tau_{il}}{\pi_k \pi_l} + \frac{\tau_{ik}\tau_{iK}}{\pi_k \pi_K} + \frac{\tau_{iK}\tau_{il}}{\pi_K \pi_l} + \frac{\tau_{iK}(1 - \tau_{iK})}{\pi_K^2}\right)$$

3.2.7.2 Matrix $cov\left(S_c(\pi), S_c(\beta)\right)$

$cov\left(S_c(\pi), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\beta_k)\right)$ for all groups k which dimension is $(K - 1) \times n_{\beta_k}$. Thus, the dimension of the first matrix is $(K - 1) \times K n_{\beta_k}$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\beta_k)\right)$ that is a matrix with elements, for $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^{n} \left(BL_{ikl}\tau_{ik} \left(\frac{1-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' = k \\ \sum_{i=1}^{n} \left(BL_{ikl}\tau_{ik} \left(-\frac{\tau_{ik'}}{\pi_{k'}} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\beta_K)\right)_{k'l} = \sum_{i=1}^{n} \left(BL_{iKl}\tau_{iK} \left(\frac{1-\tau_{iK}}{\pi_K} + \frac{\tau_{ik'}}{\pi'_k} \right) \right)$$

where $BL_{ikl} = \sum_{t=1}^{T} a_{it}^{l-1} \left(y_{it} - \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \right)$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$.

Proof. For $1 \leq k \leq K-1$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{k'l} = cov\left(\sum_{i=1}^{n} \left(\frac{Z_{ik'}}{\pi_{k'}} - \frac{Z_{iK}}{\pi_K} \right), \sum_{i=1}^{n} Z_{ik} BL_{ikl} \right) \tag{3.200}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \left(cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jk} BL_{jkl} \right) - cov\left(\frac{Z_{iK}}{\pi_K}, Z_{jk} BL_{jkl} \right) \right) \tag{3.201}$$

$$= \begin{cases} \sum_{i=1}^{n} \left(BL_{ikl}\tau_{ik} \left(\frac{1-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' = k \\ \sum_{i=1}^{n} \left(BL_{ikl}\tau_{ik} \left(-\frac{\tau_{ik'}}{\pi'_k} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' \neq k \end{cases} \tag{3.202}$$

For $k = K$,

$$cov\left(S_c(\pi), S_c(\beta_K)\right)_{k'l} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jK} BL_{jKl} \right) - cov\left(\frac{Z_{iK}}{\pi_K}, Z_{jK} BL_{jKl} \right) \right) \tag{3.203}$$

$$= \sum_{i=1}^{n} \left(BL_{iKl}\tau_{iK} \left(\frac{1-\tau_{iK}}{\pi_K} + \frac{\tau_{ik'}}{\pi'_k} \right) \right) \tag{3.204}$$

□

3.2.7.3 Matrix $cov\left(S_c(\pi), S_c(\delta)\right)$

$cov\left(S_c(\pi), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\delta_k)\right)$ for all groups k which dimension is $(K-1) \times n_\delta$. Thus, the dimension of the first matrix is $(K-1) \times Kn_\delta$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\delta_k)\right)$ that is a matrix with elements, for

$1 \leq k' \leq K-1$ and $1 \leq l \leq n_\delta$

$$cov\left(S_c(\pi), S_c(\delta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^n \left(DL_{ikl}\tau_{ik}\left(\frac{1-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right)\right), k' = k \\ \sum_{i=1}^n \left(DL_{ikl}\tau_{ik}\left(-\frac{\tau_{ik'}}{\pi_{k'}} + \frac{\tau_{iK}}{\pi_K}\right)\right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\delta_K)\right)_{k'l} = \sum_{i=1}^n \left(DL_{iKl}\tau_{iK}\left(\frac{1-\tau_{iK}}{\pi_K} + \frac{\tau_{ik'}}{\pi_k'}\right)\right)$$

where $DL_{ikl} = \sum_{t=1}^T w_{it}^l \left(y_{it} - \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}}\right)$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_\delta$.

3.2.7.4 Matrix $cov\left(S_c(\theta)\right)$

If we use predictors for the membership probability we have to calculate the matrix with θ parameters.

$cov\left(S_c(\theta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\theta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_\theta$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\theta$.

A diagonal matrix, for $1 \leq k \leq K$ and $1 \leq p, q \leq n_\theta$ is done by

$$\left(cov\left(S_c(\theta_k), S_c(\theta_k)\right)\right)_{pq} = \sum_{i=1}^n x_{ip}x_{iq}\tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$ and $1 \leq p, q \leq n_\theta$, is done by

$$\left(cov\left(S_c(\theta_k), S_c(\theta_l)\right)\right)_{pq} = -\sum_{i=1}^n x_{ip}x_{iq}\tau_{ik}\tau_{il}$$

3.2.7.5 Matrix $cov\left(S_c(\theta), S_c(\beta)\right)$

$cov\left(S_c(\theta), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\beta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_{\beta_l}$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\beta$.

A diagonal matrix, for $1 \leq k \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\beta_k}$ is done by

$$\left(cov\left(S_c(\theta_k), S_c(\beta_k)\right)\right)_{pq} = \sum_{i=1}^n x_{ip}BL_{ikq}\tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\beta_l}$, is done by

$$(cov\,(S_c(\theta_k), S_c(\beta_l)))_{pq} = \sum_{i=1}^{n} -x_{ip} BL_{ilq} \tau_{ik} \tau_{il}$$

Proof. For a diagonal matrix, $1 \leq q \leq n_{\beta_k}$ and $1 \leq p \leq n_\theta$

$$\begin{aligned}
(cov\,(S_c(\theta_k), S_c(\beta_k)))_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\,(Z_{ik} - \pi_{ik})\,, \sum_{i=1}^{n} Z_{ik} BL_{ikq}\right) && (3.205)\\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} cov\,(x_{ip} Z_{ik}, Z_{jk} BL_{jkq}) && (3.206)\\
&= \sum_{i=1}^{n} x_{ip} BL_{ikq} \tau_{ik} (1 - \tau_{ik}) && (3.207)
\end{aligned}$$

For a non diagonal matrix, $1 \leq q \leq n_{\beta_l}$ and $1 \leq p \leq n_\theta$,

$$\begin{aligned}
(cov\,(S_c(\theta_k), S_c(\beta_l)))_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\,(Z_{ik} - \pi_{ik})\,, \sum_{i=1}^{n} Z_{il} BL_{ilq}\right) && (3.208)\\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} cov\,(x_{ip} Z_{ik}, Z_{jl} BL_{jlq}) && (3.209)\\
&= \sum_{i=1}^{n} -x_{ip} BL_{ilq} \tau_{ik} \tau_{il} && (3.210)
\end{aligned}$$

□

3.2.7.6 Matrix $cov\,(S_c(\theta), S_c(\delta))$

$cov\,(S_c(\theta), S_c(\delta))$ is composed by the matrix $cov\,(S_c(\theta_k), S_c(\delta_l))$ for all groups k and l which dimension is $n_\theta \times n_\delta$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\delta$.

A diagonal matrix, for $1 \leq k \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_\delta$ is done by

$$(cov\,(S_c(\theta_k), S_c(\delta_k)))_{pq} = \sum_{i=1}^{n} x_{ip} DL_{ikq} \tau_{ik} (1 - \tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_\delta$, is done by

$$(cov\,(S_c(\theta_k), S_c(\delta_l)))_{pq} = \sum_{i=1}^{n} -x_{ip} DL_{ilq} \tau_{ik} \tau_{il}$$

Proof. For a diagonal matrix, $1 \leq q \leq n_\delta$ and $1 \leq p \leq n_\theta$

$$\left(cov\left(S_c(\theta_k), S_c(\delta_k)\right)\right)_{pq} = cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} Z_{ik} DL_{ikq}\right) \quad (3.211)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(x_{ip}Z_{ik}, Z_{jk}DL_{jkq}\right) \quad (3.212)$$

$$= \sum_{i=1}^{n} x_{ip} DL_{ikq} \tau_{ik}(1-\tau_{ik}) \quad (3.213)$$

For a non diagonal matrix, $1 \leq q \leq n_\delta$ and $1 \leq p \leq n_\theta$,

$$\left(cov\left(S_c(\theta_k), S_c(\delta_l)\right)\right)_{pq} = cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n} Z_{il} DL_{ilq}\right) \quad (3.214)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(x_{ip}Z_{ik}, Z_{jl}DL_{jlq}\right) \quad (3.215)$$

$$= \sum_{i=1}^{n} -x_{ip} DL_{ilq} \tau_{ik}\tau_{il} \quad (3.216)$$

□

3.2.7.7 Matrix $cov\left(S_c(\beta)\right)$

$cov\left(S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\beta_l)\right)$ for all groups k and l which dimension is $n_{\beta_k} \times n_{\beta_l}$. Thus, the dimension of the first matrix is $Kn_{\beta_k} \times Kn_{\beta_l}$.

A diagonal matrix, for $1 \leq k \leq K$ and $1 \leq p, q \leq n_{\beta_k}$ is done by

$$\left(cov\left(S_c(\beta_k), S_c(\beta_k)\right)\right)_{pq} = \sum_{i=1}^{n} BL_{ikp} BL_{ikq} \tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\beta_l}$, is done by

$$\left(cov\left(S_c(\beta_k), S_c(\beta_l)\right)\right)_{pq} = \sum_{i=1}^{n} -BL_{ikp} BL_{ilq}\left(\tau_{ik}\tau_{il}\right)$$

Proof. For a diagonal matrix, for $1 \leq p, q \leq n_{\beta_k}$

$$\left(cov\left(S_c(\beta_k), S_c(\beta_k)\right)\right)_{pq} = cov\left(\sum_{i=1}^{n} Z_{ik} BL_{ikp}, \sum_{i=1}^{n} Z_{ik} BL_{ikq}\right) \quad (3.217)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z_{ik}BL_{ikp}, Z_{jk}BL_{jkq}\right) \quad (3.218)$$

$$= \sum_{i=1}^{n} BL_{ikp}BL_{ikq}\tau_{ik}(1-\tau_{ik}) \tag{3.219}$$

For a non diagonal matrix, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\beta_l}$,

$$\begin{aligned}
\left(cov\left(S_c(\beta_k), S_c(\beta_l)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} Z_{ik}BL_{ikp}, \sum_{i=1}^{n} Z_{il}BL_{ilq}\right) && (3.220)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z_{ik}BL_{ikp}, Z_{jl}BL_{jlq}\right) && (3.221)\\
&= \sum_{i=1}^{n} -BL_{ikp}BL_{ilq}\left(\tau_{ik}\tau_{il}\right) && (3.222)
\end{aligned}$$

□

3.2.7.8 Matrix $cov\left(S_c(\beta), S_c(\delta)\right)$

$cov\left(S_c(\beta), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\delta_l)\right)$ for $1 \leq k \leq n_{\beta_k}$ and $1 \leq l \leq n_\delta$ which dimension is $n_{\beta_k} \times n_\delta$. Thus, the dimension of the first matrix is $Kn_{\beta_k} \times kn_\delta$.

An element of the matrix $cov\left(S_c(\beta_k), S_c(\delta_l)\right)$ for $1 \leq k, l \leq K$ is, for $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\delta_l}$

$$cov\left(S_c(\beta_k), S_c(\delta_l)\right)_{pq} = \begin{cases} \sum_{i=1}^{n} BL_{ikp}DL_{ikq}\tau_{ik}(1-\tau_{ik}), \ k = l \\ \sum_{i=1}^{n} -BL_{ikp}DL_{ilq}\tau_{ik}\tau_{il}, \ k \neq l \end{cases}$$

Proof. Let $1 \leq k, l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\delta_l}$

$$\begin{aligned}
cov\left(S_c(\beta_k), S_c(\delta_l)\right)_{pq} &= cov\left(\sum_{i=1}^{n} Z_{ik}BL_{ikp}, \sum_{i=1}^{n} Z_{il}DL_{ilq}\right) && (3.223)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} cov\left(Z_{ik}BL_{ikp}, Z_{jl}DL_{jlq}\right) && (3.224)\\
&= \begin{cases} \sum_{i=1}^{n} BL_{ikp}DL_{ikq}\tau_{ik}(1-\tau_{ik}), \ k = l \\ \sum_{i=1}^{n} -BL_{ikp}DL_{ilq}\tau_{ik}\tau_{il}, \ k \neq l \end{cases} && (3.225)
\end{aligned}$$

□

3.2.7.9 Matrix $cov\left(S_c(\delta)\right)$

$cov\left(S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\delta_k), S_c(\delta_l)\right)$ for all groups k and l which dimension is $n_\delta \times n_\delta$. Thus, the dimension of the first matrix is $Kn_\delta \times Kn_\delta$.

A diagonal matrix, for $1 \leq p, q \leq n_\delta$ is done by

$$(cov\left(S_c(\delta_k), S_c(\delta_k)\right))_{pq} = \sum_{i=1}^{n} DL_{ikp} DL_{ikq} \tau_{ik}(1 - \tau_{ik})$$

A non diagonal matrix, $1 \leq p, q \leq n_\delta$, is done by

$$(cov\left(S_c(\delta_k), S_c(\delta_l)\right))_{pq} = \sum_{i=1}^{n} -DL_{ikp} DL_{ilq} \left(\tau_{ik}\tau_{il}\right)$$

3.2.8 Numerical application

In the following sections, we will assess the accuracy of the various equations by comparing their results to those obtained from the SAS procedure traj. In each example, we generate a sample dataset consisting of K clusters, each comprising 500 observations of a binary variable Y. The trajectories of these observations are structured to follow a group-specific pattern. This pattern is achieved by modeling the probability $\rho_{ikt} = P(Y_{it} = 1|W_i = w_i, C_i = k)$ using a polynomial function, specifically $\rho_{ikt} = \dfrac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}}$.

This approach allows us to compare the theoretically expected parameter values with those obtained from the SAS procedure traj, EM algorithm, EM IRWLS, and the Likelihood method presented earlier.

3.2.8.1 Two groups

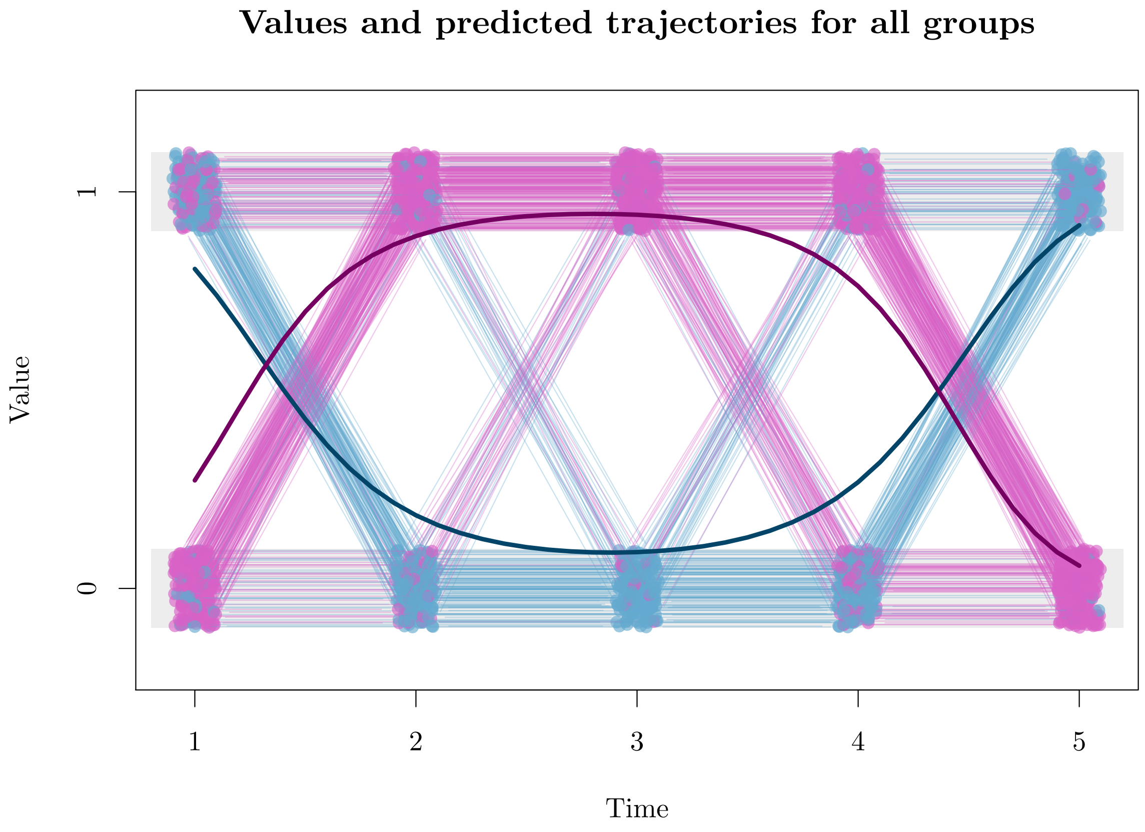
We set theoretical values as

	Polynomial		
Cluster	Degree	Shape	Probability π_k
1	2	$\beta_1 = (6.32, -5.8, 1)$	0.32
2	2	$\beta_2 = (-6.69, 6.92, -1.23)$	0.68

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms. The default evaluation of Traj yields the following values: $\beta_1 = (-1.37214, 0, 0)$, $\beta_2 = (1.87751, 0, 0)$, $\pi_1 = 0.5$. Consequently, we obtain the following results:

	Theoretical	Likelihood Traj		Likelihood		EM		EM - IRWLS	
			SE		SE		SE		SE
β_{11}	6.32	6.44400	0.48633	6.44401	0.4826	6.4392	0.48145	6.44401	0.48171
β_{12}	-5.80	-6.07143	0.41677	-6.07144	0.41127	-6.06715	0.4112	-6.07144	0.41141
β_{13}	1.00	1.05217	0.07239	1.05217	0.07125	1.05145	0.0714	1.05217	0.07144
β_{21}	-6.69	-6.36800	0.37241	-6.368	0.36963	-6.37234	0.36869	-6.368	0.36847
β_{22}	6.92	6.55591	0.34949	6.55591	0.34533	6.56008	0.3451	6.55591	0.34486
β_{23}	-1.23	-1.16831	0.06374	-1.16831	0.06266	-1.16905	0.06289	-1.16831	0.06284
π_1	0.32	0.377368	0.023127	0.37737	0.02338	0.37744	0.02207	0.37737	0.02207
π_2	0.68	0.622632	0.023127	0.62263	0.02338	0.62256	0.02207	0.62263	0.02207

To visually represent the data, we have provided the graphic below. Given that the values are restricted to 0 or 1, they all occupy the same location, making it challenging to distinguish individual points. To address this issue, we applied a slight random shift to each point and included a band to indicate the 0 and 1 values. This approach allows us to highlight different points for each value and illustrate the trajectories leading to these points. The shape is depicted more prominently with darker shading.



3.2.8.2 Three groups

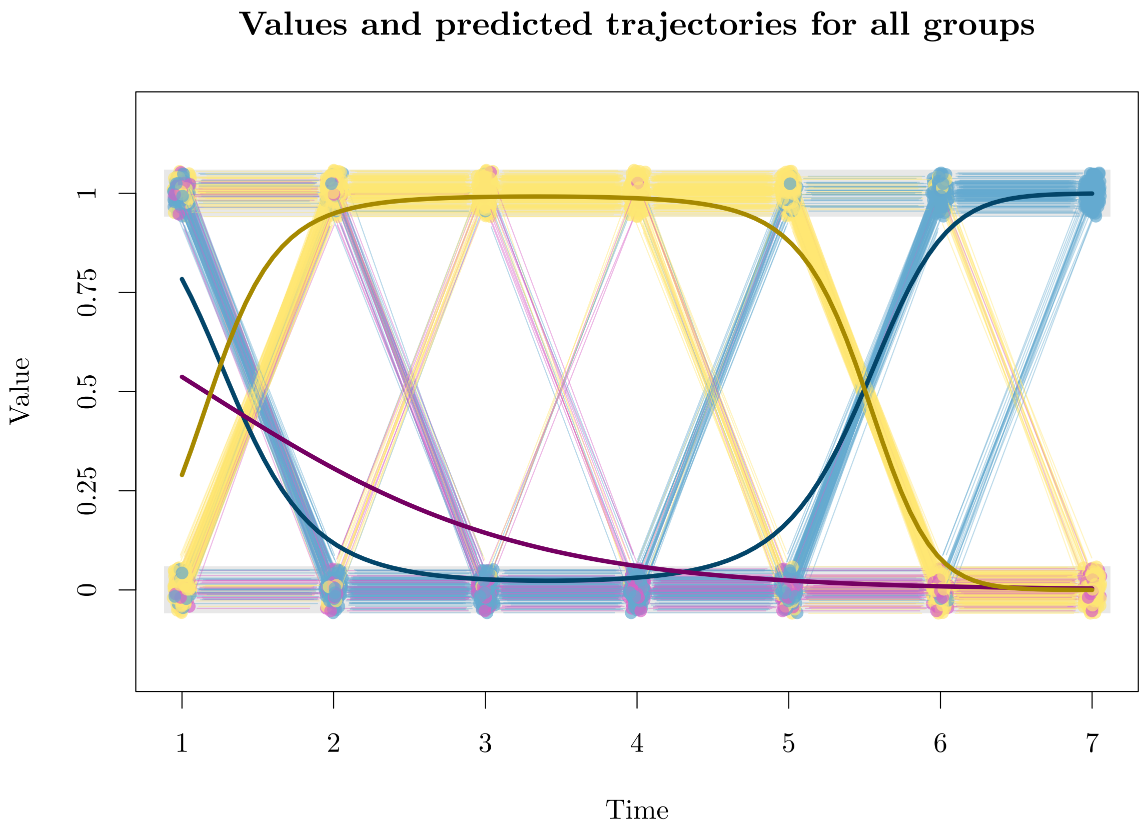
We set theoretical values as

	Polynomial		
Cluster	Degree	Shape	Probability π_k
1	2	$\beta_1 = (6.32, -5.8, 0.85)$	0.32
2	2	$\beta_2 = (-6.69, 6.92, -1.03)$	0.54
3	1	$\beta_3 = (1, -0.83)$	0.14

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms. The default evaluation of Traj yields the following values: $\beta_1 = (-3, 0, 0)$, $\beta_2 = (3.99763, 0, 0)$, $\beta_3 = (-0.07089, 0)$, $\pi_1 = \pi_2 = \pi_3 = 1/3$. Consequently, we obtain the following results:

	Theoretical	Likelihood Traj		Likelihood		EM		EM - IRWLS	
			SE		SE		SE		SE
β_{11}	6.32	6.31493	0.41218	6.31493	0.39688	6.3155	0.41167	6.31493	0.41165
β_{21}	-5.80	-5.89028	0.32610	-5.89029	0.31635	-5.89085	0.32571	-5.89029	0.32568
β_{13}	0.85	0.86311	0.04761	0.86311	0.04636	0.86319	0.04755	0.86311	0.04755
β_{21}	-6.69	-6.79063	0.34699	-6.79062	0.33389	-6.79312	0.34657	-6.79063	0.34648
β_{22}	6.92	6.92935	0.31837	6.92935	0.30366	6.93168	0.31799	6.92935	0.31788
β_{23}	-1.03	-1.03462	0.04751	-1.03462	0.04527	-1.03496	0.04745	-1.03462	0.04744
β_{31}	1	1.11515	0.30950	1.11515	0.29868	1.11512	0.30908	1.11515	0.30909
β_{32}	-0.83	-0.96600	0.13215	-0.966	0.12662	-0.96596	0.13194	-0.966	0.13195
π_1	0.32	0.32185	0.209739	0.32185	0.02113	0.32185	0.02094	0.32185	0.02094
π_2	0.54	0.53989	0.224251	0.53989	0.0227	0.53988	0.0261	0.53989	0.0261
π_3	0.14	0.13826	0.157080	0.13826	0.01527	0.13827	0.01558	0.13826	0.01558

To visually represent the data, we have provided the graphic below. Given that the values are restricted to 0 or 1, they all occupy the same location, making it challenging to distinguish individual points. To address this issue, we applied a slight random shift to each point and included a band to indicate the 0 and 1 values. This approach allows us to highlight different points for each value and illustrate the trajectories leading to these points. The shape is depicted more prominently with darker shading.



3.3 ZIP distribution

3.3.1 Definition

The Zero Inflated Poisson (ZIP) model is applied to situations where a random event exhibits an excess of zero counts within a specific unit of time. Typically, we use a Poisson distribution to model scenarios that are rare and therefore result in many zero values. However, in some cases, the number of zeros is much higher than what a Poisson distribution can adequately capture. In such instances, the ZIP model proves valuable in addressing this phenomenon.

For example, when dealing with insurance claims, one might encounter zero-inflated data. You can refer to sources such as Mouatassim and Ezzahid (2012) or Sarul (2015) for more information on this topic. The ZIP model incorporates two distinct processes: a binary distribution that generates structural zeros, representing the excess zeros in the data, and a Poisson distribution that generates count data.

In this model, we consider the random variable Y_{it}, where $1 \leq i \leq n$ and $1 \leq t \leq T$. Y_{it} can only take non-negative integer values. We assume that the probability of observing Y_{it}, denoted as y_{it}, given a group k and covariate W_i, follows a Poisson distribution with a parameter λ_{ikt}. This parameter can be viewed as a deviation from the mean rate, λ_{kt}, of the event's occurrence for all individuals in group k at time t. Therefore, for the count part of the model, we use a conditional Poisson distribution given the group k and covariate W_i, represented as $\mathcal{P}\left(\lambda_{ikt}\right)$.

$$P\left(Y_{it} = y_{it} | W_i = w_i, C_i = k\right) = \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!}$$

Moreover, for a given time period t, we can consider that a member i of a group k is inactive with a rate ρ_{ikt}. This can be seen as a deviation from the mean rate ρ_{kt} of event occurrence for all individuals in group k at time t.

In this model, we estimate the parameters with a conditional binary distribution given the group k and covariate W_i, denoted as $\mathcal{B}\left(\rho_{ikt}\right)$:

$$P(\text{excess zero}_{it} | W_i = w_i, C_i = k) = \rho_{ikt}$$

Consequently, in this model, zero values can arise from two sources. One possibility is that the count is equal to zero with a probability $P(Y_{it} = 0)$ when $Y_{it} \sim \mathcal{P}(\lambda_{ikt})$. The second source is due to the binary process, which produces zero values with a probability ρ_{ikt}.

$$(Y_{it} = y_{it} | W_i = wi, C_i = k) \sim \rho_{ikt} + (1 - \rho_{ikt})\mathcal{P}(\lambda_{ikt})$$

Finally, we have

$$P\left(Y_{it} = y_{it} | W_i = wi, C_i = k\right) = \begin{cases} \rho_{ikt} + (1 - \rho_{ikt})e^{-\lambda_{ikt}}, \ y_{it} = 0 \\ (1 - \rho_{ikt})\frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!}, \ y_{it} > 0 \end{cases} \tag{3.226}$$

It's evident that when $\rho_{ikt} = 0$, the ZIP model simplifies to the classical Poisson model.

3.3.2 Link function

We select a polynomial function as the link between the trajectory and the time variable. To be specific, we assume that λ_{ikt} varies over time according to the following equation:

$$\log\left(\lambda_{ikt}\right) = \beta_k A_{it} + \delta_k W_{it} \tag{3.227}$$

or

$$\lambda_{ikt} = e^{\beta_k A_{it} + \delta_k W_{it}} \tag{3.228}$$

where $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1})^t$, $W_{it} = (w_{i1}, \cdots, w_{in_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ and $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$.

We employ the logarithm of λ_{ikt}, denoted as $\log(\lambda_{ikt})$, because it ensures that λ_{ikt} remains positive, which is a requirement for rates.

Furthermore, we employ a polynomial function to describe the relationship between time and ρ_{ikt}. This relationship is defined as:

$$\log\left(\frac{\rho_{ikt}}{1-\rho_{ikt}}\right) = \nu_k A_{it} \tag{3.229}$$

This equation is equivalent to:

$$\rho_{ikt} = \frac{e^{\nu_k A_{it}}}{1 + e^{\nu_k A_{it}}} \tag{3.230}$$

where $\nu_k = \left(\nu_{k1}, \cdots, \nu_{kn_{\nu_k}}\right)$.

Thus the log-likelihood 2.10 becomes

$$l(\psi; y) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \nu_k)\right) \tag{3.231}$$

where

$$g_k(y_i; \beta_k, \delta_k, \nu_k) = \prod_{y_{it}=0} \left(\rho_{ikt} + (1-\rho_{ikt})e^{-\lambda_{ikt}}\right) \prod_{y_{it}>0} (1-\rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!} \tag{3.232}$$

3.3.3 Likelihood

To estimate the parameters, we utilize quasi-Newton methods, and we need to solve the equations 2.16 and 2.17. In this particular case, these equations become:

$$\frac{\partial l(\psi; y)}{\partial \theta_{kl}} = \sum_{i=1}^{n} \frac{\dfrac{\partial \pi_k}{\partial \theta_{kl}} g_k(y_i; \beta_k, \delta_k, \nu_k)}{\displaystyle\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \nu_k)} = 0 \quad 1 \le k \le K, \text{ and } 1 \le l \le n_\theta \tag{3.233}$$

$$\frac{\partial l(\psi; y)}{\partial \beta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \dfrac{\partial}{\partial \beta_{kl}} g_k(y_i; \beta_k, \delta_k, \nu_k)}{\displaystyle\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \nu_k)} = 0 \quad 1 \le k \le K, \text{ and } 1 \le l \le n_{\beta_k} \tag{3.234}$$

$$\frac{\partial l(\psi; y)}{\partial \delta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \dfrac{\partial}{\partial \delta_{kl}} g_k(y_i; \beta_k, \delta_k, \nu_k)}{\displaystyle\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \nu_k)} = 0 \quad 1 \le k \le K, \text{ and } 1 \le l \le n_\delta \tag{3.235}$$

$$\frac{\partial l(\psi;y)}{\partial \nu_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \nu_{kl}} g_k(y_i;\beta_k,\delta_k,\nu_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i;\beta_k,\delta_k,\nu_k)} = 0 \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_{\nu_k} \tag{3.236}$$

When we employ likelihood to fit the model, the probability membership takes the form $\pi_k = \frac{e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}}$. We will calculate the equation above in several steps. However, it should be noted that there is no closed-form solution.

3.3.3.1 Differential by θ_k

Same as section 2.2.

3.3.3.2 Differential by β_{kl}

Let $1 \leq l \leq n_{\beta_k}$, we start by calculating the derivatives of the two blocks by posing:

$$d_0^{\beta_{kl}} = \frac{\partial}{\partial \beta_{kl}} \left(\prod_{y_{it}=0} \left(\rho_{ikt} + (1-\rho_{ikt}) e^{-\lambda_{ikt}} \right) \right)$$

$$d_1^{\beta_{kl}} = \frac{\partial}{\partial \beta_{kl}} \left(\prod_{y_{it}>0} (1-\rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!} \right)$$

Noting that $\frac{\partial \lambda_{ikt}}{\partial \beta_{kl}} = a_{it}^{l-1} \lambda_{ikt}$ and $\frac{\partial e^{-\lambda_{ikt}}}{\partial \beta_{kl}} = -a_{it}^{l-1} \lambda_{ikt} e^{-\lambda_{ikt}}$ we can deduce the following equations:

$$d_0^{\beta_{kl}} = \sum_{y_{it}=0} -a_{it}^{l-1} \lambda_{ikt} (1-\rho_{ikt}) e^{-\lambda_{ikt}} \prod_{\substack{y_{it'}=0 \\ y_{it'} \neq y_{it}}} \left(\rho_{ikt'} + (1-\rho_{ikt'}) e^{-\lambda_{ikt'}} \right) \tag{3.237}$$

$$d_1^{\beta_{kl}} = \sum_{y_{it}>0} \frac{a_{it}^{l-1} \lambda_{ikt}^{y_{it}} (1-\rho_{ikt}) e^{-\lambda_{ikt}} \left(y_{it} - \lambda_{ikt} \right)}{y_{it}!} \prod_{\substack{y_{it'}=0 \\ y_{it'} \neq y_{it}}} (1-\rho_{ikt'}) \frac{\lambda_{ikt'}^{y_{it'}} e^{-\lambda_{ikt'}}}{y_{it'}!} \tag{3.238}$$

Next, we calculate the derivative of the function $g_k(y_i;\Theta_k)$:

$$\frac{\partial}{\partial \beta_{kl}} g_k(y_i;\beta_k,\delta_k,\nu_k) = d_0^{\beta_{kl}} \prod_{y_{it}>0} (1-\rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!} + d_1^{\beta_{kl}} \prod_{y_{it}=0} \left(\rho_{ikt} + (1-\rho_{ikt}) e^{-\lambda_{ikt}} \right)$$

3.3.3.3 Differential by δ_{kl}

Let $1 \leq l \leq n_{\delta}$, we start by calculating the derivatives of the two blocks by posing:

$$d_0^{\delta_{kl}} = \frac{\partial}{\partial \delta_{kl}} \left(\prod_{y_{it}=0} \left(\rho_{ikt} + (1-\rho_{ikt}) e^{-\lambda_{ikt}} \right) \right)$$

$$d_1^{\delta_{kl}} = \frac{\partial}{\partial \delta_{kl}} \left(\prod_{y_{it}>0} (1-\rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!} \right)$$

we have

$$d_0^{\delta_{kl}} = \sum_{y_{it}=0} -w_{it}^l \lambda_{ikt} (1-\rho_{ikt}) e^{-\lambda_{ikt}} \prod_{\substack{y_{it'}=0 \\ y_{it'} \neq y_{it}}} \left(\rho_{ikt'} + (1-\rho_{ikt'}) e^{-\lambda_{ikt'}} \right) \tag{3.239}$$

$$d_1^{\delta_{kl}} = \sum_{y_{it}>0} \frac{w_{it}^l \lambda_{ikt}^{y_{it}} (1-\rho_{ikt}) e^{-\lambda_{ikt}} \left(y_{it} - \lambda_{ikt} \right)}{y_{it}!} \prod_{\substack{y_{it'}>0 \\ y_{it'} \neq y_{it}}} (1-\rho_{ikt'}) \frac{\lambda_{ikt'}^{y_{it'}} e^{-\lambda_{ikt'}}}{y_{it'}!} \tag{3.240}$$

Next, we calculate the derivative of the function $g_k(y_i; \Theta_k)$:

$$\frac{\partial}{\partial \delta_{kl}} g_k(y_i; \delta_k, \delta_k) = d_0^{\delta_{kl}} \prod_{y_{it}>0} (1-\rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!} + d_1^{\delta_{kl}} \prod_{y_{it}=0} \left(\rho_{ikt} + (1-\rho_{ikt}) e^{-\lambda_{ikt}} \right)$$

3.3.3.4 Differential by ν_{kl}

Let $1 \leq l \leq n_{\nu_k}$, we start by calculating the derivatives of the two blocks by posing:

$$d_0^{\nu_{kl}} = \frac{\partial}{\partial \nu_{kl}} \left(\prod_{y_{it}=0} \left(\rho_{ikt} + (1-\rho_{ikt}) e^{-\lambda_{ikt}} \right) \right)$$

$$d_1^{\nu_{kl}} = \frac{\partial}{\partial \nu_{kl}} \left(\prod_{y_{it}>0} (1-\rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!} \right)$$

Recognizing that $\frac{\partial \rho_{ikt}}{\partial \nu_{kl}} = \frac{a_{it}^{l-1} \rho_{ikt}}{1+e^{\nu_k A_{it}}}$, we can proceed with the calculation.

$$d_0^{\nu_{kl}} = \sum_{y_{it}=0} \frac{a_{it}^{l-1} \rho_{ikt} \left(1 - e^{-\lambda_{ikt}} \right)}{\left(1 + e^{\nu_k A_{it}} \right)} \prod_{\substack{y_{it'}=0 \\ y_{it'} \neq y_{it}}} \left(\rho_{ikt'} + (1-\rho_{ikt'}) e^{-\lambda_{ikt'}} \right) \tag{3.241}$$

$$d_1^{\nu_{kl}} = \sum_{y_{it}>0} \frac{-a_{it}^{l-1} \rho_{ikt} \lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{\left(1 + e^{\nu_k A_{it}} \right) y_{it}!} \prod_{\substack{y_{it'}>0 \\ y_{it'} \neq y_{it}}} (1-\rho_{ikt'}) \frac{\lambda_{ikt'}^{y_{it'}} e^{-\lambda_{ikt'}}}{y_{it'}!} \tag{3.242}$$

Secondly we calculate the differential of g_k,

$$\frac{\partial}{\partial \nu_{kl}} g_k(y_i; \beta_k, \delta_k, \nu_k) = d_0^{\nu_{kl}} \prod_{y_{it}>0} (1-\rho_{ikt}) \frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!} + d_1^{\nu_{kl}} \prod_{y_{it}=0} \left(\rho_{ikt} + (1-\rho_{ikt}) e^{-\lambda_{ikt}} \right)$$

3.3.4 EM algorithm

We can consider the problem as having two missing data: the group to which Y_{it} belongs and the zero state of Y_{it}. Let S_{it} be the covariate, which, given a group k, is 0 if Y_{it} is in a Poisson state and 1 if Y_{it} is in an excess state. Thus, $(S_{it}|Z_{ik}) \sim \mathcal{B}(\rho_{ikt})$. Moreover, we suppose that the covariates $(Y_{it}, S_{it})_{it}$ are independent conditionally to Z_{ik}, meaning that, given a group k, the different values of $(Y_{it}, S_{it})_{it}$ are independent.

We define $X_c = (Y, S, Z)$ as the complete data, where Y is the given data, $S = (S_1, \cdots S_n)^t$ with $S_i = (S_{i1}, \cdots, S_{iT})$, and $Z = (Z_1, \cdots Z_n)^t$ with $Z_i = (Z_{i1}, \cdots, Z_{iK})$.

The EM algorithm can be expressed as:

$$L_C(\psi; y) = P(Y, S, Z) \tag{3.243}$$

$$= \prod_{i=1}^{n} P(Y_i, S_i, Z_i) \tag{3.244}$$

$$= \prod_{i=1}^{n} P(Y_i, S_i|Z_i) P(Z_i) \tag{3.245}$$

Or,

$$P(Z_i) = P(Z_{i1} = z_{i1}, \cdots, Z_{iK} = z_{iK}) = \prod_{k=1}^{K} \pi_k^{Z_{ik}} \tag{3.246}$$

Thus we have

$$L_C(\psi; y) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(P(Y_i, S_i|Z_{ik} = 1) \pi_k\right)^{Z_{ik}} \tag{3.247}$$

Remember that that the covariate $(Y_{it}, S_{it})_{it}$ are independent conditionally to Z_{ik}.

$$P(Y_i, S_i|Z_{ik} = 1) = \prod_{t=1}^{T} P(Y_{it}, S_{it}|Z_{ik} = 1) \tag{3.248}$$

$$= \prod_{t=1}^{T} P(Y_{it}|Z_{ik} = 1, S_{it}) P(S_{it}|Z_{ik} = 1) \tag{3.249}$$

$(S_{it}|Z_{ik} = 1)$ can take only 2 values : 0 and 1 similar to a binomial distribution $\mathcal{B}(\rho_{ikt})$. Consequently

$$P(S_{it}|Z_{ik} = 1) = (\rho_{ikt})^{S_{it}} (1 - \rho_{ikt})^{(1-S_{it})} \tag{3.250}$$

Given S_{it}, we know that Y_{it} comes from either a zero-excess state or a Poisson state. Thus,

$$P\left(Y_{it}|Z_{ik}=1,S_{it}\right)=\delta_0(Y_{it})^{S_{it}}\left(P_{Poisson}(\lambda_{ikt},Y_{it})\right)^{1-S_{it}}$$

We observe that $\delta_0(Y_{it})^{S_{it}}=1$. In other words, if $S_{it}=0$, i.e., Y_{it} is in a Poisson state, the expression simplifies to 1. If $S_{it}=1$, i.e., Y_{it} is in a zero-excess state, then $\delta_0(Y_{it})=1$, and again, the expression simplifies to 1. Hence, we have:

$$P\left(Y_{it}|Z_{ik}=1,S_{it}\right)=\left(\frac{\lambda_{ikt}^{Y_{it}}e^{-\lambda_{ikt}}}{Y_{it}!}\right)^{1-S_{it}}$$

By substituting this expression into equation (3.247), we can derive the complete likelihood.

$$L_C\left(\psi;y\right)=\prod_{i=1}^{n}\prod_{k=1}^{K}\left(\prod_{t=1}^{T}\left[(\rho_{ikt})^{S_{it}}\left(1-\rho_{ikt}\right)^{(1-S_{it})}\left(\frac{\lambda_{ikt}^{Y_{it}}e^{-\lambda_{ikt}}}{Y_{it}!}\right)^{1-S_{it}}\right]\pi_k\right)^{Z_{ik}} \tag{3.251}$$

$$=\prod_{i=1}^{n}\prod_{k=1}^{K}\prod_{t=1}^{T}(\rho_{ikt})^{Z_{ik}S_{it}}\left(1-\rho_{ikt}\right)^{Z_{ik}(1-S_{it})}\left(\frac{\lambda_{ikt}^{Y_{it}}e^{-\lambda_{ikt}}}{Y_{it}!}\right)^{Z_{ik}(1-S_{it})}\pi_k^{\frac{Z_{ik}}{T}} \tag{3.252}$$

And, finally, the complete log-likelihood.

$$l_C\left(\psi;y\right)=\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}[Z_{ik}(1-S_{it})\left(Y_{it}\log\left(\lambda_{ikt}\right)-\lambda_{ikt}-\log\left(Y_{it}!\right)\right) \tag{3.253}$$

$$+Z_{ik}S_{it}\log\left(\rho_{ikt}\right)+Z_{ik}(1-S_{it})\log\left(1-\rho_{ikt}\right)+\frac{Z_{ik}}{T}\log\left(\pi_k\right)] \tag{3.254}$$

$$=\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}[Z_{ik}(1-S_{it})\left(Y_{it}\log\left(\lambda_{ikt}\right)-\lambda_{ikt}-\log\left(Y_{it}!\right)\right) \tag{3.255}$$

$$+Z_{ik}S_{it}\left(\nu_kA_{it}-\log\left(1+e^{\nu_kA_{it}}\right)\right)-Z_{ik}(1-S_{it})\log\left(1+e^{\nu_kA_{it}}\right) \tag{3.256}$$

$$+\frac{Z_{ik}}{T}\log\left(\pi_k\right)] \tag{3.257}$$

$$=\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}[Z_{ik}(1-S_{it})\left(Y_{it}\left(\beta_kA_{it}+\delta_kW_t\right)-e^{\beta_kA_{it}+\delta_kW_t}-\log\left(Y_{it}!\right)\right) \tag{3.258}$$

$$+Z_{ik}S_{it}\nu_kA_{it}-Z_{ik}\log\left(1+e^{\nu_kA_{it}}\right)+\frac{Z_{ik}}{T}\log\left(\pi_k\right)] \tag{3.259}$$

$$=\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{k=1}^{K}[l_{C,\beta,\delta}\left(\psi;y\right)+l_{C,\nu}\left(\psi;y\right)+\frac{Z_{ik}}{T}\log\left(\pi_k\right)] \tag{3.260}$$

Where $l_{C,\beta,\delta}\left(\psi;y\right)=Z_{ik}(1-S_{it})\left(Y_{it}\left(\beta_kA_{it}+\delta_kW_t\right)-e^{\beta_kA_{it}+\delta_kW_t}-\log\left(Y_{it}!\right)\right)$ and $l_{C,\nu}\left(\psi;y\right)=Z_{ik}S_{it}\nu_kA_{it}-Z_{ik}\log\left(1+e^{\nu_kA_{it}}\right)$.

To compute $Q\left(\psi;\psi^{(t)}\right)$ we need $E_{\psi^{(t)}}(Z_{ik}|Y_{it}=y_{it})$, which is calculated using the formula

(2.44). Now, we must compute $E_{\psi^{(t)}}(Z_{ik}S_{it}|Y_{it}=y_{it})$.

$$E_{\psi^{(t)}}(Z_{ik}S_{it}|Y_i) = P(Z_{ik}S_{it}=1|Y_i) \tag{3.261}$$

$$= P(Z_{ik}=1, S_{it}=1|Y_i) \tag{3.262}$$

$$= \frac{P(Y_i, S_{it}=1|Z_{ik}=1)P(Z_{ik}=1)}{P(Y_i)} \tag{3.263}$$

Clearly, for $t' \neq t$, we can observe that $Y_{it'}$ are independent given Z_{ik} to $(Y_{it}, S_{it}=1)$. We know that $(Y_{it})_{it}$ are independent conditionally to Z_{ik}, and $S_{it}=1$ does not provide any additional information about $Y_{it'}$, only about Y_{it}. So we can write:

$$E_{\psi^{(t)}}(Z_{ik}S_{it}|Y_i) = \frac{\prod_{\substack{t'=1\\t'\neq t}}^{T} P(Y_{it'}|Z_{ik}=1)P(Y_{it}, S_{it}=1|Z_{ik}=1)\pi_k}{\sum_{k=1}^{K} P(Y_i|Z_{ik}=1)P(Z_{ik}=1)} \tag{3.264}$$

$$= \frac{\prod_{\substack{t'=1\\t'\neq t}}^{T} P(Y_{it'}|Z_{ik}=1)P(Y_{it}, S_{it}=1|Z_{ik}=1)\pi_k}{\sum_{k=1}^{K} g_k(y_i,\beta_k,\delta_k,\nu_k)\pi_k} \tag{3.265}$$

$$= \frac{\prod_{\substack{t'=1\\t'\neq t}}^{T} P(Y_{it'}|Z_{ik}=1)P(Y_{it}|Z_{ik}=1, S_{it}=1)P(S_{it}=1|Z_{ik}=1)\pi_k}{\sum_{k=1}^{K} g_k(y_i,\beta_k,\delta_k,\nu_k)\pi_k} \tag{3.266}$$

$$= \frac{\prod_{t=1}^{T}[P(Y_{it}|Z_{ik}=1)]P(Y_{it}|Z_{ik}=1, S_{it}=1)P(S_{it}=1|Z_{ik}=1)\pi_k}{P(Y_{it}|Z_{ik}=1)\sum_{k=1}^{K} g_k(y_i,\beta_k,\delta_k,\nu_k)\pi_k} \tag{3.267}$$

$$= \frac{g_k(y_i,\beta_k,\delta_k,\nu_k)\pi_k P(Y_{it}|Z_{ik}=1, S_{it}=1)P(S_{it}=1|Z_{ik}=1)}{P(Y_{it}|Z_{ik}=1)\sum_{k=1}^{K} g_k(y_i,\beta_k,\delta_k,\nu_k)\pi_k} \tag{3.268}$$

$$= \frac{E_{\psi^{(t)}}(Z_{ik}|Y_i=y_i)P(Y_{it}|Z_{ik}=1, S_{it}=1)P(S_{it}=1|Z_{ik}=1)}{P(Y_{it}|Z_{ik}=1)} \tag{3.269}$$

When $(S_{it}=1, Z_{ik}=1)$, it implies that the value Y_{it} originates from a zero-excess state. If $Y_{it>0}$, we can be certain that this value originates from the Poisson state, as the zero-excess state only produces zero values, meaning $P(Y_{it}=y_{it}|S_{it}=1, Z_{ik}=1)=0$. On the other hand, if $Y_{it}=0$, this value could originate from either the Poisson state or the zero-excess state. We have:

$$P(Y_{it}=y_{it}|S_{it}=1, Z_{ik}=1) = \begin{cases} 0, \text{ if } y_{it}>0 \\ 1, \text{ if } y_{it}=0 \end{cases} \tag{3.270}$$

For $y_{it}>0$, $E_{\psi^{(t)}}(Z_{ik}S_{it}|Y_{it}=y_{it})=0$.

For $y_{it}=0$, we have $P(Y_{it}=y_{it}|Z_{ik}=1)=\rho_{ikt}+(1-\rho_{ikt})e^{-\lambda_{ikt}}$ and

$$E_{\psi^{(t)}}(Z_{ik}S_{it}|Y_{it}=y_{it}) = \frac{\rho_{ikt}}{\rho_{ikt}+(1-\rho_{ikt})e^{-\lambda_{ikt}}}E_{\psi^{(t)}}(Z_{ik}|Y_i) \tag{3.271}$$

$$= \frac{1}{1+e^{-\nu_k A_{it}-\lambda_{ikt}}}E_{\psi^{(t)}}(Z_{ik}|Y_i) \tag{3.272}$$

Finally, we have

$$E_{\psi^{(t)}}(Z_{ik}S_{it}|Y_{it}=y_{it}) = \begin{cases} 0 \text{ if } y_{it} > 0 \\ \frac{E_{\psi^{(t)}}(Z_{ik}|Y_i)}{1+e^{-\nu_k A_{it}-\lambda_{ikt}}} \text{ if } y_{it}=0 \end{cases} \tag{3.273}$$

Following the EM methods developed in section 2.3 we perform the two steps E and M:

- E step :

  Calculation of

$$E_{\psi^{(t)}}(Z_{ik}|Y_{it}=y_{it}) = \tau_{ik}^{(t)} = \frac{g_k(y_i,\beta_k^{(t)},\delta_k^{(t)},\nu_k^{(t)})\pi_k^{(t)}}{\sum_{k=1}^{K} g_k(y_i,\beta_k^{(t)},\delta_k^{(t)},\nu_k^{(t)})\pi_k^{(t)}} \tag{3.274}$$

$$E_{\psi^{(t)}}(Z_{ik}S_{it}|Y_{it}=y_{it}) = \tau_{ikt}^{(t)} = \begin{cases} 0 \text{ if } y_{it} > 0 \\ \frac{\tau_{ik}^{(t)}}{1+e^{-\nu_k^{(t)} A_{it}-\lambda_{ikt}^{(t)}}} \text{ if } y_{it}=0 \end{cases} \tag{3.275}$$

- M step :

  Calculate of $\psi^{(t+1)} = \arg\max_{\psi} \sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}^{(t)} \log\left(\pi_k g_k\left(y_i,\beta_k,\delta_k,\nu_k\right)\right)$ which is done by

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T}\tau_{ik}^{(t)}}{Tn} = \frac{\sum_{i=1}^{n}\tau_{ik}^{(t)}}{n} \quad 1 \leq k \leq K \tag{3.276}$$

$$\beta_k^{(t+1)} = \arg\max_{\beta_k} \sum_{i=1}^{n}\sum_{t=1}^{T}\left[(\tau_{ik}^{(t)}-\tau_{ikt}^{(t)})\left(Y_{it}\left(\beta_k A_{it}+\delta_k W_t\right) - e^{\beta_k A_{it}+\delta_k W_t} - \log\left(Y_{it}!\right)\right)\right] \tag{3.277}$$

$$\delta_k^{(t+1)} = \arg\max_{\delta_k} \sum_{i=1}^{n}\sum_{t=1}^{T}\left[(\tau_{ik}^{(t)}-\tau_{ikt}^{(t)})\left(Y_{it}\left(\beta_k A_{it}+\delta_k W_t\right) - e^{\beta_k A_{it}+\delta_k W_t} - \log\left(Y_{it}!\right)\right)\right] \tag{3.278}$$

$$\nu_k^{(t+1)} = \arg\max_{\nu_k} \sum_{i=1}^{n}\sum_{t=1}^{T} \tau_{ikt}^{(t)}\nu_k A_{it} - \tau_{ik}^{(t)}\log\left(1+e^{\nu_k A_{it}}\right) \tag{3.279}$$

We compute the partial derivatives of the functions above to find their roots. The derivative with respect to β_{kl}, where $1 \leq l \leq n_{\beta_k}$, is as follows:

$$\frac{\partial l_{C,\beta,\delta}\left(\psi;y\right)}{\partial \beta_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{l-1} Z_{ik}(1-S_{it})\left(Y_{it} - e^{\beta_k A_{it}+\delta_k W_t}\right) \tag{3.280}$$

The differential by δ_{kl} is, for $1 \leq l \leq n_{\delta}$,

$$\frac{\partial l_{C,\beta,\delta}\left(\psi;y\right)}{\partial \delta_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{l} Z_{ik}(1-S_{it})\left(Y_{it} - e^{\beta_k A_{it}+\delta_k W_t}\right) \tag{3.281}$$

The differential by ν_{kl} is, for $1 \leq l \leq n_{\nu}$,

$$\frac{\partial l_{C,\nu}\left(\psi;y\right)}{\partial \nu_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{l-1} Z_{ik}\left(S_{it} - \frac{e^{\nu_k A_{it}}}{1+e^{\nu_k A_{it}}}\right) \tag{3.282}$$

There are no closed-form solutions for β_k, δ_k or ν_k. Therefore, we search for approximations of these parameters using the quasi-Newton method, such as BFGS method.

3.3.5 Iteratively Reweighted Least Squares

We present a modified version of Iteratively Reweighted Least Squares (IRLS) to solve the two equations above, as discussed in Section 3.2.4.

We define the matrix

$$A_{\beta} = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{11}^{n_{\beta_k}-1} & \cdots & a_{1T}^{n_{\beta_k}-1} & \cdots & a_{n1}^{n_{\beta_k}-1} & \cdots & a_{nT}^{n_{\beta_k}-1} \end{pmatrix} \tag{3.283}$$

$$A_{\nu} = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ \vdots & & \vdots & & \vdots & & \vdots \\ a_{11}^{n_{\nu_k}-1} & \cdots & a_{1T}^{n_{\nu_k}-1} & \cdots & a_{n1}^{n_{\nu_k}-1} & \cdots & a_{nT}^{n_{\nu_k}-1} \end{pmatrix} \tag{3.284}$$

$$W = \begin{pmatrix} w_{11}^{1} & \cdots & w_{1T}^{1} & \cdots & w_{11}^{n_{\delta}} & \cdots & w_{1T}^{n_{\delta}} \\ \vdots & & \vdots & & \vdots & & \vdots \\ w_{n1}^{1} & \cdots & w_{nT}^{1} & \cdots & w_{n1}^{n_{\delta_k}} & \cdots & w_{nT}^{n_{\delta_k}} \end{pmatrix} \tag{3.285}$$

Z is a diagonal matrix with diagonal elements $\left(\underbrace{\tau_{1k},\cdots,\tau_{1k}}_{T},\cdots\underbrace{\tau_{nk},\cdots,\tau_{nk}}_{T}\right)$.

We define too,

$$s_{ikt} = \begin{cases} 0 \text{ if } y_{it} > 0 \\ \frac{1}{1+e^{-\nu_k A_{it} - \lambda_{ikt}}} \text{ if } y_{it} = 0 \end{cases} \tag{3.286}$$

3.3.5.1 Logistic part

In this case, $S(\nu_k)$ is a vector which elements l is $\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{l-1}\tau_{ik}\left(s_{ikt} - \frac{e^{\nu_k A_{it}}}{1+e^{\nu_k A_{it}}}\right)$ for $1 \leq l \leq n_{\nu_k}$. We write it like $\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{l-1}\tau_{ik}\left(s_{ikt} - \rho_{ikt}\right)$ and, for $1 \leq l' \leq n_{\nu_k}$, $\frac{\partial S(\nu_k)}{\partial \nu_{kl'}}$ is the vector which elements l is

$$\left(\frac{\partial S(\nu_k)}{\partial \nu_{kl'}}\right)_l = \sum_{i=1}^{n}\sum_{t=1}^{T} -\tau_{ik}a_{it}^{l} - 1a_{it}^{l'-1}\rho_{ikt}(1-\rho_{ikt})$$

$$W_\rho = \begin{pmatrix} \rho_{1k1}(1-\rho_{1k1}) & 0 & & \cdots & & 0 \\ 0 & \ddots & & & & \\ & & \rho_{1kT}(1-\rho_{1kT}) & & & \\ \vdots & & & \ddots & & \vdots \\ & & & \rho_{nk1}(1-\rho_{nk1}) & & \\ & & & & \ddots & 0 \\ 0 & & & \cdots & & \rho_{nkT}(1-\rho_{nkT}) \end{pmatrix} \tag{3.287}$$

$S = (s_{1k1}, \cdots, s_{1kT}, \cdots, s_{nk1}, \cdots, s_{nkT})^t$ and $P = (\rho_{1k1}, \cdots \rho_{1kT}, \cdots, \rho_{nk1}, \cdots, \rho_{nkT})^t$.

Thus, we can write

$$S(\nu_k) = A_\nu Z\,(S-P) \tag{3.288}$$

$$\frac{\partial S(\nu_k)}{\partial \nu_k} = -A_\nu Z W_\rho A_\nu^t \tag{3.289}$$

We substitute these quantities into equation (3.184) to obtain

$$\nu_k^{(t+1)} = \left(A_\nu Z^{(t)} W_\rho^{(t)} A_\nu^t\right)^{-1} A_\nu Z^{(t)} W_\rho^{(t)} \left(A_\nu^t \nu_k^{(t)} + W_\rho^{(t)^{-1}}\left(S^{(t)} - P^{(t)}\right)\right) \tag{3.290}$$

It can be observed that $A_\nu^t \nu_k^{(t)} + W_\rho^{(t)^{-1}}\left(S^{(t)} - P^{(t)}\right)$ is a vector whose elements correspond to the notations and indices mentioned earlier,

$$\nu_k^{(t)} A_{\nu,it} + \frac{s_{ikt}^{(t)} - \rho_{ikt}^{(t)}}{\rho_{ikt}^{(t)}(1-\rho_{ikt}^{(t)})} \tag{3.291}$$

3.3.5.2 Poisson part

In this case, $S(\beta_k)$ is a vector which elements l, $1 \leq l \leq n_\beta$, is $\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{(l-1)}\tau_{ik}(1-s_{ikt})\left(Y_{it} - e^{\beta_k A_{it}+\delta_k W_t}\right)$ for $1 \leq l \leq n_{\beta_k}$. We write it like $\sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{(l-1)}\tau_{ik}(1-s_{ikt})\left(Y_{it} - \lambda_{ikt}\right)$ and, for $1 \leq l' \leq n_{\beta_k}$,

$\frac{\partial S(\beta_k)}{\partial \beta_{kl'}}$ is the vector which elements l is

$$\left(\frac{\partial S(\beta_k)}{\partial \beta_{kl'}}\right)_l = \sum_{i=1}^{n}\sum_{t=1}^{T} -\tau_{ik}(1-s_{ikt})a_{it}^{l-1}a_{it}^{l'-1}\lambda_{ikt}$$

W_λ and S_p are diagonal matrix: $W_\lambda = diag\,(1-s_{1k1},\cdots,1-s_{1kT},\cdots,1-s_{nk1},\cdots,1-s_{nkT})$ and $S_\lambda = diag\,(\lambda_{1k1},\cdots\lambda_{1kT},\cdots,\lambda_{nk1},\cdots,\lambda_{nkT})$. Y is the matrix with all values Y_i, $Y = (Y_{11},\cdots Y_{1T},\cdots,Y_{n1},\cdots,Y_{nT})^t$ and $\Lambda = (\lambda_{1k1},\cdots\lambda_{1kT},\cdots,\lambda_{nk1},\cdots,\lambda_{nkT})^t$.

Thus we can write

$$S(\beta_k) = A_\beta Z S_\lambda\,(Y-\Lambda) \tag{3.292}$$

$$\frac{\partial S(\beta_k)}{\partial \beta_k} = -A_\beta S_\lambda Z W_\lambda A_\beta^t \tag{3.293}$$

We substitute these quantities into equation (3.184) to obtain

$$\beta_k^{(t+1)} = \left(A_\beta S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)} A_\beta^t\right)^{-1} A_\beta S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)} \left(A_\beta^t \beta_k^{(t)} + W_\lambda^{(t)^{-1}}\left(Y-\Lambda^{(t)}\right)\right) \tag{3.294}$$

It can be observed that $A_\beta^t\beta_k^{(t)} + W_p^{(t)^{-1}}\left(Y-\Lambda^{(t)}\right)$ is a vector whose elements correspond to the notations and indices mentioned earlier,

$$\beta_k^{(t)} A_{it} + \frac{Y_{it}}{\lambda_{ikt}^{(t)}} - 1 \tag{3.295}$$

In the same way, we have

$$S(\delta_k) = W Z S_\lambda\,(Y-\Lambda) \tag{3.296}$$

$$\frac{\partial S(\delta_k)}{\partial \delta_k} = -W S_\lambda Z W_\rho W^t \tag{3.297}$$

We substitute these quantities into equation (3.184) to obtain

$$\delta_k^{(t+1)} = \left(W S_\lambda^{(t)} Z^{(t)} W_\rho^{(t)} W^t\right)^{-1} W S_\lambda^{(t)} Z^{(t)} W_\rho^{(t)} \left(W^t \delta_k^{(t)} + W_\rho^{(t)^{-1}}\left(Y-\Lambda^{(t)}\right)\right) \tag{3.298}$$

It can be observed that $W^t\delta_k^{(t)} + W_\rho^{(t)^{-1}}\left(Y-\Lambda^{(t)}\right)$ is a vector whose elements correspond to the notations and indices mentioned earlier,

$$\delta_k^{(t)} W_{it} + \frac{Y_{it}}{\lambda_{ikt}^{(t)}} - 1 \tag{3.299}$$

3.3.6 Numerical method

To prevent issues related to singular matrices when attempting to compute $\left(A_\nu Z^{(t)} W_\rho^{(t)} A_\nu^t\right)^{-1}$, we can apply the QR method to determine $\nu_k^{(t+1)}$. As discussed in Section 2.2.3, $\nu_k^{(t+1)}$ must satisfy the following equation:

$$\left(A_\nu (Z^{(t)} W_\rho^{(t)})^{1/2} \left((Z^{(t)} W_\rho^{(t)})^{1/2}\right)^t A_\nu^t\right) \nu_k^{(t+1)} \tag{3.300}$$

$$= A_\nu (Z^{(t)} W\rho^{(t)})^{1/2} \left[(Z^{(t)} W_\rho^{(t)})^{1/2} \left(A_\nu^t \nu_k^{(t)} + W\rho^{(t)^{-1}} \left(S^{(t)} - P^{(t)}\right)\right)\right] \tag{3.301}$$

It's worth noting that $Z^{(t)} W_\rho^{(t)}$ is a diagonal matrix with positive terms. Consequently, $(Z^{(t)} W_\rho^{(t)})^{1/2}$ is a diagonal matrix containing the square roots of the diagonal elements of $Z^{(t)} W_\rho^{(t)}$.

Thus, we compute

$$E^{(t)} \leftarrow QR\left(\left(A_\nu (Z^{(t)} W\rho^{(t)})^{1/2}\right)^t\right)$$

$$Q^{(t)} \leftarrow QR.Q(E^{(t)})$$

$$R^{(t)} \leftarrow QR.R(E^{(t)})$$

$$\text{backsolve}\left(R^{(t)}, Q^{(t)^t} (Z^{(t)} W_\rho^{(t)})^{1/2} \left(A_\nu^t \nu_k^{(t)} + W\rho^{(t)^{-1}} \left(S^{(t)} - P^{(t)}\right)\right)\right)$$

To prevent issues related to singular matrices when attempting to compute $\left(A_\beta S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)} A_\beta^t\right)^{-1}$ we can apply the QR method to determine $\beta_k^{(t+1)}$.
As discussed in Section 2.2.3, $\beta_k^{(t+1)}$ must satisfy the following equation:

$$\left(A_\beta (S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)})^{1/2} \left((S_\lambda^{(t)} Z^{(t)} W_\lambda)^{1/2}\right)^t A_\beta^t\right) \beta_k^{(t+1)} \tag{3.302}$$

$$= A_\beta (S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)})^{1/2} \left[(S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)})^{1/2} \left(A_\beta^t \beta_k^{(t)} + W_\lambda^{(t)^{-1}} \left(Y - \Lambda^{(t)}\right)\right)\right] \tag{3.303}$$

It's worth noting that $Z^{(t)} W_\lambda^{(t)}$ is a diagonal matrix with positive terms. Consequently, $(Z^{(t)} W_\lambda^{(t)})^{1/2}$is a diagonal matrix containing the square roots of the diagonal elements of $Z^{(t)} W_\lambda^{(t)}$.

Thus we compute

$$E^{(t)} \leftarrow QR\left(\left(A_\beta (S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)})^{1/2}\right)^t\right)$$

$$Q^{(t)} \leftarrow QR.Q(E^{(t)})$$

$$R^{(t)} \leftarrow QR.R(E^{(t)})$$

$$\text{backsolve}\left(R^{(t)}, Q^{(t)^t} (S_\lambda^{(t)} Z^{(t)} W_\lambda^{(t)})^{1/2} \left(A_\beta^t \beta_k^{(t)} + W_\lambda^{(t)^{-1}} \left(Y - \Lambda^{(t)}\right)\right)\right)$$

To prevent issues related to singular matrices when attempting to compute $\left(W S_\lambda^{(t)} Z^{(t)} W_\rho^{(t)} W^t\right)^{-1}$

we can apply the QR method to determine $\delta_k^{(t+1)}$.
As discussed in Section 2.2.3, $\delta_k^{(t+1)}$ must satisfy the following equation:

$$\left(W(S_\lambda^{(t)}Z^{(t)}W_\rho^{(t)})^{1/2}\left((S_\lambda^{(t)}Z^{(t)}W_\rho^{(t)})^{1/2}\right)^t W^t\right)\delta_k^{(t+1)} \tag{3.304}$$

$$= W(S_\lambda^{(t)}Z^{(t)}W_\rho^{(t)})^{1/2}\left[(S_\lambda^{(t)}Z^{(t)}W_\rho^{(t)})^{1/2}\left(W^t\delta_k^{(t)} + W_\rho^{(t)^{-1}}\left(Y - \Lambda^{(t)}\right)\right)\right] \tag{3.305}$$

Thus, we compute

$$E^{(t)} \leftarrow QR\left(\left(W_\delta(S_\lambda^{(t)}Z^{(t)}W_\rho^{(t)})^{1/2}\right)^t\right)$$
$$Q^{(t)} \leftarrow QR.Q(E^{(t)})$$
$$R^{(t)} \leftarrow QR.R(E^{(t)})$$
$$\text{backsolve}\left(R^{(t)}, Q^{(t)^t}(S_\lambda^{(t)}Z^{(t)}W_\rho^{(t)})^{1/2}\left(W^t\delta_k^{(t)} + W_\rho^{(t)^{-1}}\left(Y - \Lambda^{(t)}\right)\right)\right)$$

3.3.7 Estimation of standard error

We use the method describe in subsection 2.3.4.

We compute the complete score function

$$S_C(\psi; y) = \left(\frac{\partial l_C(\psi; y)}{\partial \pi}, \frac{\partial l_C(\psi; y)}{\partial \beta}, \frac{\partial l_C(\psi; y)}{\partial \nu}, \frac{\partial l_C(\psi; y)}{\partial \delta}\right)^t$$

where $\frac{\partial l_C(\psi;y)}{\partial \pi} = \left(\frac{\partial l_C(\psi;y)}{\partial \pi_1}, \dots, \frac{\partial l_C(\psi;y)}{\partial \pi_K}\right)$, $\frac{\partial l_C(\psi;y)}{\partial \beta} = \left(\frac{\partial l_C(\psi;y)}{\partial \beta_{11}}, \dots, \frac{\partial l_C(\psi;y)}{\partial \beta_{Kn_\beta}}\right)$, $\frac{\partial l_C(\psi;y)}{\partial \nu} = \left(\frac{\partial l_C(\psi;y)}{\partial \nu_1}, \dots, \frac{\partial l_C(\psi;y)}{\partial \nu_K}\right)$ and $\frac{\partial l_C(\psi;y)}{\partial \delta} = \left(\frac{\partial l_C(\psi;y)}{\partial \delta_{11}}, \dots, \frac{\partial l_C(\psi;y)}{\partial \delta_{Kn_\delta}}\right)$.

It is important to recall that the sum of all the values in the sequence π is equal to 1. Using this fact, we can express π_K as $\sum_{k=1}^K \pi_k = 1$. Additionally, when j is not equal to K, the derivative of π_K with respect to π_j is equal to -1. Consequently, we can conclude that for all values of k from 1 to $K-1$, the following equation holds true:

$$\frac{\partial l_C(\psi; y)}{\partial \pi_k} = \sum_{i=1}^n \left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K}\right)$$

If we utilize equation 2.64 to forecast the likelihood of membership, with $1 \leq k \leq K$ and $1 \leq l \leq n_\theta$,

$$\frac{\partial l_C\left(\psi; y\right)}{\partial \theta_{kl}} = \sum_{i=1}^n x_{il}\left(Z_{ik} - \pi_{ik}\right)$$

For $1 \leq k \leq K$ and $1 \leq l \leq n_{\beta_k}$

$$\frac{\partial l_C\left(\psi; y\right)}{\partial \beta_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{l-1} Z_{ik}(1-S_{it})\left(Y_{it}-\lambda_{ikt}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T} p_{iklt}^{\beta} Z_{ik}(1-S_{it})$$

where $p_{iklt}^{\beta} = a_{it}^{l-1}\left(Y_{it}-\lambda_{ikt}\right)$.

For $1 \leq k \leq K$ and $1 \leq l \leq n_{\nu_k}$

$$\frac{\partial l_C\left(\psi; y\right)}{\partial \nu_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{l-1} Z_{ik}\left(S_{it}-\rho_{ikt}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T} p_{iklt}^{\nu} Z_{ik}$$

where $p_{iklt}^{\nu} = a_{it}^{l-1}\left(S_{it}-\rho_{ikt}\right)$.

For $1 \leq k \leq K$ and $1 \leq l \leq n_{\delta}$

$$\frac{\partial l_C\left(\psi; y\right)}{\partial \delta_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{l} Z_{ik}(1-S_{it})\left(Y_{it}-\lambda_{ikt}\right) = \sum_{i=1}^{n}\sum_{t=1}^{T} p_{iklt}^{\delta} Z_{ik}(1-S_{it})$$

where $p_{iklt}^{\delta} = w_{it}^{l}\left(Y_{it}-\lambda_{ikt}\right)$.

3.3.7.1 Computation of the negative second derivative matrix

The negative of the second derivative matrix of the complete likelihood is,

$$-B_C(y;\hat{\psi}) = -\begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \pi^2} & \frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \beta} & \frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \nu} & \frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \delta} \\ \frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \pi} & \frac{\partial^2 l_C(\psi;y)}{\partial \beta^2} & \frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \nu} & \frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \delta} \\ \frac{\partial^2 l_C(\psi;y)}{\partial \nu \partial \pi} & \frac{\partial^2 l_C(\psi;y)}{\partial \nu \partial \beta} & \frac{\partial^2 l_C(\psi;y)}{\partial \nu^2} & \frac{\partial^2 l_C(\psi;y)}{\partial \nu \partial \delta} \\ \frac{\partial^2 l_C(\psi;y)}{\partial \delta \partial \pi} & \frac{\partial^2 l_C(\psi;y)}{\partial \delta \partial \beta} & \frac{\partial^2 l_C(\psi;y)}{\partial \delta \partial \nu} & \frac{\partial^2 l_C(\psi;y)}{\partial \delta^2} \end{pmatrix}$$

The dimensions of this matrix can be determined by evaluating the expression: $\left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_{\delta}\right)+K-1\right) \times \left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_{\delta}+\right)+K-1\right)$.
If we are using a predictor for probability, we need to make a modification: we replace all instances of the derivative π with θ. Consequently, the dimensions of the matrix will be given by $\left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_{\delta}\right)+(K-1)n_{\theta}\right) \times \left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_{\delta}\right)+(K-1)n_{\theta}\right)$.

3.3.7.1.1 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \pi^2}$ or $\frac{\partial^2 l_C(\psi;y)}{\partial \theta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \pi^2}$ has for size $(K-1)\times(K-1)$.

For $1 \leq k, l \leq K-1$ is

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi^2} = \left(\frac{\partial^2 l_C(\psi; y)}{\partial \pi_k \partial \pi_l}\right)_{kl}$$

which elements are

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi_k \partial \pi_l} = \begin{cases} \sum_{i=1}^n -\left(\frac{Z_{ik}}{\pi_k^2} + \frac{Z_{iK}}{\pi_K^2}\right),\ k = l \\ \sum_{i=1}^n -\frac{Z_{iK}}{\pi_K^2},\ k \neq l \end{cases}$$

The second derivative matrix for $\frac{\partial^2 l_C(\psi; y)}{\partial \theta^2}$ has for size $Kn_\theta \times Kn_\theta$ and is composed by

$$\frac{\partial^2 l_C(\psi; y)}{\partial \theta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \theta_1 \partial \theta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial \theta_K \partial \theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \theta_K \partial \theta_K} \end{pmatrix}$$

which elements, for $1 \leq l, l' \leq n_\theta$, are

$$\left(\frac{\partial^2 l_C(\psi; y)}{\partial \theta_k \partial \theta_{k'}}\right)_{ll'} = \frac{\partial^2 l_C(\psi; y)}{\partial \theta_{kl} \partial \theta_{k'l'}} = \begin{cases} \sum_{i=1}^n -x_{il} x_{il'} \left(\frac{e^{\theta_k x_i}}{\sum_{k=1}^K e^{\theta_k x_i}}\right)\left(1 - \frac{e^{\theta_k x_i}}{\sum_{k=1}^K e^{\theta_k x_i}}\right),\ k = k' \\ \sum_{i=1}^n x_{il} x_{il'} \frac{e^{\theta_k x_i}}{\sum_{k=1}^K e^{\theta_k x_i}} \frac{e^{\theta_{k'} x_i}}{\sum_{k=1}^K e^{\theta_k x_i}},\ k \neq k' \end{cases}$$

The outcomes remain the same for both π and θ in the subsequent sections.

3.3.7.1.2 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \beta}$

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi \partial \beta} = 0$$

3.3.7.1.3 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \nu}$

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi \partial \nu} = 0$$

3.3.7.1.4 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \pi \partial \delta}$

$$\frac{\partial^2 l_C(\psi; y)}{\partial \pi \partial \delta} = 0$$

3.3.7.1.5 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \beta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \beta^2}$ has for size $\left(\sum_{k=1}^K n_{\beta_k}\right) \times \left(\sum_{k=1}^K n_{\beta_k}\right)$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial \beta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \beta_1 \partial \beta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial \beta_K \partial \beta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \beta_K \partial \beta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$, $1 \leq l \leq n_{\beta_k}$ and $1 \leq l' \leq n_{\beta_{k'}}$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial \beta_{k'} \partial \beta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial \beta_{k'l'} \partial \beta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T -Z_{ik}(1 - S_{it}) a_{it}^{l+l'-2} \lambda_{ikt}, \ k = k' \\ 0, \ k \neq k' \end{cases}$$

3.3.7.1.6 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \nu}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \nu}$ has for size $\left(\sum_{k=1}^K n_{\beta_k}\right) \times \left(\sum_{k=1}^K n_{\nu_k}\right)$.

$$\frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \nu} = 0$$

3.3.7.1.7 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \beta \delta}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \delta}$ has for size $\left(\sum_{k=1}^K n_{\beta_k}\right) \times K n_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial \beta \partial \delta} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \beta_1 \partial \delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \beta_1 \partial \delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial \beta_K \partial \delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \beta_K \partial \delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$, $1 \leq l \leq n_\beta$ and $1 \leq l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial \beta_{k'} \partial \delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial \beta_{k'l'} \partial \delta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T -Z_{ik}(1 - S_{it}) a_{it}^{l-1} w_{it}^{l'-1} \lambda_{ikt}, \ k = k' \\ 0, \ k \neq k' \end{cases}$$

3.3.7.1.8 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \nu^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \nu^2}$ has for size $K n_\nu \times K n_\nu$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial \nu^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \nu_1 \partial \nu_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \nu_1 \partial \nu_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial \nu_K \partial \nu_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \nu_K \partial \nu_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l, l' \leq n_\nu$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial \nu_{k'} \partial \nu_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial \nu_{k'l'} \partial \nu_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T -Z_{ik} a_{it}^{l+l'-2} \rho_{ikt}(1-\rho_{ikt}),\ k = k' \\ 0,\ k \neq k' \end{cases}$$

3.3.7.1.9 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \nu \partial \delta}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \nu \partial \delta}$ has for size $Kn_\nu \times Kn_\delta$.

$$\frac{\partial^2 l_C(\psi;y)}{\partial \nu \partial \delta} = 0$$

3.3.7.1.10 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial \delta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial \delta^2}$ has for size $Kn_\delta \times Kn_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial \delta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial \delta_1 \partial \delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \delta_1 \partial \delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial \delta_K \partial \delta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial \delta_K \partial \delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l, l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial \delta_{k'} \partial \delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial \delta_{k'l'} \partial \delta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T -Z_{ik}(1-S_{it}) w_{it}^{l+l'-2} \lambda_{ikt},\ k = k' \\ 0,\ k \neq k' \end{cases}$$

3.3.8 Computation of $cov\left(S_C(\hat{\psi};x)|X=x\right)$

The conditional matrix of the score vector is given by

$$I_{y/x}\left(y;\hat{\psi}\right)=\begin{pmatrix} cov\left(S_c(\pi)\right) & cov\left(S_c(\pi),S_c(\beta)\right) & cov\left(S_c(\pi),S_c(\nu)\right) & cov\left(S_c(\pi),S_c(\delta)\right)\\ cov\left(S_c(\beta),S_c(\pi)\right) & cov\left(S_c(\beta)\right) & cov\left(S_c(\beta),S_c(\nu)\right) & cov\left(S_c(\beta),S_c(\delta)\right)\\ cov\left(S_c(\nu),S_c(\pi)\right) & cov\left(S_c(\nu),S_c(\beta)\right) & cov\left(S_c(\nu)\right) & cov\left(S_c(\nu),S_c(\delta)\right)\\ cov\left(S_c(\delta),S_c(\pi)\right) & cov\left(S_c(\delta),S_c(\beta)\right) & cov\left(S_c(\delta),S_c(\nu)\right) & cov\left(S_c(\delta)\right) \end{pmatrix}$$

The dimensions of this matrix can be determined by evaluating the expression: $\left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_\delta\right)+K-1\right)\times\left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_\delta+\right)+K-1\right)$.
If we are using a predictor for probability, we need to make a modification: we replace all instances of the derivative π with θ. Consequently, the dimensions of the matrix will be given by $\left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_\delta\right)+(K-1)n_\theta\right)\times\left(\sum_{k=1}^{K}\left(n_{\beta_k}+n_{\nu_k}+n_\delta\right)+(K-1)n_\theta\right)$.

As a reminder (see proposition 2 page 45),

- $E(Z_{ik})=\tau_{ik}$
- $var\left(Z_{ik}\right)=\tau_{ik}(1-\tau_{ik})$
- $cov\left(Z_{ik},Z_{il}\right)=-\tau_{ik}\tau_{il}$ for $k\neq l$
- $cov\left(Z_{ik},Z_{jl}\right)=0$

We have to compute $cov\left(Z_{ik'},Z_{ik}S_{it}\right)$ and $cov\left(Z_{ik'}S_{it},Z_{ik}S_{it}\right)$.

Proposition 3. *Let* $1\leq k,k'\leq K$, $1\leq i\leq n$ *and* $1\leq t\leq T$,

$$cov\left(Z_{ik'},Z_{ik}S_{it}\right)=\begin{cases}-\tau_{ik'}\tau_{ik}s_{ikt},\ k\neq k'\\ \tau_{ik}s_{ikt}(1-\tau_{ik}),\ k=k'\end{cases}\tag{3.306}$$

$$cov\left(Z_{ik'}S_{it},Z_{ik}S_{it}\right)=\begin{cases}-\tau_{ik'}s_{ik't}\tau_{ik}s_{ikt},\ k\neq k'\\ \tau_{ik}s_{ikt}\left(1-\tau_{ik}s_{ikt}\right),\ k=k'\end{cases}\tag{3.307}$$

Proof. We observe that $cov\left(Z_{jk'},Z_{ik}S_{it}\right)=0$ for $i\neq j$ because $Z_{jk'}$ and $Z_{ik}S_{it}$ are independent. In other words, if we are aware of the group membership of individual j, it provides no information about individual i.

$$cov\left(Z_{ik'},Z_{ik}S_{it}\right)=E\left(Z_{ik'}Z_{ik}S_{it}\right)-E\left(Z_{ik'}\right)E\left(Z_{ik}S_{it}\right)\tag{3.308}$$

$$=P\left(Z_{ik'}Z_{ik}S_{it}=1\right)-\tau_{ik'}\tau_{ik}s_{ikt}\quad\text{see (3.275) and (3.273)}\tag{3.309}$$

$$=P\left(Z_{ik'}=1,Z_{ik}=1,S_{it}=1\right)-\tau_{ik'}\tau_{ik}s_{ikt}\tag{3.310}$$

If $k \neq k'$ the $Z_{ik'} = 1$ and $Z_{ik} = 1$ are not possible. Thus, $cov\left(Z_{ik'}, Z_{ik}S_{it}\right) = -\tau_{ik'}\tau_{ik}s_{ikt}$ and if $k = k'$ then $P\left(Z_{ik'} = 1, Z_{ik} = 1, S_{it} = 1\right) = P\left(Z_{ik} = 1, S_{it} = 1\right) = E\left(Z_{ik} = 1, S_{it} = 1\right) = \tau_{ik}s_{ikt}$.

Finally,

$$cov\left(Z_{ik'}, Z_{ik}S_{it}\right) = \begin{cases} -\tau_{ik'}\tau_{ik}s_{ikt},\ k \neq k' \\ \tau_{ik}s_{ikt}(1-\tau_{ik}),\ k = k' \end{cases} \tag{3.311}$$

$$cov\left(Z_{ik'}S_{it}, Z_{ik}S_{it}\right) = E\left(Z_{ik'}S_{it}Z_{ik}S_{it}\right) - E\left(Z_{ik'}S_{it}\right)E\left(Z_{ik}S_{it}\right) \tag{3.312}$$

$$= E\left(Z_{ik'}S_{it}Z_{ik}S_{it}\right) - \tau_{ik'}s_{ik't}\tau_{ik}s_{ikt} \tag{3.313}$$

If $k \neq k'$ then $E\left(Z_{ik'}S_{it}Z_{ik}S_{it}\right) = 0$ and if $k = k'$ then $E\left(Z_{ik'}S_{it}Z_{ik}S_{it}\right) = E\left(\left(Z_{ik}S_{it}\right)^2\right) = P\left(\left(Z_{ik}S_{it}\right)^2 = 1\right) = P\left(Z_{ik}S_{it} = 1\right) = E\left(Z_{ik}S_{it}\right) = \tau_{ik}s_{ikt}$.

Thus,

$$cov\left(Z_{ik'}S_{it}, Z_{ik}S_{it}\right) = \begin{cases} -\tau_{ik'}s_{ik't}\tau_{ik}s_{ikt},\ k \neq k' \\ \tau_{ik}s_{ikt}\left(1-\tau_{ik}s_{ikt}\right),\ k = k' \end{cases} \tag{3.314}$$

□

3.3.8.1 Matrix $cov\left(S_c(\pi)\right)$

The matrix as for dimension $(K-1)\times(K-1)$.

For a diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write $1 \leq k \leq K-1$

$$cov\left(S_c(\pi)\right)_{kk} = \sum_{i=1}^{n}\left(\frac{\tau_{ik}(1-\tau_{ik})}{\pi_k^2} + \frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2} - 2\frac{\tau_{ik}\tau_{iK}}{\pi_k\pi_K}\right)$$

where $1 \leq k \leq K-1$.

For a non-diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write

$$cov\left(S_c(\pi)\right)_{kl} = \sum_{i=1}^{n}\left(-\frac{\tau_{ik}\tau_{il}}{\pi_k\pi_l} + \frac{\tau_{ik}\tau_{iK}}{\pi_k\pi_K} + \frac{\tau_{iK}\tau_{il}}{\pi_K\pi_l} + \frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2}\right)$$

3.3.8.2 Matrix $cov\left(S_c(\pi), S_c(\beta)\right)$

$cov\left(S_c(\pi), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\beta_k)\right)$ for all groups k which dimension is $(K-1)\times n_{\beta_k}$. Thus, the dimension of the first matrix is $(K-1)\times\sum_{k=1}^{K} n_{\beta_k}$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\beta_k)\right)$ that is a matrix with elements, for $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{n} B^{\pi}_{ikl}\tau_{ik}\left(\frac{(1-\tau_{ik})}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right), k' = k \\ \sum_{i=1}^{n}\sum_{j=1}^{n} B^{\pi}_{ikl}\tau_{ik}\left(\frac{\tau_{ik'}}{\pi_{k'}} - \frac{\tau_{iK}}{\pi_K}\right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\beta_K)\right)_{k'l} = \sum_{i=1}^{n}\sum_{j=1}^{n} B^{\pi}_{iKl}\tau_{iK}\left(\frac{\tau_{ik'}}{\pi_{k'}} + \frac{1-\tau_{iK}}{\pi_K}\right)$$

where $B^{\pi}_{ikl} = \sum_{t=1}^{T} p^{\beta}_{iklt}\left(s_{ikt} - 1\right)$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$.

Proof. Let $1 \leq k, k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$,

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{k'l} = cov\left(\sum_{i=1}^{n}\left(\frac{Z_{ik'}}{\pi_{k'}} - \frac{Z_{iK}}{\pi_K}\right), \sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1-S_{it})p^{\beta}_{iklt}\right) \tag{3.315}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jk}(1-S_{jt})p^{\beta}_{jklt}\right) - cov\left(\frac{Z_{iK}}{\pi_K}, Z_{jk}(1-S_{jt})p^{\beta}_{jklt}\right) \tag{3.316}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \frac{p^{\beta}_{iklt}}{\pi_{k'}}\left(cov\left(Z_{ik'}, Z_{ik}\right) - cov\left(Z_{ik'}, Z_{ik}S_{it}\right)\right) \tag{3.317}$$

$$- \frac{p^{\beta}_{iklt}}{\pi_K}\left(cov\left(Z_{iK}, Z_{ik}\right) - cov\left(Z_{iK}, Z_{ik}S_{it}\right)\right) \tag{3.318}$$

if $k = k'$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{kl} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \frac{p^{\beta}_{iklt}}{\pi_k}\left(\tau_{ik}(1-\tau_{ik}) - \tau_{ik}s_{ikt}(1-\tau_{ik})\right) - \frac{p^{\beta}_{iklt}}{\pi_K}\left(-\tau_{iK}\tau_{ik} + \tau_{iK}\tau_{ik}s_{ikt}\right) \tag{3.319}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} p^{\beta}_{iklt}\left(1 - s_{ikt}\right)\tau_{ik}\left(\frac{(1-\tau_{ik})}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right) \tag{3.320}$$

if $k \neq k'$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{kl} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \frac{p^{\beta}_{iklt}}{\pi_{k'}}\left(-\tau_{ik}\tau_{ik'} + \tau_{ik'}\tau_{ik}s_{ikt}\right) - \frac{p^{\beta}_{iklt}}{\pi_K}\left(-\tau_{iK}\tau_{ik} + \tau_{iK}\tau_{ik}s_{ikt}\right) \tag{3.321}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} p_{iklt}^{\beta}\tau_{ik}\left(1 - s_{ikt}\right)\left(\frac{-\tau_{ik'}}{\pi_{k'}} + \frac{\tau_{iK}}{\pi_K}\right) \tag{3.322}$$

For $k = K$,

$$cov\left(S_c(\pi), S_c(\beta_K)\right)_{k'l} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jK}(1 - S_{jt})p_{jKt}^{\beta}\right) - cov\left(\frac{Z_{iK}}{\pi_K}, Z_{jK}(1 - S_{jt})p_{jKt}^{\beta}\right) \tag{3.323}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \frac{p_{iKt}^{\beta}}{\pi_{k'}}\left(cov\left(Z_{ik'}, Z_{iK}\right) - cov\left(Z_{ik'}, Z_{iK}S_{it}\right)\right) \tag{3.324}$$

$$- \frac{p_{iKt}^{\beta}}{\pi_K}\left(cov\left(Z_{iK}, Z_{iK}\right) - cov\left(Z_{iK}, Z_{iK}S_{it}\right)\right) \tag{3.325}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \frac{p_{iKt}^{\beta}}{\pi_{k'}}\left(-\tau_{ik'}\tau_{iK} + \tau_{ik'}\tau_{iK}s_{iKt}\right) - \frac{p_{iKt}^{\beta}}{\pi_K}\left(\tau_{iK}(1 - \tau_{iK}) - \tau_{iK}s_{iKt}(1 - \tau_{iK})\right) \tag{3.326}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} -p_{iKt}^{\beta}\tau_{iK}\left(1 - s_{iKt}\right)\left(\frac{\tau_{ik'}}{\pi_{k'}} + \frac{1 - \tau_{iK}}{\pi_K}\right) \tag{3.327}$$

□

3.3.8.3 Matrix $cov\left(S_c(\pi), S_c(\nu)\right)$

$cov\left(S_c(\pi), S_c(\nu)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\nu_k)\right)$ for all groups k which dimension is $(K-1) \times n_{\nu_k}$. Thus, the dimension of the first matrix is $(K-1) \times \sum_{k=1}^{K} n_{\nu_k}$.

Given $1 \le k \le K-1$ we compute $cov\left(S_c(\pi), S_c(\nu_k)\right)$ that is a matrix with elements, for $1 \le k' \le K-1$ and $1 \le l \le n_{\nu_k}$

$$cov\left(S_c(\pi), S_c(\nu_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^{n} N_{ikl}\tau_{ik}\left(\frac{1-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right), k' = k \\ \sum_{i=1}^{n} N_{ikl}\tau_{ik}\left(-\frac{\tau_{ik'}}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\nu_K)\right)_{k'l} = \sum_{i=1}^{n} -N_{iKl}\tau_{iK}\left(\frac{1 - \tau_{iK}}{\pi_K} + \frac{\tau_{ik'}}{\pi_k'}\right)$$

where $N_{ikl} = \sum_{t=1}^{T} a_{it}^{l-1}\left(s_{ikt} - \rho_{ikt}\right)$, $1 \le k' \le K-1$ and $1 \le l \le n_{\nu_k}$.

Proof. Let $1 \le k, k' \le K-1$ and $1 \le l \le n_{\beta_k}$,

$$cov\left(S_c(\pi), S_c(\nu_k)\right)_{k'l} = cov\left(\sum_{i=1}^{n}\left(\frac{Z_{ik'}}{\pi_{k'}} - \frac{Z_{iK}}{\pi_K}\right), \sum_{i=1}^{n} Z_{ik}\sum_{t=1}^{T} a_{it}^{l-1}\left(S_{it} - \rho_{ikt}\right)\right) \tag{3.328}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\left(cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jk}a_{jt}^{l-1}\left(S_{jt}-\rho_{jkt}\right)\right)\right. \tag{3.329}$$

$$\left.-cov\left(\frac{Z_{iK}}{\pi_{K}}, Z_{jk}a_{jt}^{l-1}\left(S_{jt}-\rho_{jkt}\right)\right)\right) \tag{3.330}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\left(\frac{a_{jt}^{l-1}}{\pi_{k'}}cov\left(Z_{ik'}, Z_{jk}S_{jt}\right)-\frac{\rho_{jkt}a_{jt}^{(l-1)}}{\pi_{k}'}cov\left(Z_{ik'}, Z_{jk}\right)\right. \tag{3.331}$$

$$\left.-\frac{a_{jt}^{l-1}}{\pi_{K}}cov\left(Z_{iK}, Z_{jk}S_{jt}\right)+\frac{\rho_{jkt}a_{jt}^{(l-1)}}{\pi_{K}}cov\left(Z_{iK}, Z_{jk}\right)\right) \tag{3.332}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T}\left(\frac{a_{it}^{l-1}}{\pi_{k'}}cov\left(Z_{ik'}, Z_{ik}S_{it}\right)-\frac{\rho_{ikt}a_{it}^{(l-1)}}{\pi_{k}'}cov\left(Z_{ik'}, Z_{ik}\right)\right. \tag{3.333}$$

$$\left.-\frac{a_{it}^{l-1}}{\pi_{K}}cov\left(Z_{iK}, Z_{ik}S_{it}\right)+\frac{\rho_{ikt}a_{it}^{(l-1)}}{\pi_{K}}cov\left(Z_{iK}, Z_{ik}\right)\right) \tag{3.334}$$

$$= \begin{cases} \sum_{i=1}^{n}\sum_{t=1}^{T}\left(a_{it}^{l-1}\tau_{ik}\left(s_{ikt}\left(\frac{\tau_{iK}}{\pi_K}-\frac{\tau_{ik'}}{\pi_k}\right)-\rho_{ikt}\left(\frac{\tau_{iK}}{\pi_K}-\frac{\tau_{ik'}}{\pi_k}\right)\right)\right), k' \neq k \\ \sum_{i=1}^{n}\sum_{t=1}^{T}\left(a_{it}^{l-1}\tau_{ik}\left(s_{ikt}\left(\frac{\tau_{iK}}{\pi_K}-\frac{\tau_{ik'}}{\pi_k}\right)-\rho_{ikt}\left(\frac{\tau_{iK}}{\pi_K}-\frac{1-\tau_{ik}}{\pi_k}\right)\right)\right), k' = k \end{cases} \tag{3.335}$$

$$= \begin{cases} \sum_{i=1}^{n}\sum_{t=1}^{T}\left(a_{it}^{l-1}\tau_{ik}\left(s_{ikt}-\rho_{ikt}\right)\left(\frac{\tau_{iK}}{\pi_K}-\frac{\tau_{ik'}}{\pi_k}\right)\right), k' \neq k \\ \sum_{i=1}^{n}\sum_{t=1}^{T}\left(a_{it}^{l-1}\tau_{ik}\left(s_{ikt}-\rho_{ikt}\right)\left(\frac{\tau_{iK}}{\pi_K}-\frac{1-\tau_{ik}}{\pi_k}\right)\right), k' = k \end{cases} \tag{3.336}$$

For $k = K$,

$$cov\left(S_c(\pi), S_c(\nu_K)\right)_{k'l} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T}\left(cov\left(\frac{Z_{ik'}}{\pi_{k'}}, Z_{jK}a_{it}^{l-1}\left(S_{jt}-\rho_{jKt}\right)\right)-cov\left(\frac{Z_{iK}}{\pi_{K}}, Z_{jK}a_{jt}^{l-1}\left(S_{jt}-\rho_{jKt}\right)\right)\right) \tag{3.337}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T}\left(-a_{it}^{l-1}\tau_{iK}\left(s_{iKt}-\rho_{iKt}\right)\left(\frac{1-\tau_{iK}}{\pi_K}+\frac{\tau_{ik'}}{\pi_k'}\right)\right) \tag{3.338}$$

□

3.3.8.4 Matrix $cov\left(S_c(\pi), S_c(\delta)\right)$

$cov\left(S_c(\pi), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\delta_k)\right)$ for all groups k which dimension is $(K-1) \times n_\delta$. Thus, the dimension of the first matrix is $(K-1) \times Kn_\delta$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\delta_k)\right)$ that is a matrix with elements, for $1 \leq k' \leq K-1$ and $1 \leq l \leq n_\delta$

$$cov\left(S_c(\pi), S_c(\delta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{n} D_{ikt}^{\pi}\tau_{ik}\left(\frac{(1-\tau_{ik})}{\pi_k}+\frac{\tau_{iK}}{\pi_K}\right), k' = k \\ \sum_{i=1}^{n}\sum_{j=1}^{n} D_{ikt}^{\pi}\tau_{ik}\left(\frac{\tau_{ik'}}{\pi_{k'}}-\frac{\tau_{iK}}{\pi_K}\right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\delta_K)\right)_{k'l} = \sum_{i=1}^{n}\sum_{j=1}^{n} D_{iKt}^{\pi}\tau_{iK}\left(\frac{\tau_{ik'}}{\pi_{k'}} + \frac{1-\tau_{iK}}{\pi_K}\right)$$

where $D_{ikt}^{\pi} = \sum_{t=1}^{T} p_{ikt}^{\delta}\left(s_{ikt}-1\right)$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\delta}$.

Proof. Same as β. □

3.3.8.5 Matrix $cov\left(S_c(\theta)\right)$ si on choist comme référence le gr k on enlève la mtrice cov thetak et on met thetak à 0

$cov\left(S_c(\theta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\theta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_\theta$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\theta$.

A diagonal matrix, for $1 \leq k \leq K$ and $1 \leq p, q \leq n_\theta$ is done by

$$\left(cov\left(S_c(\theta_k), S_c(\theta_k)\right)\right)_{pq} = \sum_{i=1}^{n} x_{ip}x_{iq}\tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$ and $1 \leq p, q \leq n_\theta$, is done by

$$\left(cov\left(S_c(\theta_k), S_c(\theta_l)\right)\right)_{pq} = -\sum_{i=1}^{n} x_{ip}x_{iq}\tau_{ik}\tau_{il}$$

See proof page 49.

3.3.8.6 Matrix $cov\left(S_c(\theta), S_c(\beta)\right)$

$cov\left(S_c(\theta), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\beta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_{\beta_l}$. Thus, the dimension of the first matrix is $Kn_\theta \times \sum_{l=1}^{K} n_{\beta_l}$.

A diagonal matrix, for $1 \leq k \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\beta_k}$ is done by

$$\left(cov\left(S_c(\theta_k), S_c(\beta_k)\right)\right)_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}p_{ikqt}^{\beta}\tau_{ik}(1-\tau_{ik})(1-s_{ikt})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\beta_l}$, is done by

$$\left(cov\left(S_c(\theta_k), S_c(\beta_l)\right)\right)_{pq} = -\sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}p_{ilqt}^{\beta}\tau_{ik}\tau_{il}(1+s_{ilt})$$

Proof. For a diagonal matrix, let $1 \leq k \leq K$ and $1 \leq q \leq n_{\beta_k}$ and $1 \leq p \leq n_\theta$

$$\begin{aligned}
\left(cov\left(S_c(\theta_k), S_c(\beta_k)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1 - S_{it})p^{\beta}_{ikqt}\right) && (3.339)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(x_{ip}Z_{ik}, Z_{jk}(1 - S_{jt})p^{\beta}_{jkqt}\right) && (3.340)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} cov\left(x_{ip}Z_{ik}, Z_{ik}(1 - S_{it})p^{\beta}_{ikqt}\right) && (3.341)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} \left(x_{ip}p^{\beta}_{ikqt}cov\left(Z_{ik}, Z_{ik}\right) - x_{ip}p^{\beta}_{ikqt}cov\left(Z_{ik}, Z_{ik}S_{it}\right)\right) && (3.342)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}p^{\beta}_{ikqt}\left(\tau_{ik}(1 - \tau_{ik}) - \tau_{ik}s_{ikt}(1 - \tau_{ik})\right) && (3.343)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}p^{\beta}_{ikqt}\left(\tau_{ik}(1 - \tau_{ik})(1 - s_{ikt})\right) && (3.344)
\end{aligned}$$

For a non diagonal matrix, let $1 \leq k, l \leq K$ and $1 \leq q \leq n_{\beta_l}$ and $1 \leq p \leq n_\theta$

$$\begin{aligned}
\left(cov\left(S_c(\theta_k), S_c(\beta_l)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik} - \pi_{ik}\right), \sum_{i=1}^{n}\sum_{t=1}^{T} Z_{il}(1 - S_{it})p^{\beta}_{ilqt}\right) && (3.345)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(x_{ip}Z_{ik}, Z_{jl}(1 - S_{jt})p^{\beta}_{jlqt}\right) && (3.346)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} cov\left(x_{ip}Z_{ik}, Z_{il}(1 - S_{it})p^{\beta}_{ilqt}\right) && (3.347)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} \left(x_{ip}p_{ilqt}cov\left(Z_{ik}, Z_{il}\right) - x_{ip}p^{\beta}_{ilqt}cov\left(Z_{ik}, Z_{il}S_{it}\right)\right) && (3.348)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}p^{\beta}_{ilqt}\left(-\tau_{ik}\tau_{il} - \tau_{ik}\tau_{il}s_{ilt}\right) && (3.349)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} -x_{ip}p^{\beta}_{ilqt}\tau_{ik}\tau_{il}(1 + s_{ilt}) && (3.350)
\end{aligned}$$

□

3.3.8.7 Matrix $cov\left(S_c(\theta), S_c(\delta)\right)$

$cov\left(S_c(\theta), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\delta_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_\delta$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\delta$.

A diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_\delta$ is done by

$$(cov\left(S_c(\theta_k), S_c(\beta_k)\right))_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}p^{\delta}_{ikqt}\tau_{ik}(1-\tau_{ik})(1-s_{ikt})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_\delta$, is done by

$$(cov\left(S_c(\theta_k), S_c(\beta_l)\right))_{pq} = -\sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}p^{\delta}_{ilqt}\tau_{ik}\tau_{il}(1+s_{ilt})$$

Proof. Same as above □

3.3.8.8 Matrix $cov\left(S_c(\theta), S_c(\nu)\right)$

$cov\left(S_c(\theta), S_c(\nu)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\nu_l)\right)$ for all groups k and l which dimension is $n_\theta \times n_{\nu_l}$. Thus, the dimension of the first matrix is $Kn_\theta \times \sum_{l=1}^{K} n_{\nu_l}$.

A diagonal matrix, for $1 \leq k \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\nu_k}$ is done by

$$(cov\left(S_c(\theta_k), S_c(\nu_k)\right))_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a^{q-1}_{it}\tau_{ik}(1-\tau_{ik})(s_{ikt}-\rho_{ikt})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\nu_l}$, is done by

$$(cov\left(S_c(\theta_k), S_c(\nu_l)\right))_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a^{q-1}_{it}\tau_{ik}\tau_{il}\left(\rho_{ilt}-s_{ilt}\right)$$

Proof. For a diagonal matrix, let $1 \leq k \leq K$ and $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\nu_k}$

$$\begin{aligned}
(cov\left(S_c(\theta_k), S_c(\nu_k)\right))_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik}-\pi_{ik}\right), \sum_{i=1}^{n} Z_{ik}\sum_{t=1}^{T} a^{q-1}_{it}\left(S_{it}-\rho_{ikt}\right)\right) && (3.351)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(x_{ip}\left(Z_{ik}-\pi_{ik}\right), Z_{jk}a^{q-1}_{jt}\left(S_{jt}-\rho_{jkt}\right)\right) && (3.352)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a^{q-1}_{it} cov\left(\left(Z_{ik}-\pi_{ik}\right), Z_{ik}\left(S_{it}-\rho_{ikt}\right)\right) && (3.353)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a^{q-1}_{it}\left(cov\left(Z_{ik}, Z_{ik}S_{it}\right) - \rho_{ikt} cov\left(Z_{ik}, Z_{ik}\right)\right) && (3.354)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a^{q-1}_{it}\left(\tau_{ik}s_{ikt}(1-\tau_{ik}) - \rho_{ikt}\tau_{ik}(1-\tau_{ik})\right) && (3.355)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a^{q-1}_{it}\tau_{ik}(1-\tau_{ik})(s_{ikt}-\rho_{ikt}) && (3.356)
\end{aligned}$$

For a non diagonal matrix, let $1 \leq k, l \leq K$ and $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\nu_l}$

$$
\begin{aligned}
\left(cov\left(S_c(\theta_k), S_c(\nu_l)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n} x_{ip}\left(Z_{ik}-\pi_{ik}\right), \sum_{i=1}^{n} Z_{il}\sum_{t=1}^{T} a_{it}^{q-1}\left(S_{it}-\rho_{ilt}\right)\right) && (3.357)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(x_{ip}\left(Z_{ik}-\pi_{ik}\right), Z_{jl}a_{jt}^{q-1}\left(S_{jt}-\rho_{jlt}\right)\right) && (3.358)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a_{it}^{q-1} cov\left(\left(Z_{ik}-\pi_{ik}\right), Z_{il}\left(S_{it}-\rho_{ilt}\right)\right) && (3.359)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a_{it}^{q-1}\left(cov\left(Z_{ik}, Z_{il}S_{it}\right)-\rho_{ilt}cov\left(Z_{ik}, Z_{il}\right)\right) && (3.360)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a_{it}^{q-1}\left(-\tau_{ik}\tau_{il}s_{ilt}+\rho_{ilt}\tau_{ik}\tau_{il}\right) && (3.361)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} x_{ip}a_{it}^{q-1}\tau_{ik}\tau_{il}\left(\rho_{ilt}-s_{ilt}\right) && (3.362)
\end{aligned}
$$

□

3.3.8.9 Matrix $cov\left(S_c(\beta)\right)$

$cov\left(S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\beta_l)\right)$ for all groups k and l which dimension is $n_{\beta_k} \times n_{\beta_l}$. Thus, the dimension of the first matrix is $\sum_{k=1}^{K} n_{\beta_k} \times \sum_{l=1}^{K} n_{\beta_l}$.

A diagonal matrix, for $1 \leq k \leq K$ and $1 \leq p, q \leq n_{\beta_k}$ is done by

$$
\left(cov\left(S_c(\beta_k), S_c(\beta_k)\right)\right)_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta}p_{ikqt}^{\beta}\tau_{ik}\left((1-\tau_{ik})(1-2s_{ikt})-s_{ikt}(1-\tau_{ik}s_{ikt})\right)
$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\beta_l}$, is done by

$$
\left(cov\left(S_c(\beta_k), S_c(\beta_l)\right)\right)_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta}p_{ilqt}^{\beta}\tau_{ik}\tau_{il}(s_{ikt}-1)(1-s_{ilt})
$$

Proof. For a diagonal matrix, let $1 \leq k \leq K$ and $1 \leq p, q \leq n_{\beta_k}$

$$
\begin{aligned}
\left(cov\left(S_c(\beta_k), S_c(\beta_k)\right)\right)_{pq} &= cov\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1-S_{it})p_{ikpt}^{\beta}, \sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1-S_{it})p_{ikqt}^{\beta}\right) && (3.363)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta}p_{ikqt}^{\beta} cov\left(Z_{ik}(1-S_{it}), Z_{ik}(1-S_{it})\right) && (3.364)
\end{aligned}
$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta} p_{ikqt}^{\beta} \left(\tau_{ik}(1-\tau_{ik}) - 2\tau_{ik}s_{ikt}(1-\tau_{ik}) - \tau_{ik}s_{ikt}(1-\tau_{ik}s_{ikt})\right) \tag{3.365}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta} p_{ikqt}^{\beta} \tau_{ik} \left((1-\tau_{ik})(1-2s_{ikt}) - s_{ikt}(1-\tau_{ik}s_{ikt})\right) \tag{3.366}$$

For a non diagonal matrix, let $1 \leq k, l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\beta_l}$

$$\left(cov\left(S_c(\beta_k), S_c(\beta_l)\right)\right)_{pq} = cov\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1-S_{it})p_{ikpt}^{\beta}, \sum_{i=1}^{n}\sum_{t=1}^{T} Z_{il}(1-S_{it})p_{ilqt}^{\beta}\right) \tag{3.367}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta} p_{ilqt}^{\beta} cov\left(Z_{ik}(1-S_{it}), Z_{il}(1-S_{it})\right) \tag{3.368}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta} p_{ilqt}^{\beta} \left(-\tau_{ik}\tau_{il} + \tau_{ik}\tau_{il}s_{ilt} + \tau_{il}\tau_{ik}s_{ikt} - \tau_{ik}s_{ikt}\tau_{il}s_{ilt}\right) \tag{3.369}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} p_{ikpt}^{\beta} p_{ilqt}^{\beta} \tau_{ik}\tau_{il}(s_{ikt}-1)(1-s_{ilt}) \tag{3.370}$$

□

3.3.8.10 Matrix $cov\left(S_c(\beta), S_c(\nu)\right)$

$cov\left(S_c(\beta), S_c(\nu)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\nu_l)\right)$ for $1 \leq k \leq K$ and $1 \leq l \leq K$ which dimension is $n_{\beta_k} \times n_{\nu_l}$. Thus, the dimension of the first matrix is $\sum_{k=1}^{K} n_{\beta_k} \times \sum_{l=1}^{K} n_{\nu_l}$.

An element of the matrix $cov\left(S_c(\beta_k), S_c(\nu_l)\right)$ for $1 \leq k \leq K$, $1 \leq l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\nu_l}$ is

$$cov\left(S_c(\beta_k), S_c(\nu_l)\right)_{pq} = \begin{cases} \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta} \tau_{ik}(s_{ikt}-1)(\tau_{ik}s_{ikt} + \rho_{ikt}(1-\tau_{ik})),\ k = l \\ \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta} \tau_{ik}\tau_{il}\left(s_{ikt}-1\right)\left(s_{ilt}-\rho_{ilt}\right),\ k \neq l \end{cases}$$

Proof. Let $1 \leq k \leq K$, $1 \leq l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\nu_l}$.
For $k = l$,

$$cov\left(S_c(\beta_k), S_c(\nu_k)\right)_{pq} = cov\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1-S_{it})p_{ikpt}^{\beta}, \sum_{i=1}^{n} Z_{ik}\sum_{t=1}^{T} a_{it}^{q-1}\left(S_{it}-\rho_{ikt}\right)\right) \tag{3.371}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} cov\left(Z_{ik}(1-S_{it})p_{ikpt}^{\beta}, Z_{ik}a_{it}^{q-1}\left(S_{it}-\rho_{ikt}\right)\right) \tag{3.372}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta} (cov\,(Z_{ik}S_{it}, Z_{ik}) - cov\,(Z_{ik}S_{it}, Z_{ik}S_{it}) - cov\,(\rho_{ikt}Z_{ik}, Z_{ik}) \tag{3.373}$$

$$+cov\,(\rho_{ikt}Z_{ik}, Z_{ik}S_{it})) \tag{3.374}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta} (\tau_{ik}s_{ikt}(1-\tau_{ik}) - \tau_{ik}s_{ikt}(1-\tau_{ik}s_{ikt}) - \rho_{ikt}\tau_{ik}(1-\tau_{ik}) \tag{3.375}$$

$$+\rho_{ikt}\tau_{ik}s_{ikt}(1-\tau_{ik})) \tag{3.376}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta}\tau_{ik}(s_{ikt}-1)(\tau_{ik}s_{ikt}+\rho_{ikt}(1-\tau_{ik})) \tag{3.377}$$

For $k \neq l$,

$$cov\,(S_c(\beta_k), S_c(\nu_l))_{pq} = cov\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1-S_{it})p_{ikpt}^{\beta}, \sum_{i=1}^{n} Z_{il}\sum_{t=1}^{T} a_{it}^{q-1}\,(S_{it}-\rho_{ilt})\right) \tag{3.378}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} cov\,\Big(Z_{ik}(1-S_{it})p_{ikpt}^{\beta}, Z_{il}a_{it}^{q-1}\,(S_{it}-\rho_{ilt})\Big) \tag{3.379}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta} (cov\,(Z_{il}S_{it}, Z_{ik}) - cov\,(Z_{il}S_{it}, Z_{ik}S_{it}) - cov\,(\rho_{ilt}Z_{il}, Z_{ik}) \tag{3.380}$$

$$+cov\,(\rho_{ilt}Z_{il}, Z_{ik}S_{it})) \tag{3.381}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta} (-\tau_{ik}\tau_{il}s_{ilt} + \tau_{il}s_{ilt}\tau_{ik}s_{ikt} + \rho_{ilt}\tau_{il}\tau_{ik} - \rho_{ilt}\tau_{il}\tau_{ik}s_{ikt}) \tag{3.382}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta}\tau_{ik}\tau_{il}\,(-s_{ilt} + s_{ilt}s_{ikt} + \rho_{ilt} - \rho_{ilt}s_{ikt}) \tag{3.383}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{q-1} p_{ikpt}^{\beta}\tau_{ik}\tau_{il}\,(s_{ikt}-1)\,(s_{ilt}-\rho_{ilt}) \tag{3.384}$$

□

3.3.8.11 Matrix $cov\,(S_c(\beta), S_c(\delta))$

$cov\,(S_c(\beta), S_c(\delta))$ is composed by the matrix $cov\,(S_c(\beta_k), S_c(\delta_l))$ for $1 \leq k \leq K$ and $1 \leq l \leq n_\delta$ which dimension is $n_{\beta_k} \times n_\delta$. Thus, the dimension of the first matrix is $\sum_{k=1}^{K} n_{\beta_k} \times Kn_\delta$.

A diagonal matrix, for $1 \leq k \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\delta_k}$ is done by

$$(cov\left(S_c(\beta_k), S_c(\delta_k)\right))_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} p^{\beta}_{ikpt} p^{\delta}_{ikqt} \tau_{ik}\left((1-\tau_{ik})(1-2s_{ikt}) - s_{ikt}(1-\tau_{ik}s_{ikt})\right)$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\delta_l}$, is done by

$$(cov\left(S_c(\beta_k), S_c(\delta_l)\right))_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} p^{\beta}_{ikpt} p^{\delta}_{ikqt} \tau_{ik}\tau_{il}(s_{ikt}-1)(1-s_{ilt})$$

3.3.8.12 Matrix $cov\left(S_c(\nu)\right)$

$cov\left(S_c(\nu)\right)$ is composed by the matrix $cov\left(S_c(\nu_k), S_c(\nu_l)\right)$ for $1 \leq k, l \leq K$ which dimension is $n_{\nu_k} \times n_{\nu_l}$. Thus, the dimension of the first matrix is $\sum_{k=1}^{K} n_{\nu_k} \times \sum_{l=1}^{K} n_{\nu_l}$.

A diagonal matrix, for $1 \leq k \leq K$ and $1 \leq p, q \leq n_{\nu_k}$ is done by

$$(cov\left(S_c(\nu_k), S_c(\nu_k)\right))_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1} a_{it}^{q-1} \tau_{ik}\left(s_{ikt}(1-\tau_{ik}s_{ikt}) + \rho_{ikt}^2(1-\tau_{ik})\right)$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_{\nu_k}$ and $1 \leq q \leq n_{\nu_l}$ is done by

$$(cov\left(S_c(\nu_k), S_c(\nu_l)\right))_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1} a_{it}^{q-1} \tau_{ik}\tau_{il}\left(-s_{ikt}s_{ilt} + \rho_{ilt}s_{ikt} + \rho_{ikt}s_{ilt} - \rho_{ikt}\rho_{ilt}\right)$$

Proof. For a diagonal matrix, let $1 \leq k \leq K$ and $1 \leq p, q \leq n_{\nu_k}$,

$$(cov\left(S_c(\nu_k), S_c(\nu_k)\right))_{pq} = cov\left(\sum_{i=1}^{n} Z_{ik}\sum_{t=1}^{T} a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), \sum_{i=1}^{n} Z_{ik}\sum_{t=1}^{T} a_{it}^{q-1}\left(S_{it}-\rho_{ikt}\right)\right) \tag{3.385}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(Z_{ik}a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), Z_{jk}a_{jt}^{q-1}\left(S_{jt}-\rho_{jkt}\right)\right) \tag{3.386}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1} var\left(Z_{ik}\left(S_{it}-\rho_{ikt}\right)\right) \tag{3.387}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1}\left(var\left(Z_{ik}S_{it}\right) + \rho_{ikt}^2 var\left(Z_{ik}\right)\right. \tag{3.388}$$

$$\left. + 2\rho_{ikt} cov\left(Z_{ik}S_{it}, Z_{ik}\right)\right) \tag{3.389}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1}\left(\tau_{ik}s_{ikt}(1-\tau_{ik}s_{ikt}) + \rho_{ikt}^2\tau_{ik}(1-\tau_{ik})\right. \tag{3.390}$$

$$\left. + 2\rho_{ikt}\tau_{ik}s_{ikt}(1-\tau_{ik})\right) \tag{3.391}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1}\tau_{ik}\left(s_{ikt}(1-\tau_{ik}s_{ikt}) + \rho_{ikt}^{2}(1-\tau_{ik}) + 2\rho_{ikt}s_{ikt}(1-\tau_{ik})\right) \tag{3.392}$$

For a non diagonal matrix, let $1 \leq k, l \leq K$, $1 \leq p \leq n_{\nu_k}$ and $1 \leq q \leq n_{\nu_l}$,

$$\left(cov\left(S_c(\nu_k), S_c(\nu_l)\right)\right)_{pq} = cov\left(\sum_{i=1}^{n} Z_{ik}\sum_{t=1}^{T} a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), \sum_{i=1}^{n} Z_{il}\sum_{t=1}^{T} a_{it}^{q-1}\left(S_{it}-\rho_{ilt}\right)\right) \tag{3.393}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} cov\left(Z_{ik}a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), Z_{jl}a_{jt}^{q-1}\left(S_{jt}-\rho_{jlt}\right)\right) \tag{3.394}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1} cov\left(Z_{ik}\left(S_{it}-\rho_{ikt}\right), Z_{jl}\left(S_{jt}-\rho_{jlt}\right)\right) \tag{3.395}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1}\left(cov\left(Z_{ik}S_{it}, Z_{il}S_{it}\right) - \rho_{ilt}cov\left(Z_{ik}S_{it}, Z_{il}\right)\right. \tag{3.396}$$

$$\left. -\rho_{ikt}cov\left(Z_{ik}, Z_{il}S_{it}\right) + \rho_{ikt}\rho_{ilt}cov\left(Z_{ik}, Z_{il}\right)\right) \tag{3.397}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1}\left(-\tau_{ik}s_{ikt}\tau_{il}s_{ilt} + \rho_{ilt}\tau_{il}\tau_{ik}s_{ikt} + \rho_{ikt}\tau_{ik}\tau_{il}s_{ilt} - \rho_{ikt}\rho_{ilt}\tau_{ik}\tau_{il}\right) \tag{3.398}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}a_{it}^{q-1}\tau_{ik}\tau_{il}\left(-s_{ikt}s_{ilt} + \rho_{ilt}s_{ikt} + \rho_{ikt}s_{ilt} - \rho_{ikt}\rho_{ilt}\right) \tag{3.399}$$

□

3.3.8.13 Matrix $cov\left(S_c(\nu), S_c(\delta)\right)$

$cov\left(S_c(\nu), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\nu_k), S_c(\delta_l)\right)$ for $1 \leq k \leq K$ and $1 \leq l \leq K$ which dimension is $n_{\nu_l} \times n_{\delta_l}$. Thus, the dimension of the first matrix is $\sum_{k=1}^{K} n_{\nu_k} \times \sum_{l=1}^{K} n_{\delta_l}$.

An element of the matrix $cov\left(S_c(\nu_k), S_c(\delta_l)\right)$ for $1 \leq k, l \leq K$, $1 \leq p \leq n_{\nu_k}$ and $1 \leq q \leq n_{\delta_l}$ is

$$cov\left(S_c(\nu_k), S_c(\delta_l)\right)_{pq} = \begin{cases} \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ikqt}^{\delta}\tau_{ik}(s_{ikt}-1)(\tau_{ik}s_{ikt}+\rho_{ikt}(1-\tau_{ik})), \ k = l \\ \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ilqt}^{\delta}\tau_{ik}\tau_{il}\left(s_{ilt}-1\right)\left(s_{ikt}-\rho_{ikt}\right), \ k \neq l \end{cases}$$

Proof. Let $1 \leq k \leq K$, $1 \leq l \leq K$, $1 \leq p \leq n_{\nu_k}$ and $1 \leq q \leq n_{\delta_l}$.

For $k = l$,

$$cov\left(S_c(\nu_k), S_c(\delta_k)\right)_{pq} = cov\left(\sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), \sum_{i=1}^{n}\sum_{t=1}^{T} Z_{ik}(1-S_{it})p_{ikqt}^{\delta}\right) \tag{3.400}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} cov\left(Z_{ik}a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), Z_{ik}(1-S_{it})p_{ikqt}^{\delta}\right) \tag{3.401}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ikqt}^{\delta}\left(cov\left(Z_{ik}S_{it}, Z_{ik}\right) - cov\left(Z_{ik}S_{it}, Z_{ik}S_{it}\right)\right. \tag{3.402}$$

$$\left.-cov\left(\rho_{ikt}Z_{ik}, Z_{ik}\right) + cov\left(\rho_{ikt}Z_{ik}, Z_{ik}S_{it}\right)\right) \tag{3.403}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ikqt}^{\delta}\left(\tau_{ik}s_{ikt}(1-\tau_{ik}) - \tau_{ik}s_{ikt}(1-\tau_{ik}s_{ikt}) - \rho_{ikt}\tau_{ik}(1-\tau_{ik})\right. \tag{3.404}$$

$$\left.+\rho_{ikt}\tau_{ik}s_{ikt}(1-\tau_{ik})\right) \tag{3.405}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ikqt}^{\delta}\tau_{ik}(s_{ikt}-1)(\tau_{ik}s_{ikt}+\rho_{ikt}(1-\tau_{ik})) \tag{3.406}$$

For $k \neq l$,

$$cov\left(S_c(\nu_k), S_c(\delta_l)\right)_{pq} = cov\left(\sum_{i=1}^{n} Z_{ik}\sum_{t=1}^{T} a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), \sum_{i=1}^{n}\sum_{t=1}^{T} Z_{il}(1-S_{it})p_{ilqt}^{\delta}\right) \tag{3.407}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} cov\left(Z_{ik}a_{it}^{p-1}\left(S_{it}-\rho_{ikt}\right), Z_{il}(1-S_{it})p_{ilqt}^{\delta}\right) \tag{3.408}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ilqt}^{\delta}\left(cov\left(Z_{ik}S_{it}, Z_{il}\right) - cov\left(Z_{ik}S_{it}, Z_{il}S_{it}\right) - cov\left(\rho_{ikt}Z_{ik}, Z_{il}\right)\right. \tag{3.409}$$

$$\left.+cov\left(\rho_{ikt}Z_{ik}, Z_{il}S_{it}\right)\right) \tag{3.410}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ilqt}^{\delta}\left(-\tau_{ik}\tau_{il}s_{ikt} + \tau_{il}s_{ilt}\tau_{ik}s_{ikt} + \rho_{ikt}\tau_{il}\tau_{ik} - \rho_{ikt}\tau_{il}\tau_{ik}s_{ilt}\right) \tag{3.411}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ilqt}^{\delta}\tau_{ik}\tau_{il}\left(-s_{ikt} + s_{ilt}s_{ikt} + \rho_{ikt} - \rho_{ikt}s_{ilt}\right) \tag{3.412}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} a_{it}^{p-1}p_{ilqt}^{\delta}\tau_{ik}\tau_{il}\left(s_{ilt}-1\right)\left(s_{ikt}-\rho_{ikt}\right) \tag{3.413}$$

□

3.3.8.14 Matrix $cov\left(S_c(\delta)\right)$

$cov\left(S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\delta_k), S_c(\delta_l)\right)$ $1 \leq k, l \leq K$ which dimension is $n_\delta \times n_\delta$. Thus, the dimension of the first matrix is $Kn_\delta \times Kn_\delta$.

A diagonal matrix, for $1 \leq p, q \leq n_\delta$ is done by

$$\left(cov\left(S_c(\delta_k), S_c(\delta_k)\right)\right)_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} p^{\delta}_{ikpt} p^{\delta}_{ikqt} \tau_{ik}\left((1-\tau_{ik})(1-2s_{ikt}) - s_{ikt}(1-\tau_{ik}s_{ikt})\right)$$

A non diagonal matrix, $1 \leq p, q \leq n_\delta$, is done by

$$\left(cov\left(S_c(\delta_k), S_c(\delta_l)\right)\right)_{pq} = \sum_{i=1}^{n}\sum_{t=1}^{T} p^{\delta}_{ikpt} p^{\delta}_{ikqt} \tau_{ik}\tau_{il}(s_{ikt}-1)(1-s_{ilt})$$

3.3.9 Numerical application

In the following sections, we will test the different equations by comparing them to the results obtained from the SAS procedure "traj". We conducted the same procedure in each example, where we constructed a sample consisting of a structure with K clusters, each containing 500 values of a ZIP variable Y. These variables follow a group-controlled pattern, achieved by linking the probability of the Poisson component to $\lambda_{ikt} = e^{\beta_k A_{it} + \delta_k W_{it}}$ and the probability of the zero excess state to $\rho_{ikt} = \frac{e^{\nu_k A_{it}}}{1+e^{\nu_k A_{it}}}$.

Therefore, we can compare the theoretical parameter values with those obtained from the SAS procedure "traj", the EM algorithm, EM IRLS, and the likelihood method described above.

3.3.9.1 Two groups

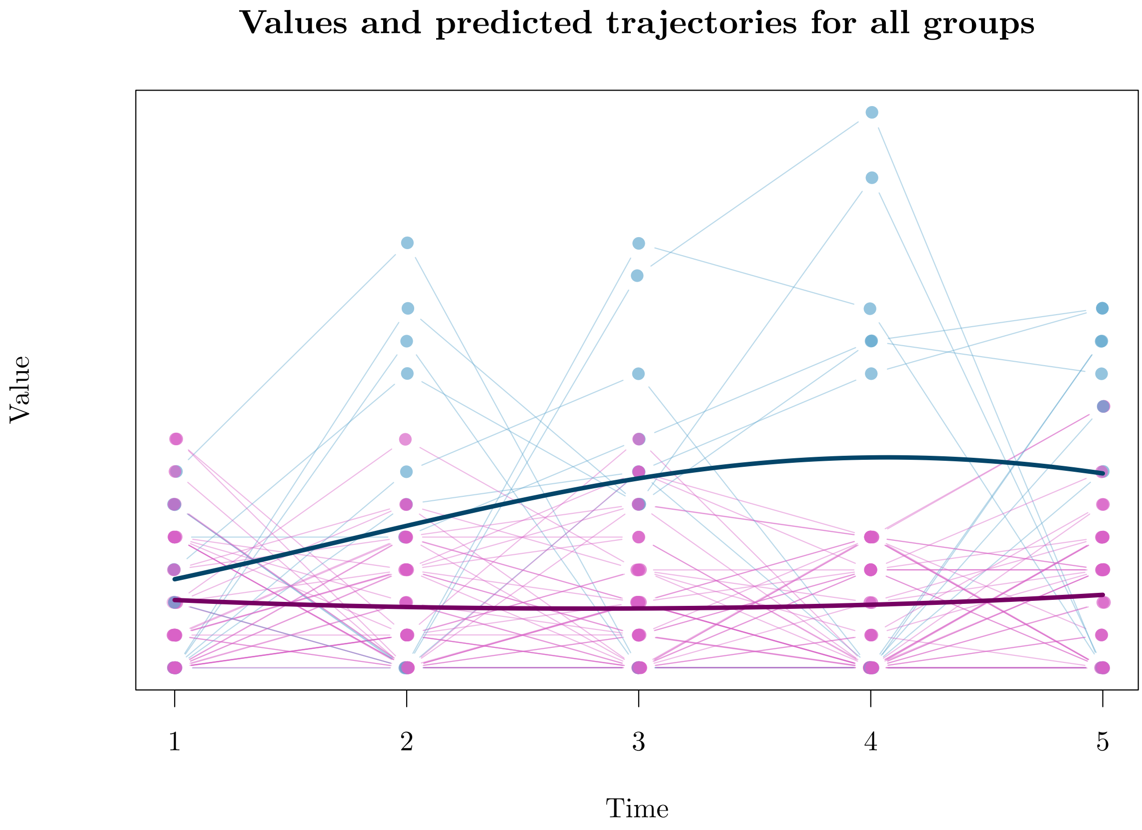
We set theoretical values as

	Polynomial		
Cluster	Degree	Shape	Probability π_k
1	2	$\beta_1 = (1.2, 0.5, -0.06)$	0.4
	1	$\nu_1 = (-0.2, -0.1)$	
2	2	$\beta_2 = (.89, 0.01, 0.01)$	0.6
	1	$\nu_2 = (-1, 0.01)$	

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms. The default evaluation of Traj yields the following values: $\beta_1 = (-0.91659, 0, 0)$, $\beta_2 = (0.41657, 0, 0)$, $\pi_1 = \pi_2 = 0.5$, $\nu_1 = (-3, 0)$ and $\nu_2 = (-3, 0)$. Consequently, we obtain the following results:

	Theoretical	Likelihood Traj		Likelihood		EM		EM - IRWLS	
			SE		SE		SE		SE
β_{11}	1.2	1.20813	0.27160	1.20813	0.38364	1.20853	0.22685	1.20813	0.22686
β_{12}	0.5	-0.21036	0.20755	-0.21036	0.25447	-0.21066	0.173	-0.21036	0.17301
β_{13}	-0.06	0.04368	0.03350	0.04368	0.03894	0.04373	0.02812	0.04368	0.02812
β_{21}	0.89	0.84365	0.38653	0.84365	0.26589	0.8422	0.31234	0.84365	0.31227
β_{22}	0.01	0.74046	0.25960	0.74046	0.20205	0.74125	0.20373	0.74046	0.20369
β_{23}	0.01	-0.09178	0.03979	-0.09178	0.03253	-0.09188	0.03094	-0.09178	0.03093
ν_{11}	-0.2	-1.11050	0.43066	-1.1105	0.65786	-1.11041	0.42564	-1.1105	0.42566
ν_{12}	-0.1	0.10653	0.12406	0.10653	0.19801	0.10651	0.123	0.10653	0.123
ν_{21}	-1	-0.42696	0.67909	-0.42696	0.38557	-0.42716	0.70395	-0.42696	0.70394
ν_{22}	0.01	-0.01780	0.20510	-0.0178	0.11375	-0.01775	0.21318	-0.0178	0.21318
π_1	0.4	0.77495	0.6239	0.77495	0.06064	0.77495	0.05977	0.77495	0.05977
π_2	0.6	0.22505	0.6239	0.22505	0.06064	0.22505	0.05977	0.22505	0.05977

We have created graphs that display the values and the trajectory shapes for all the groups.



3.3.9.2 Three groups

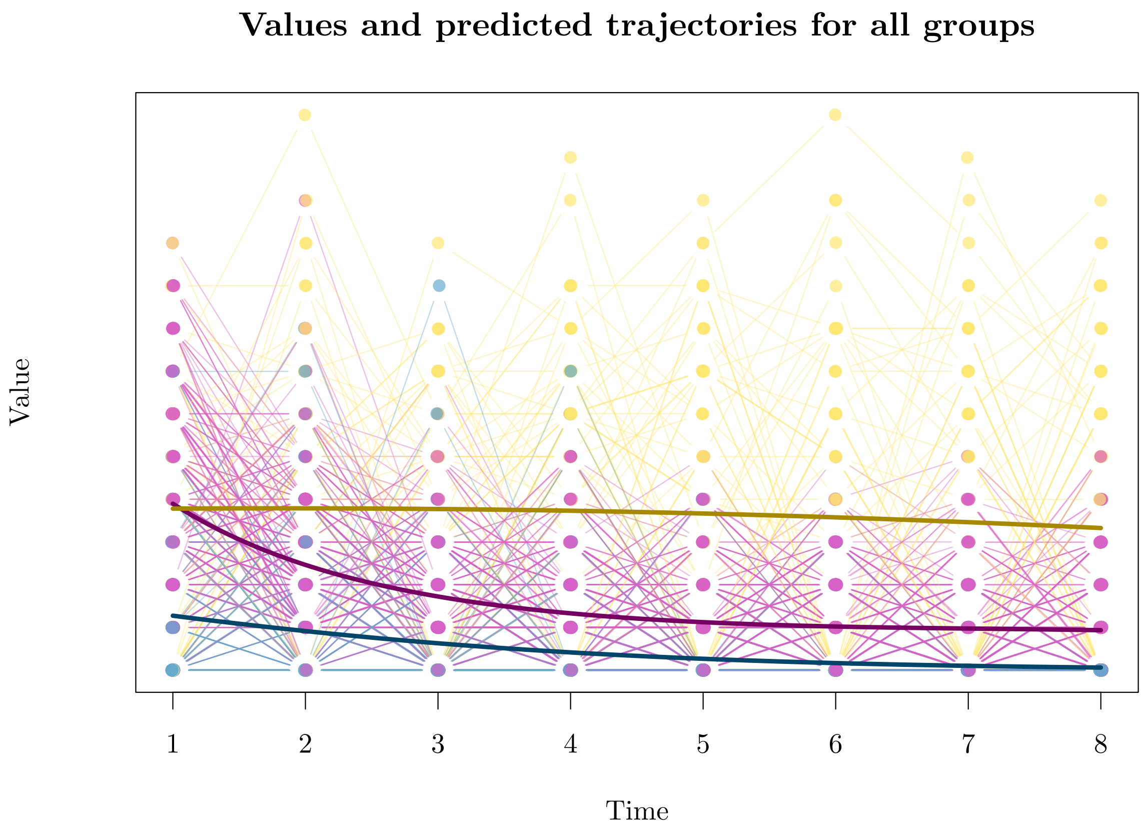
We set theoretical values as

	Polynomial		
Cluster	Degree	Shape	Probability π_k
1	2	$\beta_1 = (1.88998, -0.584769, 0.0492597)$	0.54
	1	$\nu_1 = (-10.4919, 1.2217)$	
2	2	$\beta_2 = (1.15384, -0.131227, -0.000409647)$	0.32
	1	$\nu_2 = (-0.57109, 0.429338)$	
3	1	$\beta_3 = (1.50031, 0.0536479)$	0.14
	1	$\nu_3 = (-1.32511, 0.161581)$	

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms. The default evaluation of Traj yields the following values: $\beta_1 = (-3, 0, 0)$, $\beta_2 = (-0.40104, 0, 0)$, $\beta_3 = (0.31986, 0)$, $\pi_1 = \pi_2 = \pi_3 = 1/3$, $\nu_1 = (-3, 0)$, $\nu_2 = (-3, 0)$ and $\nu_3 = (-3, 0)$. Consequently, we obtain the following results:

	Theoretical	Likelihood Traj		Likelihood		EM		EM - IRWLS	
			SE		SE		SE		SE
β_{11}	.88998	1.09270	0.16275	1.0927	0.16497	1.09266	0.11397	1.0927	0.11402
β_{12}	-0.584769	-0.09181	0.11294	-0.09181	0.11634	-0.09178	0.06943	-0.09181	0.06948
β_{13}	0.0492597	-0.00377	0.01522	-0.00377	0.01576	-0.00378	0.00837	-0.00377	0.00837
β_{21}	1.15384	1.92186	0.05409	1.92186	0.05479	1.92151	0.05191	1.92186	0.05192
β_{22}	-0.131227	-0.61193	0.03523	-0.61193	0.0358	-0.61167	0.03332	-0.61193	0.03333
β_{23}	-0.00041	0.05202	0.00431	0.05202	0.00436	0.05199	0.0039	0.05202	0.0039
β_{31}	1.50031	1.54134	0.05171	1.54134	0.05131	1.54135	0.04391	1.54134	0.04392
β_{32}	0.0536479	0.04431	0.01031	0.04431	0.01026	0.04431	0.00827	0.04431	0.00827
ν_{11}	-10.4919	-0.27347	0.17867	-0.27347	0.1825	-0.27339	0.17987	-0.27347	0.17985
ν_{12}	1.2217	0.39473	0.04675	0.39473	0.04707	0.39474	0.01876	0.39473	0.01876
ν_{21}	-0.57109	-6.96464	1.35652	-6.96466	1.35296	-6.95276	1.60232	-6.96465	1.61021
ν_{22}	0.429338	0.79069	0.17761	0.7907	0.17667	0.78914	0.21116	0.7907	0.21212
ν_{31}	-1.32511	-1.40935	0.22272	-1.40935	0.22773	-1.40935	0.23822	-1.40935	0.23822
ν_{32}	0.161581	0.17636	0.04208	0.17636	0.04195	0.17636	0.0467	0.17636	0.0467
π_1	0.54	0.3185	0.22896	0.3185	0.02284	0.31848	0.02203	0.3185	0.02203
π_2	0.32	0.54759	0.24124	0.54758	0.02428	0.5476	0.0233	0.54758	0.0233
π_3	0.14	0.13391	0.15764	0.13392	0.01587	0.13392	0.03207	0.13392	0.03207

We have created graphs that display the values and the trajectory shapes for all the groups.



3.4 Beta distribution

3.4.1 Definition

The Beta distribution serves as a valuable tool for dealing with percentages and proportions, relying on two parameters, namely α and β. Its chief strength lies in its ability to offer a versatile range of density shapes, rendering it less restrictive when compared to other distributions like the normal distribution. However, it's important to note that the Beta distribution is limited in its application, as it requires data to fall within the range of 0 to 1. It is not the case, we can use the

The Beta distribution's density can be represented using various parameterizations. In the context of regression, it's common to use the parametrization of the density based on the mean, as outlined in Ferrari and Cribari-Neto (2004). When dealing with a Beta random variable $0 < Y < 1$, we denote the parameters as μ and ϕ, such that

$$E(Y) = \mu \text{ and } V(Y) = \frac{V(\mu)}{1+\phi}$$

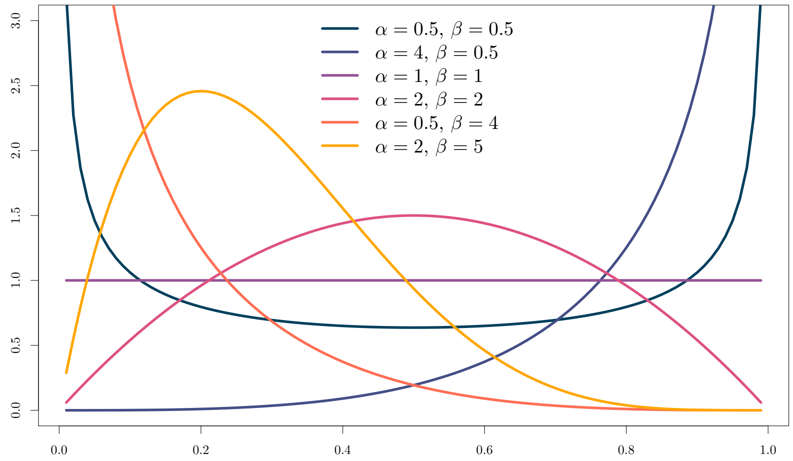


Figure 3.1: Distributions of Beta law.
Example of different shapes of the beta density for some parameters.

where $V(\mu) = \mu(1-\mu)$. The parameter ϕ can be interpreted as a precision parameter. For any μ, a larger value of ϕ results in a smaller $V(Y)$. Let f the density of Y

$$f(y;\mu;\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$$

where $0 < \mu < 1$ and $\phi > 0$.

Suppose that Y_{it} follow a beta distribution with the mean parameterization $\mathrm{Beta}(\mu_{ikt}, \phi_{ikt})$. The probability can be expressed as

$$P\left(Y_{it} = y_{it} | W_i = wi, C_i = k\right) = \frac{\Gamma(\phi_{ikt})}{\Gamma(\mu_{ikt}\phi_{ikt})\Gamma((1-\mu_{ikt})\phi_{ikt})} y_{it}^{\mu_{ikt}\phi_{ikt}-1}(1-y_{it})^{(1-\mu_{ikt})\phi_{ikt}-1} \tag{3.414}$$

In classical beta regression, we establish the relationship between the trajectory and the time variable using the following formula:

$$\mu_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \text{ and } \phi_{ikt} = e^{\zeta_k A_{\zeta,it}} \tag{3.415}$$

where $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1})^t$, $A_{\zeta,it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\zeta - 1})^t$, $W_{it} = (w_{i1}, \cdots, w_{in_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ et $\zeta_k = (\zeta_{k1}, \cdots, \zeta_{kn_\zeta})$. The formula for ϕ_{ikt} provides increased flexibility for modeling changes in the precision parameter over time.

Thus the log-likelihood 2.10 becomes

$$l(\psi; y) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \zeta_k)\right) \tag{3.416}$$

where

$$g_k(y_i;\beta_k,\delta_k,\zeta_k)=\prod_{t=1}^{T}\frac{\Gamma(\phi_{ikt})}{\Gamma(\mu_{ikt}\phi_{ikt})\Gamma((1-\mu_{ikt})\phi_{ikt})}y_{it}^{\mu_{ikt}\phi_{ikt}-1}(1-y_{it})^{(1-\mu_{ikt})\phi_{ikt}-1} \tag{3.417}$$

3.4.2 Likelihood

To estimate the parameters, we employ quasi-Newton methods, and we need to solve equations 2.16 and 2.17, which, in this specific case, become:

$$\frac{\partial l(\psi;y)}{\partial\theta_{kl}}=\sum_{i=1}^{n}\frac{\dfrac{\partial\pi_k}{\partial\theta_{kl}}g_k(y_i;\beta_k,\delta_k,\zeta_k)}{\displaystyle\sum_{k=1}^{K}\pi_k g_k(y_i;\beta_k,\delta_k,\zeta_k)}=0 \quad 1\le k\le K,\ \text{and}\ 1\le l\le n_\theta \tag{3.418}$$

$$\frac{\partial l(\psi;y)}{\partial\beta_{kl}}=\sum_{i=1}^{n}\frac{\pi_k\dfrac{\partial}{\partial\beta_{kl}}g_k(y_i;\beta_k,\delta_k,\zeta_k)}{\displaystyle\sum_{k=1}^{K}\pi_k g_k(y_i;\beta_k,\delta_k,\zeta_k)}=0 \quad 1\le k\le K,\ \text{and}\ 1\le l\le n_{\beta_k} \tag{3.419}$$

$$\frac{\partial l(\psi;y)}{\partial\delta_{kl}}=\sum_{i=1}^{n}\frac{\pi_k\dfrac{\partial}{\partial\delta_{kl}}g_k(y_i;\beta_k,\delta_k,\zeta_k)}{\displaystyle\sum_{k=1}^{K}\pi_k g_k(y_i;\beta_k,\delta_k,\zeta_k)}=0 \quad 1\le k\le K,\ \text{and}\ 1\le l\le n_\delta \tag{3.420}$$

$$\frac{\partial l(\psi;y)}{\partial\zeta_{kl}}=\sum_{i=1}^{n}\frac{\pi_k\dfrac{\partial}{\partial\zeta_{kl}}g_k(y_i;\beta_k,\delta_k,\zeta_k)}{\displaystyle\sum_{k=1}^{K}\pi_k g_k(y_i;\beta_k,\delta_k,\zeta_k)}=0 \quad 1\le k\le K,\ \text{and}\ 1\le l\le n_{\zeta_k} \tag{3.421}$$

When fitting the model using likelihood, the probability membership takes the form $\pi_k=\frac{e^{\theta_k x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}}$. However, it's important to note that there are no closed-form solutions for the equations we're about to calculate.

3.4.2.1 Differential by θ_k

Same as section 2.2.

3.4.2.2 Differential by β_{kl}

Let $1\le l\le n_{\beta_k}$, the derivative $\frac{\partial}{\partial\beta_{kl}}g_k(y_i;\beta_k,\delta_k,\zeta_k)$ is

$$\frac{\partial}{\partial\beta_{kl}}g_k(y_i;\beta_k,\delta_k,\zeta_k)=g_k(y_i;\beta_k,\delta_k,\zeta_k)\sum_{t=1}^{T}\frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{\left(1+e^{\beta_k A_{it}+\delta_k W_{it}}\right)^2}a_{it}^{l-1}\phi_{ikt}\left(y_{it}^*-\mu_{ikt}^*\right) \tag{3.422}$$

where $y^*_{it} = \log\left(\frac{y_{ikt}}{1-y_{ikt}}\right)$, $\mu^*_{ikt} = \psi(\mu_{ikt}\phi_{ikt}) - \psi((1-\mu_{ikt})\phi_{ikt})$, and $\psi(\cdot)$ is the digamma function define by $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)} = \dfrac{\int_0^\infty y^{z-1} \ln y \, \mathrm{e}^{-y} \, \mathrm{d}y}{\int_0^\infty y^{z-1} \mathrm{e}^{-y} \, \mathrm{d}y}$.

Proof. Following Ferrari and Cribari-Neto (2004), to find u' for some function u, we may compute $\log u$ since $u' = u \log u$.

Let $1 \le l \le n_{\beta_k}$, the differential of the logarithm of $g_k(y_{it}; \beta_k, \delta_k, \zeta_k) = \frac{\Gamma(\phi_{ikt})}{\Gamma(\mu_{ikt}\phi_{ikt})\Gamma((1-\mu_{ikt})\phi_{ikt})} y_{it}^{\mu_{ikt}\phi_{ikt}-1}(1-y_{it})^{(1-\mu_{ikt})\phi_{ikt}-1}$ (see page 8 for the definition of $g_k(y_{it}; \psi)$) is

$$\frac{\partial}{\partial \beta_{kl}} \log\left(g_k(y_{it}; \beta_k, \delta_k, \zeta_k)\right) = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{\left(1 + e^{\beta_k A_{it} + \delta_k W_{it}}\right)^2} a_{it}^{l-1} \phi_{ikt} \left(y^*_{it} - \mu^*_{ikt}\right) \tag{3.423}$$

By consequence

$$\frac{\partial}{\partial \beta_{kl}} g_k(y_{it}; \beta_k, \delta_k, \zeta_k) = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{\left(1 + e^{\beta_k A_{it} + \delta_k W_{it}}\right)^2} a_{it}^{l-1} \phi_{ikt} \left(y^*_{it} - \mu^*_{ikt}\right) g_k(y_{it}; \beta_k, \delta_k, \zeta_k) \tag{3.424}$$

Thus

$$\frac{\partial}{\partial \beta_{kl}} g_k(y_i; \beta_k, \delta_k, \zeta_k) = \sum_{t=1}^{T} \frac{\partial}{\partial \beta_{kl}} g_k(y_{it}; \beta_k, \delta_k, \zeta_k) \prod_{\substack{t'=1 \\ t' \neq t}}^{T} g_k(y_{it'}; \beta_k, \delta_k, \zeta_k) \tag{3.425}$$

$$= \sum_{t=1}^{T} \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{\left(1 + e^{\beta_k A_{it} + \delta_k W_{it}}\right)^2} a_{it}^{l-1} \phi_{ikt} \left(y^*_{it} - \mu^*_{ikt}\right) g_k(y_{it}; \beta_k, \delta_k, \zeta_k) \prod_{\substack{t'=1 \\ t' \neq t}}^{T} g_k(y_{it'}; \beta_k, \delta_k, \zeta_k) \tag{3.426}$$

$$= g_k(y_i; \beta_k, \delta_k, \zeta_k) \sum_{t=1}^{T} \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{\left(1 + e^{\beta_k A_{it} + \delta_k W_{it}}\right)^2} a_{it}^{l-1} \phi_{ikt} \left(y^*_{it} - \mu^*_{ikt}\right) \tag{3.427}$$

□

3.4.2.3 Differential by δ_{kl}

Let $1 \le l \le n_\delta$, the derivative $\frac{\partial}{\partial \delta_{kl}} g_k(y_i; \beta_k, \delta_k, \zeta_k)$ is

$$\frac{\partial}{\partial \delta_{kl}} g_k(y_i; \beta_k, \delta_k, \zeta_k) = g_k(y_i; \beta_k, \delta_k, \zeta_k) \sum_{t=1}^{T} \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{\left(1 + e^{\beta_k A_{it} + \delta_k W_{it}}\right)^2} w_{it}^{l} \phi_{ikt} \left(y^*_{it} - \mu^*_{ikt}\right) \tag{3.428}$$

where $y^*_{it} = \log\left(\frac{y_{ikt}}{1-y_{ikt}}\right)$, $\mu^*_{ikt} = \psi(\mu_{ikt}\phi_{ikt}) - \psi((1-\mu_{ikt})\phi_{ikt})$, and $\psi(\cdot)$ is the digamma function define by $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)} = \dfrac{\int_0^\infty y^{z-1} \ln y \, \mathrm{e}^{-y} \, \mathrm{d}y}{\int_0^\infty y^{z-1} \mathrm{e}^{-y} \, \mathrm{d}y}$.

Proof. Same as above □

3.4.2.4 Differential by ζ_{kl}

Let $1 \leq l \leq n_{\zeta_k}$, the derivative $\dfrac{\partial}{\partial \zeta_{kl}} g_k(y_i; \beta_k, \delta_k, \zeta_k)$ is

$$\frac{\partial}{\partial \zeta_{kl}} g_k(y_i; \beta_k, \delta_k, \zeta_k) \tag{3.429}$$

$$= g_k(y_i; \beta_k, \delta_k, \zeta_k) \sum_{t=1}^{T} \left(\phi_{ikt} a_{it}^{l-1} \left(\mu_{ikt} \left(y_{it}^* - \mu_{ikt}^* \right) + \psi(\phi_{ikt}) - \psi((1-\mu_{ikt})\phi_{ikt}) + \log(1-y_{it}) \right) \right) \tag{3.430}$$

where $y_{it}^* = \log \left(\frac{y_{ikt}}{1-y_{ikt}} \right)$, $\mu_{ikt}^* = \psi(\mu_{ikt}\phi_{ikt}) - \psi((1-\mu_{ikt})\phi_{ikt})$, and $\psi(\cdot)$ is the digamma function define by $\psi(z) = \frac{\Gamma'(z)}{\Gamma(z)} = \dfrac{\int_0^\infty y^{z-1} \ln y \, \mathrm{e}^{-y} \, \mathrm{d}y}{\int_0^\infty y^{z-1} \mathrm{e}^{-y} \, \mathrm{d}y}$.

Proof. Let $1 \leq l \leq n_{\zeta_k}$,

$$\frac{\partial}{\partial \zeta_{kl}} \log \left(g_k(y_{it}; \beta_k, \delta_k, \zeta_k) \right) = \phi_{ikt} a_{it}^{l-1} \left(\mu_{ikt} \left(y_{it}^* - \mu_{ikt}^* \right) + \psi(\phi_{ikt}) - \psi((1-\mu_{ikt})\phi_{ikt}) + \log(1-y_{it}) \right) \tag{3.431}$$

By consequence

$$\frac{\partial}{\partial \zeta_{kl}} g_k(y_{it}; \beta_k, \delta_k, \zeta_k) \tag{3.432}$$

$$= \left(\phi_{ikt} a_{it}^{l-1} \left(\mu_{ikt} \left(y_{it}^* - \mu_{ikt}^* \right) + \psi(\phi_{ikt}) - \psi((1-\mu_{ikt})\phi_{ikt}) + \log(1-y_{it}) \right) \right) g_k(y_{it}; \beta_k, \delta_k, \zeta_k) \tag{3.433}$$

Thus,

$$\frac{\partial}{\partial \zeta_{kl}} g_k(y_i; \beta_k, \delta_k, \zeta_k) \tag{3.434}$$

$$= g_k(y_i; \beta_k, \delta_k, \zeta_k) \sum_{t=1}^{T} \left(\phi_{ikt} a_{it}^{l-1} \left(\mu_{ikt} \left(y_{it}^* - \mu_{ikt}^* \right) + \psi(\phi_{ikt}) - \psi((1-\mu_{ikt})\phi_{ikt}) + \log(1-y_{it}) \right) \right) \tag{3.435}$$

□

3.4.3 Numerical application

We will test the different equations by comparing them to the results obtained from the SAS procedure traj. In each example, we have created a sample with a structure of K clusters, each containing 500 values of a Beta distribution using mean parametrization for variable Y. The trajectories follow a group-controlled pattern, achieved by linking the mean and dispersion parameters to the time variable as follows: $\mu_{ikt} = \frac{e^{\beta_k A_{it}+\delta_k W_{it}}}{1+e^{\beta_k A_{it}+\delta_k W_{it}}}$ and $\phi_{ikt} = e^{\zeta_k A_{\zeta,it}}$.

This allows us to compare the theoretical parameter values with the results obtained from the SAS procedure traj and the Likelihood method mentioned earlier.

3.4.3.1 Two groups

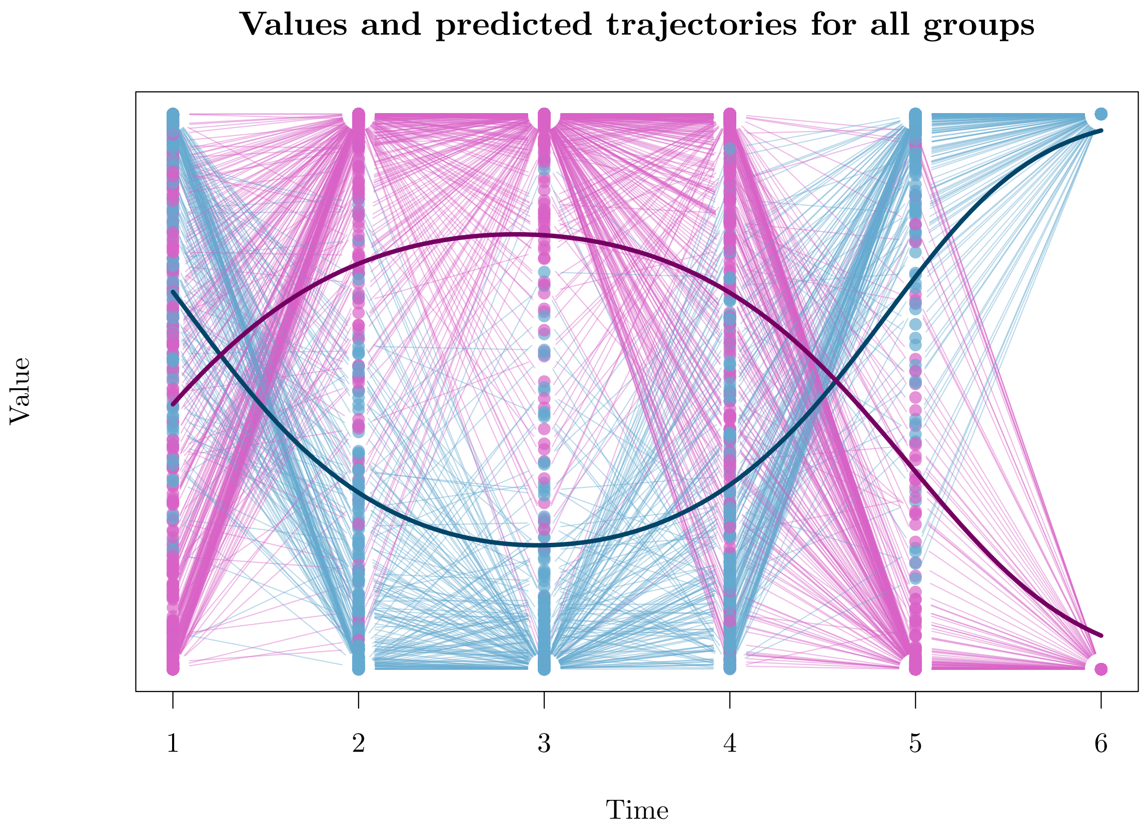
We set theoretical values as

	Polynomial		
Cluster	Degree	Shape	Probability π_k
1	2	$\beta_1 = (6.32, -5.8, 1)$	0.35
	1	$\zeta_1 = 1.5$	
2	2	$\beta_2 = (-6.69, 6.92, -1.23)$	0.65
	1	$\zeta_2 = 0.5$	

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms. The default evaluation of Traj yields the following values: $\beta_1 = (-0.69315, 0, 0)$, $\beta_2 = (0.69315, 0, 0)$, $\pi_1 = \pi_2 = 0.5$ and $\phi_1 = \phi_2 = \zeta_1 = \zeta_2 = 0.30125$. Consequently, we obtain the following results:

	Theoretical	Likelihood Traj		Likelihood	
			SE		SE
β_{11}	6.3.2	3.29783	0.17397	3.29825	0.17617
β_{12}	-5.8	-3.06134	0.12606	-3.06173	0.12821
β_{13}	1	0.51554	0.01918	0.51562	0.01948
β_{21}	-6.69	-1.98798	0.13714	-1.98786	0.13835
β_{22}	6.92	2.30117	0.09593	2.30129	0.09644
β_{23}	-1.23	-0.40448	0.01431	-0.4045	0.01435
ϕ_1	1.5	1.24648	0.05792	1.24648	0.05877
ϕ_2	0.5	0.54114	0.01695	0.54101	0.01707
π_1	0.35	0.35199	0.02139	0.35198	0.02141
π_2	0.65	0.64801	0.02139	0.64802	0.02141

We have created graphs that display the values and the trajectory shapes for all the groups.



3.4.3.2 Three groups

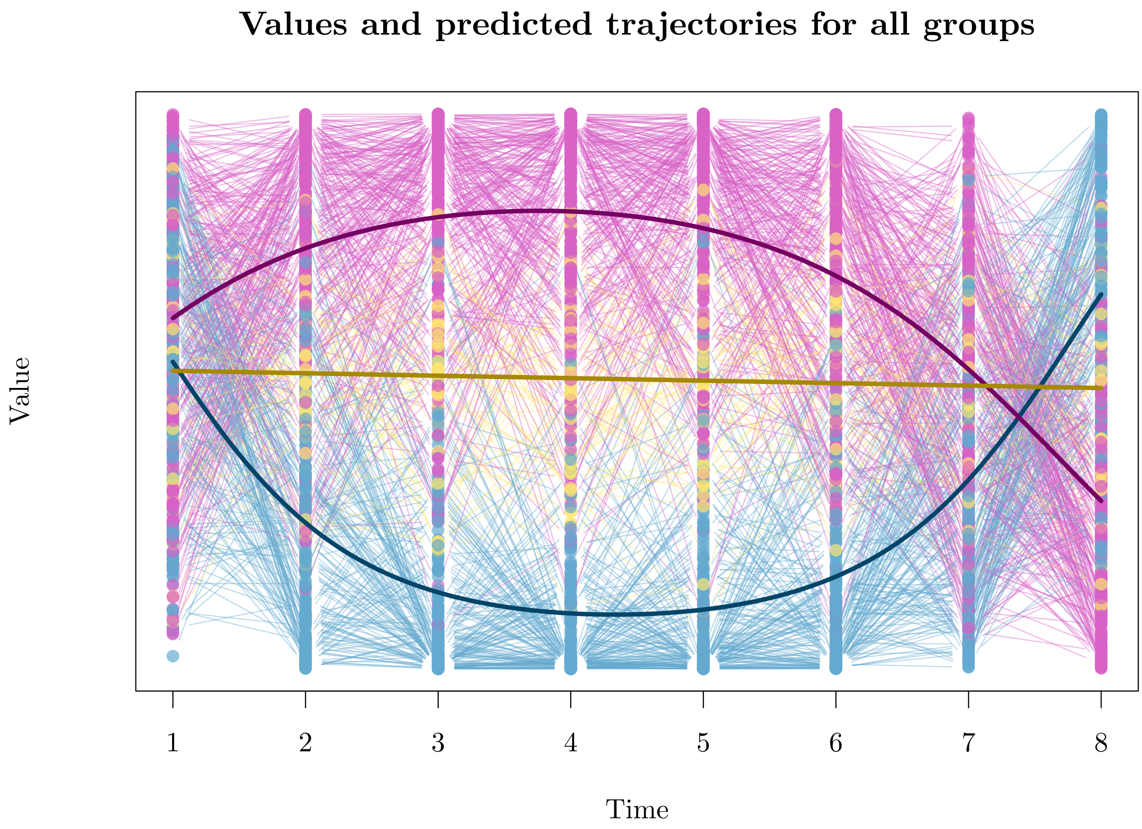
We set theoretical values as

	Polynomial		
Cluster	Degree	Shape	Probability π_k
1	2	$\beta_1 = (1.8833, -1.90201, 0.219884)$	0.25
	1	$\zeta_1 = 3$	
2	2	$\beta_2 = (-0.456576, 1.02901, -0.134615)$	0.54
	1	$\zeta_2 = 3$	
3	1	$\beta_3 = (0.279827, -0.0385803)$	0.21
	1	$\zeta_2 = 8$	

We have applied identical initial values to the Traj's Likelihood, Likelihood, and EM algorithms. The default evaluation of Traj yields the following values: $\beta_1 = (-1.09861, 0, 0)$, $\beta_3 = (0, 0, 0)$, $\beta_3 = (1.09861, 0)$, $\pi_1 = \pi_2 = \pi_3 = 1/3$ and $\phi_1 = \phi_2 = \phi_3 = \zeta_1 = \zeta_2 = \zeta_3 = 0.3539753$. Consequently, we obtain the following results:

	Theoretical	Likelihood Traj		Likelihood	
			SE		SE
β_{11}	1.8833	1.89622	0.10729	1.89621	0.10955
β_{12}	-1.90201	-1.90181	0.05960	-1.9018	0.06183
β_{13}	0.219884	0.21951	0.00654	0.21951	0.0068
β_{21}	-0.456576	-0.32613	0.08574	-0.32613	0.08908
β_{22}	1.02901	0.99940	0.04574	0.9994	0.04719
β_{23}	-0.134615	-0.13289	0.00506	-0.13289	0.00516
β_{31}	0.279827	0.16656	0.07724	0.16657	0.07776
β_{32}	-0.0385803	-0.01781	0.01507	-0.01781	0.01492
ϕ_1	3	3.01461	0.11416	3.01461	0.11636
ϕ_2	3	3.07932	0.09129	3.07932	0.09167
ϕ_3	8	7.33059	0.51445	7.33062	0.50200
π_1	0.25	0.35802	0.21577	0.35802	0.02141
π_2	0.54	0.51807	0.22964	0.51807	0.02317
π_3	0.21	0.12390	0.15801	0.1239	0.01558

We have created graphs that display the values and the trajectory shapes for all the groups.



3.5 Non linear mixture model

We assume that the variable Y_{it} is defined by

$$y_{it} = f(a_{it}; \beta_k, \delta_k) + \epsilon_{itk} \tag{3.436}$$

where $\epsilon_{itk} \sim \mathcal{N}(0; \sigma_k)$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$, $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$ and the function f is not linear in the parameters β_k.
In this case we have

$$E(Y_{it} = y_{it}|W_i = w_i, C_i = k) = f(a_{it}; \beta_k, \delta_k) \tag{3.437}$$

Assuming ϕ represents the density function of a standard normal distribution with mean 0 and standard deviation 1, and Φ represents its cumulative distribution function, the membership probability for a given group k can be expressed as follows:

$$P(Y_{it} = y_{it}|W_i = w_i, C_i = k) = \frac{1}{\sigma_k}\phi\left(\frac{y_{it} - f(a_{it}; \beta_k, \delta_k)}{\sigma_k}\right) \tag{3.438}$$

Thus the log-likelihood 2.10 becomes

$$l(\psi; y) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \sigma_k)\right) \tag{3.439}$$

where

$$g_k(y_i; \beta_k, \delta_k, \sigma_k) = \prod_{t=1}^{T} \frac{1}{\sigma_k}\phi\left(\frac{y_{it} - f(a_{it}; \beta_k, \delta_k)}{\sigma_k}\right) \tag{3.440}$$

3.5.1 Likelihood

The Newton method in likelihood optimization requires that the gradient of the function to maximize or minimize becomes zero at a certain value. In practice, this approach can be used effectively when the gradient can indeed be set to zero.

To fit the parameters, we utilize quasi-Newton methods, and the equations 2.16 and 2.17 become, in this specific case:

$$\frac{\partial l(\psi; y)}{\partial \theta_{kl}} = \sum_{i=1}^{n} \frac{\dfrac{\partial \pi_k}{\partial \theta_{kl}} g_k(y_i; \beta_k, \delta_k, \sigma_k)}{\displaystyle\sum_{k=1}^{K} \pi_k g_k(y_i; \beta_k, \delta_k, \sigma_k)} = 0 \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_\theta \tag{3.441}$$

$$\frac{\partial l(\psi;y)}{\partial \beta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \beta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i;\beta_k,\delta_k,\sigma_k)} = 0 \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_{\beta_k} \tag{3.442}$$

$$\frac{\partial l(\psi;y)}{\partial \delta_{kl}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \delta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i;\beta_k,\delta_k,\sigma_k)} = 0 \quad 1 \leq k \leq K, \text{ and } 1 \leq l \leq n_{\delta_k} \tag{3.443}$$

$$\frac{\partial l(\psi;y)}{\partial \sigma_{k}} = \sum_{i=1}^{n} \frac{\pi_k \frac{\partial}{\partial \delta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k)}{\sum_{k=1}^{K} \pi_k g_k(y_i;\beta_k,\delta_k,\sigma_k)} \quad 1 \leq k \leq K \tag{3.444}$$

In the scenario where we employ likelihood to estimate the model, the membership probability takes the form $\pi_k = \frac{e^{\theta_k x_i}}{\sum_{k=1}^{K} e^{\theta_k x_i}}$. We will compute this equation in several steps. It is essential to be aware that there are no closed-form solutions for this process.

3.5.1.1 Differential by θ_k

Same as section 2.2.

3.5.1.2 Differential by β_{kl}

Let $1 \leq l \leq n_{\beta_k}$, the derivative $\frac{\partial}{\partial \beta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k)$ is

$$\frac{\partial}{\partial \beta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k) = \tag{3.445}$$

$$\sum_{t=1}^{T} \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial \beta_{kl}} \frac{(y_{it} - f(a_{it};\beta_k,\delta_k))}{\sigma_k^3} \phi\left(\frac{y_{it} - f(a_{it};\beta_k,\delta_k)}{\sigma_k}\right) \prod_{\substack{t'=1,\\ t' \neq t}}^{T} \frac{1}{\sigma_k} \phi\left(\frac{y_{it'} - f(a_{it'};\beta_k,\delta_k)}{\sigma_k}\right) \tag{3.446}$$

3.5.1.3 Differential by δ_{kl}

Let $1 \leq l \leq n_{\delta_k}$, the derivative $\frac{\partial}{\partial \delta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k)$ is

$$\frac{\partial}{\partial \delta_{kl}} g_k(y_i;\beta_k,\delta_k,\sigma_k) = \tag{3.447}$$

$$\sum_{t=1}^{T} \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial \delta_{kl}} \frac{(y_{it} - f(a_{it};\beta_k,\delta_k))}{\sigma_k^3} \phi\left(\frac{y_{it} - f(a_{it};\beta_k,\delta_k)}{\sigma_k}\right) \prod_{\substack{t'=1,\\ t' \neq t}}^{T} \frac{1}{\sigma_k} \phi\left(\frac{y_{it'} - f(a_{it'};\beta_k,\delta_k)}{\sigma_k}\right) \tag{3.448}$$

3.5.1.4 Differential by σ_k

Let $1 \leq k \leq K$, the derivative $\frac{\partial}{\partial \sigma_k} g_k(y_i; \beta_k, \delta_k, \sigma_k)$ is

$$\frac{\partial}{\partial \sigma_k} g_k(y_i; \beta_k, \delta_k, \sigma_k) = \tag{3.449}$$

$$\sum_{t=1}^{T} \frac{(y_{it} - f(a_{it}; \beta_k, \delta_k))^2 - \sigma_k^2}{\sigma_k^4} \phi\left(\frac{y_{it} - f(a_{it}; \beta_k, \delta_k)}{\sigma_k}\right) \prod_{\substack{t'=1,\\ t' \neq t}}^{T} \frac{1}{\sigma_k} \phi\left(\frac{y_{it'} - f(a_{it'}; \beta_k, \delta_k)}{\sigma_k}\right) \tag{3.450}$$

3.5.2 EM algorithm

To apply the Likelihood method, we need to compute various differentials for each function, which can be a laborious process. To streamline this, we can utilize the EM method to estimate the parameters. In this case, we don't need to explicitly provide the differentials; except, we aim to obtain the standard errors. The complete likelihood is expressed by equation 2.38.

$$l_C(\psi; y) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k) \tag{3.451}$$

$$- \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\sum_{t=1}^{T} \log(\sigma_k) + \log(\sqrt{2\pi}) + \frac{1}{2} \left(\frac{y_{it} - f(a_{it}; \beta_k, \delta_k)}{\sigma_k} \right)^2 \right) \tag{3.452}$$

By noting

- Y the vector $(Y_{11}, \cdots, Y_{1T}, \cdots, Y_{n1}, \cdots, Y_{nT})^t$.
- F_k the vector $(f(a_{11}; \beta_k, \delta_k), \cdots, f(a_{1T}; \beta_k, \delta_k), \cdots, f(a_{n1}; \beta_k, \delta_k), \cdots, f(a_{nT}; \beta_k, \delta_k))^t$ for $1 \leq k \leq K$.
- Z_k the diagonal matrix with elements $\left(\underbrace{\tau_{1k}, \cdots, \tau_{1k}}_{T}, \cdots \underbrace{\tau_{nk}, \cdots, \tau_{nk}}_{T}\right)^t$ for $1 \leq k \leq K$.

We follow the EM method, as discussed in section 2.3, and compute the two steps: E (Expectation) and M (Maximization):

- E step :

  Calculation of $E_{\psi^{(t)}}(z_{ik} | Y_i = y_i) = \tau_{ik}^{(t)} = \dfrac{\pi_k^{(t)} g_k\left(y_i, \beta_k^{(t)}, \delta_k^{(t)}, \sigma_k^{(t)}\right)}{\sum_{k=1}^{K} \pi_k^{(t)} g_k\left(y_i, \beta_k^{(t)}, \delta_k^{(t)}, \sigma_k^{(t)}\right)}$

- M step :

Calculate of $\psi^{(t+1)} = \arg\max_{\psi} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \log\left(\pi_k g_k\left(y_i, \beta_k, \delta_k, \sigma_k\right)\right)$ which is done by

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(t)}}{n} \tag{3.453}$$

$$\beta_k^{(t+1)} = \arg\max_{\beta_k} \left\| Z_k^{1/2^{(t)}} \left(Y - F_k^{(t)}\right) \right\|^2 \tag{3.454}$$

$$\delta_k^{(t+1)} = \arg\max_{\delta_k} \left\| Z_k^{1/2^{(t)}} \left(Y - F_k^{(t)}\right) \right\|^2 \tag{3.455}$$

$$\sigma_k^{(t+1)} = \sqrt{\frac{\left\| Z_k^{1/2^{(t)}} \left(Y - F_k^{(t)}\right) \right\|^2}{T \left\| Z_k^{1/2^{(t)}} \right\|^2}} \tag{3.456}$$

3.5.2.1 Non linear regression geometry

In this context, ψ represents any parameters for a given cluster $1 \leq k \leq K$. Unlike linear regression, the expectation surface is not flat, so the estimated parameter $\hat{\psi}$ does not lie on a plane. To find $\psi^{(t+1)}$ in equation 3.454, we aim to find the value on the expectation surface that is nearest to Y in terms of a Weighted Euclidean distance.

Therefore, if some function f is twice differentiable, we can use Taylor-Lagrange approximation to express, within a neighborhood of x, the following relationship:

$$f(x + \Delta x) = f(x) + \left[\nabla f(x)\right]^t \Delta x + \frac{1}{2} \Delta x^t H(\tilde{x}) \Delta x, \tag{3.457}$$

where ∇f is the gradient of f, H is the Hessian of f and $\tilde{x} \in [0; \Delta x]$.
If Δx is close to 0,

$$f(x + \Delta x) \simeq f(x) + \left[\nabla f(x)\right]^t \Delta x$$

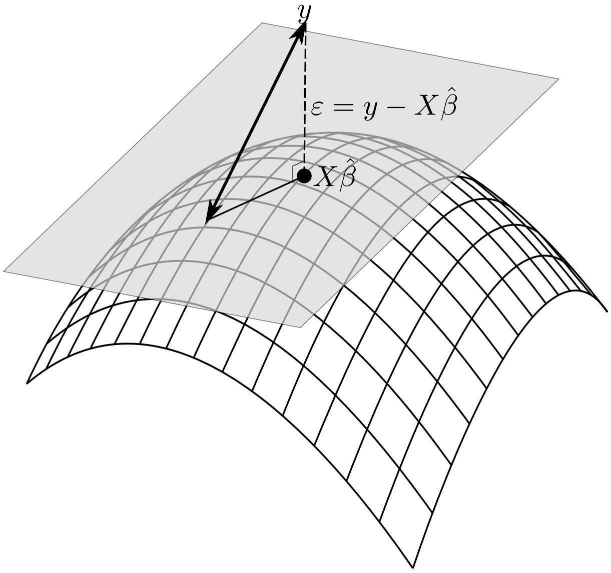
For values of x' that are close to x, we can approximate $f(x')$ as a linear function: $f(x') \approx f(x) + \left[\nabla f(x)\right]^T (x' - x)$.

In our case, which is a nonlinear model, we have $y_{it} = f(a_{it}; \beta_k, \delta_k) + \epsilon_{itk}$. This can be rewritten using vector notation as described in section 3.5 on page 123, where $Y = F + \epsilon = \eta(\psi) + \epsilon$. Here, $\eta(\psi)$ is an nT-vector with elements similar to those in F, and ϵ is a vector containing all the $\epsilon_{ik't'}$ values.

With the approximation mentioned earlier, we can write for ψ^* near to ψ:

$$\eta(\psi) = \eta(\psi^*) + V(\psi^*)(\psi - \psi^*) \tag{3.458}$$

Here, $V(\psi^*)$ represents a matrix containing the derivatives of $f(a_{it}; \beta_k, \delta_k)$ with respect to each parameter. We can replace the expectation surface $\eta(\psi)$ with the tangent plane and then find the point on this plane that minimizes the distance to the observed values Y. It's worth noting that if the expectation surface is very steep or non-linear, the tangent plane approximation may not be very accurate, potentially leading to a suboptimal approximation of the parameter values.



3.5.2.2 Gauss Newton method

The Gauss Newton method consist to

1. Obtaining a starting value ψ^0.
2. Using a linear approximation to $\eta(\psi)$ for ψ near to ψ^0.
3. Using linear regression methods to give the estimate of ψ, named ψ^1.
4. Repeat 2 and 3 until convergence.

At step (t), to find the parameter $\psi^{(t+1)}$ we use equation (3.458). Let $V^{(t)} = V(\psi^{(t)})$. For $\psi^{(t)}$ near to ψ, we have

$$\eta(\psi) = \eta(\psi^{(t)}) + V^{(t)}(\psi - \psi^{(t)})$$

It follows that

$$Y - \eta(\psi) = Y - \eta(\psi^{(t)}) - V^{(t)}(\psi - \psi^{(t)}) = U^{(t)} - V^{(t)}d^{(t)}$$

where $U^{(t)} = Y - \eta(\psi^{(t)})$ and $d^{(t)} = \psi - \psi^{(t)}$.
So, choosing ψ to minimize $S(\psi) = \left\| Z_k^{1/2^{(t)}} (Y - F) \right\|^2$ is approximately equivalent to choosing $d^{(t)}$ to minimize $\left\| Z_k^{1/2^{(t)}} \left(U^{(t)} - V^{(t)}d^{(t)}\right) \right\|^2$ which is a linear regression problem.

Thus the value of $d^{(t)}$ that minimize the expression above is

$$\hat{d}^{(t)} = \left[V^{(t)^t} Z_k^{\frac{1}{2}} V^{(t)}\right]^{-1} V^{(t)^t} Z_k^{\frac{1}{2}} U^{(t)}$$

This calculate can be estimated by using QR decomposition to avoid the search of an inverse. Next, we can obtained the update $\psi^{(t+1)}$,

$$\psi^{(t+1)} = \psi^{(t)} + \hat{d}^{(t)}$$

With this approach, there's no guarantee that the updated parameters will lead to a reduction in the objective function. If it turns out that after the update, we have $S(\psi^{(t+1)}) > S(\psi^{(t)})$, we take a different approach. We modify the update as follows: $\psi^{(t+1)} = \psi^{(t)} + \lambda \hat{d}_k^{(t)}$, where $0 < \lambda \leq 1$. We start with $\lambda = 1$, and if it doesn't result in a lower objective function, we continue to try decreasing values of λ such as $\frac{1}{2}$, $\frac{1}{4}$, and so on, until we find a λ value that leads to $S(\psi^{(t+1)}) < S(\psi^{(t)})$. This process helps us adjust the step size to improve the optimization process.

3.5.2.3 Levenberg Marquardt Method

In Gauss Newton method, at step (t), we have to solve $\left(V^{(t)^t} Z^{\frac{1}{2}} V^{(t)}\right) \hat{\delta} = V^{(t)^t} Z^{\frac{1}{2}} U^{(t)}$ called normal equation.

The inversion of this matrix depend on the condition number of the matrix, the ratio of the maximal eigenvalues in norm and the minimal one. Levenberg's contribution is to replace this equation by a "damped" version :

$$\left(V^{(t)^t} Z^{\frac{1}{2}} V^{(t)} + \lambda I\right) \hat{\delta} = V^{(t)^t} Z^{\frac{1}{2}} U^{(t)}$$

where I is the identity matrix. The non-negative damping factor λ is adjusted at each iteration. Small value of λ result in Gauss-Newton update and large value result in gradient descent update. λ is initialized to be large such that first updates are small step in the steepest-descent direction. In any iteration happens to result in a worse approximation, the parameter is increased. Otherwise, as the solution improve, λ is decreased, Lourakis et al. (2005), Madsen et al. (2004), Marquardt (1963). When the damping factor λ is large relative to $\left\|V^{(t)^t} Z^{\frac{1}{2}} V^{(t)}\right\|$, inverting $V^{(t)^t} Z^{\frac{1}{2}} V^{(t)} + \lambda I$ is not necessary, as the update is well-approximated by the small gradient step $\lambda^{-1} V^{(t)^t} Z^{\frac{1}{2}} U^{(t)}$.

In Marquardt's update relationship, see Marquardt (1963), the damping parameter λ is scaled by the diagonal of $V^{(t)^t} Z^{\frac{1}{2}} V^{(t)}$ for each parameter.

$$\left(V^{(t)^t} Z^{\frac{1}{2}} V^{(t)} + \lambda \text{diag}\left(V^{(t)^t} Z^{\frac{1}{2}} V^{(t)}\right)\right) \hat{d} = V^{(t)^t} Z^{\frac{1}{2}} U^{(t)}$$

the values of λ are normalized to the values of $V^{(t)^t} Z^{\frac{1}{2}} V^{(t)}$ and it make the solution scale invariant.

For information about numerical implementation we can see Gavin (2022) for example.

3.5.3 Estimation of standard error

We use the method describe in subsection 2.3.4.

We compute the complete score function

$$S_C(\psi; y) = \left(\frac{\partial l_C(\psi; y)}{\partial \pi}, \frac{\partial l_C(\psi; y)}{\partial \beta}, \frac{\partial l_C(\psi; y)}{\partial \delta}, \frac{\partial l_C(\psi; y)}{\partial \sigma}\right)^t$$

where $\frac{\partial l_C(\psi;y)}{\partial \pi} = \left(\frac{\partial l_C(\psi;y)}{\partial \pi_1}, \ldots, \frac{\partial l_C(\psi;y)}{\partial \pi_K}\right)$, $\frac{\partial l_C(\psi;y)}{\partial \beta} = \left(\frac{\partial l_C(\psi;y)}{\partial \beta_{11}}, \ldots, \frac{\partial l_C(\psi;y)}{\partial \beta_{Kn_{\beta_K}}}\right)$, $\frac{\partial l_C(\psi;y)}{\partial \delta} = \left(\frac{\partial l_C(\psi;y)}{\partial \delta_{11}}, \ldots, \frac{\partial l_C(\psi;y)}{\partial \delta_{Kn_{\delta_K}}}\right)$ and $\frac{\partial l_C(\psi;y)}{\partial \sigma} = \frac{\partial l_C(\psi;y)}{\partial \sigma_1}, \cdots, \frac{\partial l_C(\psi;y)}{\partial \sigma_K}$.

It is important to recall that the sum of all the values in the sequence π is equal to 1. Using this fact, we can express π_K as $\sum_{k=1}^{K} \pi_k = 1$. Additionally, when j is not equal to K, the derivative of π_K with respect to π_j is equal to -1. Consequently, we can conclude that for all values of k from 1 to $K-1$, the following equation holds true:

$$\frac{\partial l_C(\psi; y)}{\partial \pi_k} = \sum_{i=1}^{n} \left(\frac{z_{ik}}{\pi_k} - \frac{z_{iK}}{\pi_K}\right)$$

If we utilize equation 2.64 to forecast the likelihood of membership, with $1 \leq k \leq K$ and $1 \leq l \leq n_\theta$,

$$\frac{\partial l_C\left(\psi; y\right)}{\partial \theta_{kl}} = \sum_{i=1}^{n} x_{il}\left(Z_{ik} - \pi_{ik}\right)$$

For $1 \leq k \leq K$ and $1 \leq l \leq n_{\beta_k}$

$$\frac{\partial l_C(\psi; y)}{\partial \beta_{kl}} = \sum_{i=1}^{n}\sum_{t=1}^{T} \frac{z_{ik}}{\sigma_k^2} \frac{\partial f(a_{it}; \beta_k)}{\partial \beta_{kl}} \left(y_{it} - f(a_{it}; \beta_k)\right)$$

For $1 \leq k \leq K$

$$\frac{\partial l_C(\psi; y)}{\partial \sigma_k} = -\sum_{i=1}^{n}\sum_{t=1}^{T} \frac{z_{ik}\left[\sigma_k^2 - \left(y_{it} - f(a_{it}; \beta_k)\right)^2\right]}{\sigma_k^3}$$

3.5.3.1 Computation of the negative second derivative matrix

The negative of the second derivative matrix of the complete likelihood is,

$$-B_C(y;\hat{\psi}) = -\begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\pi^2} & \frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\beta} & \frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\delta} & \frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\sigma} \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\pi} & \frac{\partial^2 l_C(\psi;y)}{\partial\beta^2} & \frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta} & \frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma} \\ \frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\pi} & \frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\beta} & \frac{\partial^2 l_C(\psi;y)}{\partial\delta^2} & \frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\sigma} \\ \frac{\partial^2 l_C(\psi;y)}{\partial\sigma\partial\pi} & \frac{\partial^2 l_C(\psi;y)}{\partial\sigma\partial\beta} & \frac{\partial^2 l_C(\psi;y)}{\partial\sigma\partial\delta} & \frac{\partial^2 l_C(\psi;y)}{\partial\sigma^2} \end{pmatrix}$$

The size of this matrix is $\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+2K-1\right)\times\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+2K-1\right)$. In case we are using predictor for probability, we replace all derivatives by π by θ. The size of the matrix becomes $\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+(K-1)n_\theta)+K\right)\times\left(\sum_{k=1}^{K}(n_{\beta_k}+n_\delta)+(K-1)n_\theta)+K\right)$.

3.5.3.1.1 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi^2}$ or $\frac{\partial^2 l_C(\psi;y)}{\partial\theta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\pi^2}$ has for size $(K-1)\times(K-1)$.
For $1\le k,l\le K-1$ is

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi^2} = \left(\frac{\partial^2 l_C(\psi;y)}{\partial\pi_k\partial\pi_l}\right)_{kl}$$

which elements are for $1\le k,l\le K-1$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi_k\partial\pi_l} = \begin{cases} \sum_{i=1}^{n} -\left(\frac{z_{ik}}{\pi_k^2}+\frac{z_{iK}}{\pi_K^2}\right),\ k=l \\ \sum_{i=1}^{n} -\frac{z_{iK}}{\pi_K^2},\ k\neq l \end{cases}$$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\theta^2}$ has for size $Kn_\theta\times Kn_\theta$ and is composed by

$$\frac{\partial^2 l_C(\psi;y)}{\partial\theta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\theta_1\partial\theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\theta_1\partial\theta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\theta_K\partial\theta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\theta_K\partial\theta_K} \end{pmatrix}$$

which elements, for $1\le l,l'\le n_\theta$, are

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\theta_k\partial\theta_{k'}}\right)_{ll'} = \frac{\partial^2 l_C(\psi;y)}{\partial\theta_{kl}\partial\theta_{k'l'}} = \begin{cases} \sum_{i=1}^{n} -x_{il}x_{il'}\left(\frac{e^{\theta_k x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}}\right)\left(1-\frac{e^{\theta_k x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}}\right),\ k=k' \\ \sum_{i=1}^{n} x_{il}x_{il'}\frac{e^{\theta_k x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}}\frac{e^{\theta_{k'} x_i}}{\sum_{k=1}^{K}e^{\theta_k x_i}},\ k\neq k' \end{cases}$$

The outcomes remain the same for both π and θ in the subsequent sections.

3.5.3.1.2 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\beta}$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\beta} = 0$$

3.5.3.1.3 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\delta}$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\delta} = 0$$

3.5.3.1.4 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\sigma}$

$$\frac{\partial^2 l_C(\psi;y)}{\partial\pi\partial\sigma} = 0$$

3.5.3.1.5 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\beta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\beta^2}$ has for size $\sum_{k=1}^K n_{\beta_k} \times \sum_{k=1}^K n_{\beta_k}$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\beta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\beta_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\beta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$, $1 \leq l \leq n_{\beta_k}$ and $1 \leq l' \leq n_{\beta_{k'}}$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'}\partial\beta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial\beta_{k'l'}\partial\beta_{kl}} = \tag{3.459}$$

$$\begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{z_{ik}}{\sigma_k^2} \left(\frac{\partial^2 f(a_{it};\beta_k,\delta_k)}{\partial\beta_{kl'}\partial\beta_{kl}} (y_{it} - f(a_{it};\beta_k,\delta_k)) - \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\beta_{kl}} \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\beta_{kl'}} \right), \ k = k' \\ 0, \ k \neq k' \end{cases} \tag{3.460}$$

3.5.3.1.6 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta}$ has for size $\sum_{k=1}^K n_{\beta_k} \times K n_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\delta} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$, $1 \leq l \leq n_{\beta_k}$ and $1 \leq l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\delta_{k'}\partial\delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial\delta_{k'l'}\partial\delta_{kl}} = \tag{3.461}$$

$$\begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{z_{ik}}{\sigma_k^2} \left(\frac{\partial^2 f(a_{it};\beta_k,\delta_k)}{\partial\delta_{kl'}\partial\delta_{kl}} (y_{it} - f(a_{it};\beta_k,\delta_k)) - \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\delta_{kl}} \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\delta_{kl'}}\right), \; k = k' \\ 0, \; k \neq k' \end{cases} \tag{3.462}$$

3.5.3.1.7 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma}$ has for size $\sum_{k=1}^K n_{\beta_k} \times K$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\beta\partial\sigma} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\sigma_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_1\partial\sigma_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\sigma_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\beta_K\partial\sigma_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l \leq n_{\beta_k}$,

$$\left(\frac{\partial l_C(\psi;y)}{\partial\beta_{k'}\partial\sigma_k}\right)_{l'k} = \frac{\partial l_C(\psi;y)}{\partial\beta_{k'l}\partial\sigma_k} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{-2z_{ik}}{\sigma_k^3} \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\beta_{kl}} (y_{it} - f(a_{it};\beta_k,\delta_k)), \; k = k' \\ 0, \; k \neq k' \end{cases}$$

3.5.3.1.8 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2}$ has for size $Kn_\delta \times Kn_\delta$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\delta^2} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\delta_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\delta_1} & \dots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\delta_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l, l' \leq n_\delta$,

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\delta_{k'}\partial\delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial\delta_{k'l'}\partial\delta_{kl}} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T -z_{ik}\frac{w_{it}^l w_{it}^{l'}}{\sigma_k^2},\ k = k' \\ 0,\ k \neq k' \end{cases}$$

$$\left(\frac{\partial^2 l_C(\psi;y)}{\partial\delta_{k'}\partial\delta_k}\right)_{l'l} = \frac{\partial^2 l_C(\psi;y)}{\partial\delta_{k'l'}\partial\delta_{kl}} = \tag{3.463}$$

$$\begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{z_{ik}}{\sigma_k^2}\left(\frac{\partial^2 f(a_{it};\beta_k,\delta_k)}{\partial\delta_{kl'}\partial\delta_{kl}}\left(y_{it} - f(a_{it};\beta_k,\delta_k)\right) - \frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\delta_{kl}}\frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\delta_{kl'}}\right),\ k = k' \\ 0,\ k \neq k' \end{cases} \tag{3.464}$$

3.5.3.1.9 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\sigma}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\sigma}$ has for size $Kn_\delta \times K$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\delta\partial\sigma} = \begin{pmatrix} \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\sigma_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_1\partial\sigma_K} \\ \vdots & & \vdots \\ \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\sigma_1} & \cdots & \frac{\partial^2 l_C(\psi;y)}{\partial\delta_K\partial\sigma_K} \end{pmatrix}$$

which elements are, for $1 \leq k, k' \leq K$ and $1 \leq l \leq n_\delta$,

$$\left(\frac{\partial l_C(\psi;y)}{\partial\delta_{k'}\partial\sigma_k}\right)_{lk} = \frac{\partial l_C(\psi;y)}{\partial\delta_{k'l}\partial\sigma_k} = \begin{cases} \sum_{i=1}^n \sum_{t=1}^T \frac{-2z_{ik}}{\sigma_k^3}\frac{\partial f(a_{it};\beta_k,\delta_k)}{\partial\delta_{kl}}\left(y_{it} - f(a_{it};\beta_k,\delta_k)\right),\ k = k' \\ 0,\ k \neq k' \end{cases} \tag{3.465}$$

3.5.3.1.10 Second derivative $\frac{\partial^2 l_C(\psi;y)}{\partial\sigma^2}$

The second derivative matrix for $\frac{\partial^2 l_C(\psi;y)}{\partial\sigma^2}$ has for size $K \times K$ and is composed by block matrix

$$\frac{\partial^2 l_C(\psi;y)}{\partial\sigma^2} = \left(\frac{\partial^2 l_C(\psi;y)}{\partial\sigma_k\partial\sigma_l}\right)_{kl}$$

where

$$\frac{\partial^2 l_C(\psi;y)}{\partial\sigma_k\partial\sigma_l} = \begin{cases} -\sum_{i=1}^n \sum_{t=1}^T \frac{z_{ik}\left(-\sigma_k^2 + 3\left(y_{it} - f(a_{it};\beta_k)\right)^2\right)}{\sigma_k^4},\ k = l \\ 0,\ k \neq l \end{cases}$$

Since $E_\psi(Z_{ik}|Y_i = y_i) = \frac{\pi_k g_k(y_i,\theta)}{\sum_{k=1}^{K}\pi_k g_k(y_i,\theta_k)} = \tau_{ik}$, to compute the conditional expectation of the negative of the second derivative matrix, replace z_{ik} by τ_{ik} in the equations above. The resulting expression would be for the negative of the second derivative matrix, denoted as $-B_C(x;\hat{\Psi})$, given $X = x$. This allows to obtain a conditional expectation that takes into account the updated value τ_{ik} for z_{ik}.

3.5.4 Computation of $cov\left(S_C(\hat{\psi};x)|X=x\right)$

The conditional matrix of the score vector is given by

$$I_{y/u}\left(\hat{\psi};y\right) = \begin{pmatrix} cov\left(S_c(\pi)\right) & cov\left(S_c(\pi),S_c(\beta)\right) & cov\left(S_c(\pi),S_c(\delta)\right) & cov\left(S_c(\pi),S_c(\sigma)\right) \\ cov\left(S_c(\beta),S_c(\pi)\right) & cov\left(S_c(\beta)\right) & cov\left(S_c(\beta),S_c(\delta)\right) & cov\left(S_c(\beta),S_c(\sigma)\right) \\ cov\left(S_c(\delta),S_c(\pi)\right) & cov\left(S_c(\delta),S_c(\beta)\right) & cov\left(S_c(\delta)\right) & cov\left(S_c(\delta),S_c(\sigma)\right) \\ cov\left(S_c(\sigma),S_c(\pi)\right) & cov\left(S_c(\sigma),S_c(\beta)\right) & cov\left(S_c(\sigma),S_c(\delta)\right) & cov\left(S_c(\sigma)\right) \end{pmatrix}$$

The size of this matrix is $(K(n_\beta+n_\delta+2)-1)\times(K(n_\beta+n_\delta+2)-1)$ or $(K(n_\beta+n_\delta+n_\theta+1))\times(K(n_\beta+n_\delta+n_\theta+1))$ in case we are using predictor for probability.

As reminder (see proposition 2 page 45),

- $E_\psi(Z_{ik}) = \tau_{ik}$
- $var\left(Z_{ik}\right) = E\left(Z_{ik}^2\right) - E^2\left(Z_{ik}\right) = \tau_{ik}(1-\tau_{ik})$
- $cov\left(Z_{ik},Z_{il}\right) = E\left(Z_{ik}Z_{il}\right) - E\left(Z_{ik}\right)E\left(Z_{il}\right) = -\tau_{ik}\tau_{il}$ for $k \neq l$
- $cov\left(Z_{ik},Z_{jl}\right) = 0$

3.5.4.1 Matrix $cov\left(S_c(\pi)\right)$

The matrix as for dimension $(K-1)\times(K-1)$.

For a diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write for $1 \leq k \leq K-1$

$$cov\left(S_c(\pi)\right)_{kk} = \sum_{i=1}^{n}\left(\frac{\tau_{ik}(1-\tau_{ik})}{\pi_k^2} + \frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2} - 2\frac{\tau_{ik}\tau_{iK}}{\pi_k\pi_K}\right)$$

For a non-diagonal element of the matrix $cov\left(S_c(\pi)\right)$, we can write for $1 \leq k,l \leq K-1$

$$cov\left(S_c(\pi)\right)_{kl} = \sum_{i=1}^{n}\left(-\frac{\tau_{ik}\tau_{il}}{\pi_k\pi_l} + \frac{\tau_{ik}\tau_{iK}}{\pi_k\pi_K} + \frac{\tau_{iK}\tau_{il}}{\pi_K\pi_l} + \frac{\tau_{iK}(1-\tau_{iK})}{\pi_K^2}\right)$$

3.5.4.2 Matrix $cov\left(S_c(\pi), S_c(\beta)\right)$

$cov\left(S_c(\pi), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\beta_k)\right)$ for all groups k which dimension is $(K-1) \times n_{\beta_k}$. Thus, the dimension of the first matrix is $(K-1) \times Kn_{\beta_k}$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\beta_k)\right)$ that is a matrix with elements, for $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$

$$cov\left(S_c(\pi), S_c(\beta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^n \left(B_{ikl}\tau_{ik} \left(\frac{1-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' = k \\ \sum_{i=1}^n \left(B_{ikl}\tau_{ik} \left(-\frac{\tau_{ik'}}{\pi_{k'}} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\beta_K)\right)_{k'l} = \sum_{i=1}^n \left(B_{iKl}\tau_{iK} \left(\frac{1-\tau_{iK}}{\pi_K} + \frac{\tau_{ik'}}{\pi'_k} \right) \right)$$

where $B_{ikl} = \sum_{t=1}^T \dfrac{-\partial f(a_{it};\beta_k)}{\partial \beta_{kl}} \dfrac{(y_{it} - f(a_{it};\beta_k,\delta_k))}{\sigma_k^2}$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\beta_k}$.

3.5.4.3 Matrix $cov\left(S_c(\pi), S_c(\delta)\right)$

$cov\left(S_c(\pi), S_c(\delta)\right)$ is composed by the matrix $cov\left(S_c(\pi), S_c(\delta)\right)$ for all groups k which dimension is $(K-1) \times n_\delta$. Thus, the dimension of the first matrix is $(K-1) \times Kn_\delta$.

Given $1 \leq k \leq K-1$ we compute $cov\left(S_c(\pi), S_c(\delta_k)\right)$ that is a matrix with elements, for $1 \leq k' \leq K-1$ and $1 \leq l \leq n_\delta$

$$cov\left(S_c(\pi), S_c(\delta_k)\right)_{k'l} = \begin{cases} \sum_{i=1}^n \left(D_{ikl}\tau_{ik} \left(\frac{1-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' = k \\ \sum_{i=1}^n \left(D_{ikl}\tau_{ik} \left(-\frac{\tau_{ik'}}{\pi_{k'}} + \frac{\tau_{iK}}{\pi_K} \right) \right), k' \neq k \end{cases}$$

and for $k = K$,

$$cov\left(S_c(\pi), S_c(\delta_K)\right)_{k'l} = \sum_{i=1}^n \left(D_{iKl}\tau_{iK} \left(\frac{1-\tau_{iK}}{\pi_K} + \frac{\tau_{ik'}}{\pi'_k} \right) \right)$$

where $D_{ikl} = \sum_{t=1}^T \dfrac{-\partial f(a_{it};\beta_k)}{\partial \beta_{kl}} \dfrac{(y_{it} - f(a_{it};\beta_k,\delta_k))}{\sigma_k^2}$, $1 \leq k' \leq K-1$ and $1 \leq l \leq n_{\delta_k}$.

3.5.4.4 Matrix $cov\left(S_c(\pi), S_c(\sigma)\right)$

$cov\left(S_c(\pi), S_c(\sigma)\right)$ as for dimension $(K-1) \times K$.

For $1 \leq k \leq K-1$ and $1 \leq l \leq K-1$

$$(cov\,(S_c(\pi), S_c(\sigma)))_{kl} = \begin{cases} \sum_{i=1}^{n} \left(\tau_{ik} S_{ik} \left(\frac{(1-\tau_{ik})}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right)\right), \; k = l \\ \sum_{i=1}^{n} \left(\frac{\tau_{il} S_{il}}{\pi_l} \left(\frac{-\tau_{ik}}{\pi_k} + \frac{\tau_{iK}}{\pi_K}\right)\right), \; k \neq l \end{cases}$$

and for $l = K$

$$(cov\,(S_c(\pi), S_c(\sigma)))_{kK} = \sum_{i=1}^{n} \left(\tau_{iK} S_{iK} \left(\frac{(1-\tau_{iK})}{\pi_K} - \frac{\tau_{ik}}{\pi_k}\right)\right)$$

where $S_{ik} = -\sum_{t=1}^{T} \frac{\left[\sigma_k^2 - (y_{it} - f(a_{it}; \beta_k, \delta_k))^2\right]}{\sigma_k^3}$

3.5.4.5 Matrix $cov\,(S_c(\theta))$

If we use predictors for the membership probability we have to calculate the matrix with θ parameters.

$cov\,(S_c(\theta))$ is composed by the matrix $cov\,(S_c(\theta_k), S_c(\theta_l))$ for all groups k and l which dimension is $n_\theta \times n_\theta$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_\theta$.

A diagonal matrix, for $1 \leq k \leq K$ and $1 \leq p, q \leq n_\theta$ is done by

$$(cov\,(S_c(\theta_k), S_c(\theta_k)))_{pq} = \sum_{i=1}^{n} x_{ip} x_{iq} \tau_{ik} (1 - \tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$ $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_\theta$, is done by

$$(cov\,(S_c(\theta_k), S_c(\theta_l)))_{pq} = -\sum_{i=1}^{n} x_{ip} x_{iq} \tau_{ik} \tau_{il}$$

3.5.4.6 Matrix $cov\,(S_c(\theta), S_c(\beta))$

$cov\,(S_c(\theta), S_c(\beta))$ is composed by the matrix $cov\,(S_c(\theta_k), S_c(\beta_l))$ for all groups k and l which dimension is $n_\theta \times n_{\beta_l}$. Thus, the dimension of the first matrix is $Kn_\theta \times Kn_{\beta_l}$.

A diagonal matrix, for $1 \leq k \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\beta_k}$ is done by

$$(cov\,(S_c(\theta_k), S_c(\beta_k)))_{pq} = \sum_{i=1}^{n} x_{ip} B_{ikq} \tau_{ik} (1 - \tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_\theta$ and $1 \leq q \leq n_{\beta_l}$, is done by

$$(cov\left(S_c(\theta_k), S_c(\beta_l)\right))_{pq} = \sum_{i=1}^{n} -x_{ip}B_{ilq}\tau_{ik}\tau_{il}$$

3.5.4.7 Matrix $cov\left(S_c(\theta), S_c(\delta)\right)$

$cov\left(S_c(\theta), S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\delta_l)\right)$ for all groups k and l which dimension is $n_{\theta_k} \times n_{\delta_l}$. Thus, the dimension of the first matrix is $Kn_{\theta_k} \times Kn_{\delta_l}$.

A diagonal matrix, for $1 \leq k \leq K$, $1 \leq p \leq n_{\theta_k}$ and $1 \leq q \leq n_{\delta_l}$ is done by

$$(cov\left(S_c(\theta_k), S_c(\delta_k)\right))_{pq} = \sum_{i=1}^{n} x_{ip}D_{ikq}\tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_{\theta_k}$ and $1 \leq q \leq n_{\delta_l}$, is done by

$$(cov\left(S_c(\theta_k), S_c(\delta_l)\right))_{pq} = \sum_{i=1}^{n} -x_{ip}D_{ilq}\tau_{ik}\tau_{il}$$

3.5.4.8 Matrix $cov\left(S_c(\theta), S_c(\sigma)\right)$

$cov\left(S_c(\theta), S_c(\sigma)\right)$ is composed by the matrix $cov\left(S_c(\theta_k), S_c(\sigma_l)\right)$ for all groups k, l which dimension is $(n_\theta \times K)$. Thus, the dimension of the first matrix is $Kn_\theta \times K$.

Given $1 \leq k \leq K$ we compute $cov\left(S_c(\theta_k), S_c(\sigma_l)\right)$ that is a matrix with elements $1 \leq k', l \leq K$ and $1 \leq p \leq n_\theta$,

$$cov\left(S_c(\theta_k), S_c(\sigma_l)\right)_p = \begin{cases} \sum_{i=1}^{n} x_{ip}S_{ik}\tau_{ik}(1-\tau_{ik}), \ k = l \\ \sum_{i=1}^{n} -x_{ip}S_{il}\tau_{ik}\tau_{il}, \ k \neq l \end{cases}$$

3.5.4.9 Matrix $cov\left(S_c(\beta)\right)$

$cov\left(S_c(\beta)\right)$ is composed by the matrix $cov\left(S_c(\beta_k), S_c(\beta_l)\right)$ for all groups k and l which dimension is $n_{\beta_k} \times n_{\beta_l}$. Thus, the dimension of the first matrix is $Kn_{\beta_k} \times Kn_{\beta_l}$.

A diagonal matrix, for $1 \leq k \leq K$ and $1 \leq p, q \leq n_{\beta_k}$ is done by

$$(cov\left(S_c(\beta_k), S_c(\beta_k)\right))_{pq} = \sum_{i=1}^{n} B_{ikp}B_{ikq}\tau_{ik}(1-\tau_{ik})$$

A non diagonal matrix, for $1 \leq k, l \leq K$, $1 \leq p \leq n_{\beta_k}$ and $1 \leq q \leq n_{\beta_l}$, is done by

$$(cov\,(S_c(\beta_k), S_c(\beta_l)))_{pq} = \sum_{i=1}^{n} -B_{ikp}B_{ilq}\,(\tau_{ik}\tau_{il})$$

3.5.4.10 Matrix $cov\,(S_c(\beta), S_c(\sigma))$

$cov\,(S_c(\beta), S_c(\sigma))$ is composed by the matrix $cov\,(S_c(\beta_k), S_c(\sigma_l))$ for all groups k, l which dimension is $(n_{\beta_k} \times K)$. Thus, the dimension of the first matrix is $Kn_{\beta_k} \times K$.

Given $1 \leq k, l \leq K$ we compute $cov\,(S_c(\beta_k), S_c(\sigma_l))$ that is a matrix with elements $1 \leq p \leq n_\theta$ and

$$cov\,(S_c(\beta_k), S_c(\sigma_l))_p = \begin{cases} \sum_{i=1}^{n} B_{ikp}S_{ik}\tau_{ik}(1-\tau_{ik}),\ k = l \\ \sum_{i=1}^{n} -B_{ikp}S_{il}\tau_{ik}\tau_{il},\ k \neq l \end{cases}$$

3.5.4.11 Matrix $cov\,(S_c(\sigma))$

The matrix as for dimension $K \times K$.

For a diagonal element of the matrix $cov\,(S_c(\sigma))$, we can write for $1 \leq k \leq K$ and $1 \leq l \leq K$,

$$cov\,(S_c(\sigma))_{kl} = \begin{cases} \sum_{i=1}^{n} S_{ik}^2\tau_{ik}(1-\tau_{ik}),\ k = l \\ \sum_{i=1}^{n} -S_{ik}S_{il}\tau_{ik}\tau_{il},\ k \neq l \end{cases}$$

3.5.5 Numerical applications

In the following, we will test the procedure described above. In each example, we constructed a sample with a structure of K clusters, each containing 500 values from a variable Y. The trajectories of these values follow a nonlinear group-controlled pattern, which is determined by the definition of the function f.

3.5.5.1 Two groups

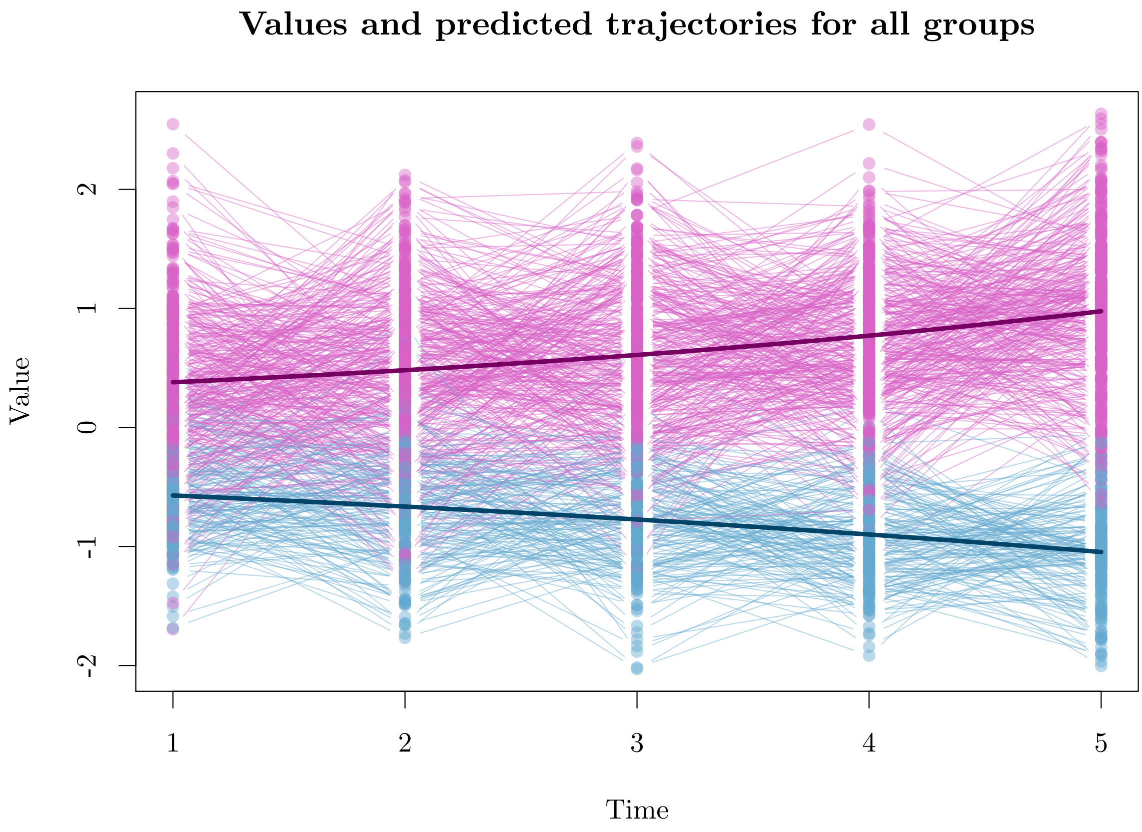
We set theoretical values as

Function	$f(t;\beta_k) = \beta_{k1}e^{\beta_{k2}t}$	
Cluster	Parameters	Probability π_k
1	$\beta_1 = (-0.5, 0.15)$ $\sigma_1 = 0.45$	0.32
2	$\beta_2 = (0.3, 0.24)$ $\sigma_2 = 0.65$	0.68

We have started the algorithm with default starting values, see chapter 5. We find

	Theoretical	Likelihood		EM	
			SE		SE
β_{11}	-0.5	-0.46084	0.02589	-0.46084	0.02589
β_{12}	0.15	0.17615	0.01446	0.17615	0.01446
β_{21}	0.3	0.31839	0.02362	0.3184	0.02362
β_{22}	0.24	0.22783	0.01833	0.22783	0.01833
σ_1	0.45	0.43453	0.02592	0.43455	0.01126
σ_2	0.65	0.66971	0.01715	0.66971	0.01149
π_1	0.32	0.31118	0.04837	0.31115	0.02071
π_2	0.68	0.68882	0.04837	0.68885	0.02071

We have created graphs that display the values and the trajectory shapes for all the groups.



3.5.5.2 Three groups

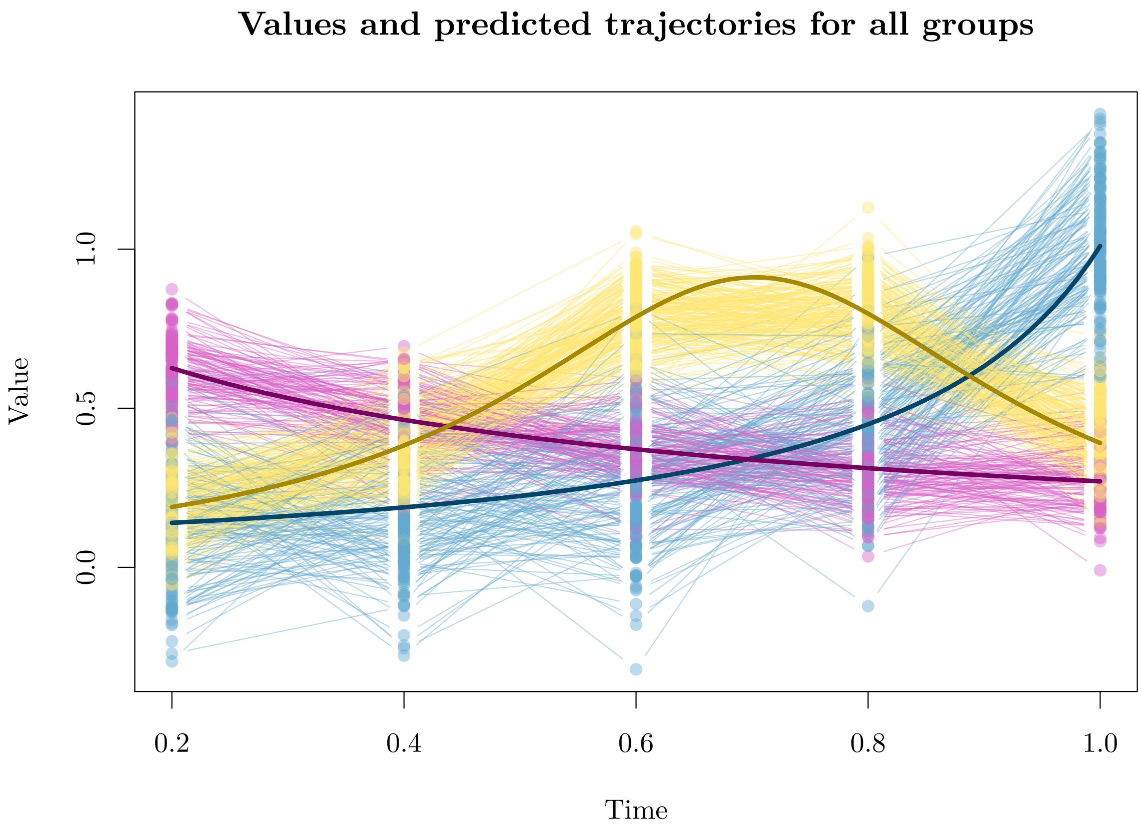
We set theoretical values as

Function	$f\left(t;\beta_k\right)=\dfrac{1}{\beta_{k1}+\beta_{k2}t+\beta_{k3}t^2}$	
Cluster	Parameters	Probability π_k
1	$\beta_1=(1,3,-0.2)$ $\sigma_1=0.1$	0.32
2	$\beta_2=(8.94,-22.4,16)$ $\sigma_2=0.1$	0.54
3	$\beta_3=(8,-8,1)$ $\sigma_2=0.2$	0.14

In this example, the Likelihood method cannot be used because the gradient of $f(t;\beta_k)$ has no root. Therefore, we use only the EM method. The initial parameters are $\pi_1=\pi_2=\pi_3=1/3$, $\beta_1=(5.59293,0,0)$, $\beta_2=(2.198027,0,0)$, $\beta_3=(1.367784,0,0)$, and $\sigma_1=\sigma_2=\sigma_3=0.285456$.

	Theoretical	EM	
		Param.	SE
β_{11}	1	1.01318	0.08021
β_{12}	3	2.97746	0.42337
β_{13}	-0.2	-0.29066	0.4232
β_{21}	8.94	9.25528	0.17665
β_{22}	-22.4	-23.22537	0.51216
β_{23}	16	16.52991	0.36616
β_{31}	8	9.19966	1.00919
β_{32}	-8	-10.75779	2.49206
β_{33}	1	2.54838	1.5088
σ_1	0.1	0.09973	0.00287
σ_2	0.1	0.10193	0.00215
σ_3	0.2	0.19555	0.00512
π_1	0.32	0.25166	0.01944
π_2	0.54	0.45157	0.02823
π_3	0.14	0.29677	0.02047

We have created graphs that display the values and the trajectory shapes for all the groups.



3.5.5.3 Two groups

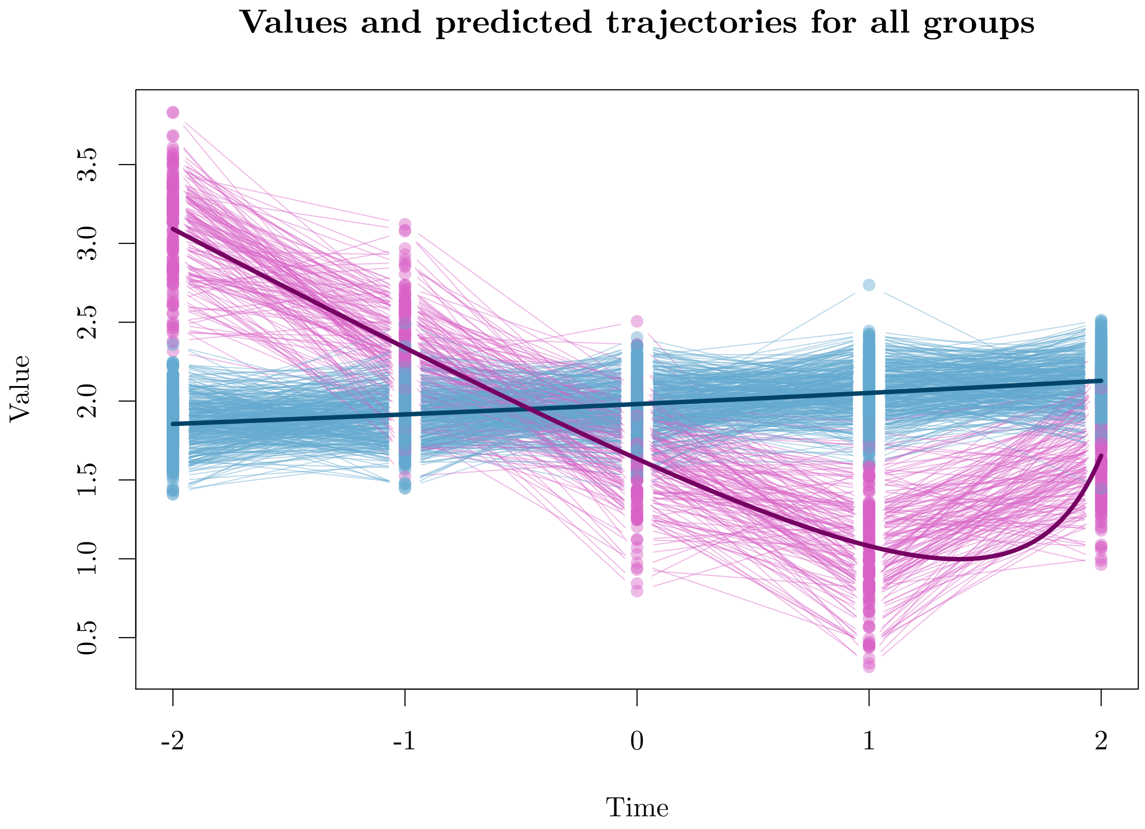
We set theoretical values as

Function	$f(t;\beta_k) = \dfrac{\beta_{k1} + \beta_{k2}(t-2)^2}{\beta_{k3} + \beta_{k4}(t-2)^2}$	
Cluster	Parameters	Probability π_k
1	$\beta_1 = (0.83, 1, 0.5, -1.23)$ $\sigma_1 = 0.32$	0.32
2	$\beta_2 = (12, 0.03, 5.54, -0.3)$ $\sigma_2 = 0.18$	0.68

Clearly, this model lacks parameter identifiability for β_k. In other words, finding a specific value for β_k that fits the model is not unique, as any multiple of β_k would also be a valid fit. Nevertheless, we can evaluate the algorithm's performance by checking whether it correctly identifies clusters, determines the appropriate values for σ_k and probabilities, and captures the correct trajectory shapes. The initial parameters used are $\pi_1 = \pi_2 = 0.5$, $\beta_1 = (1.779046, 0, 1, 0)$, $\beta_2 = (2.193121, 0, 1, 0)$, and $\sigma_1 = \sigma_2 = 0.4806698$. Some parameters might not provide valid standard errors.

	Theoretical	EM
β_{11}	0.83	0.6054
β_{12}	1	0.26984
β_{13}	0.5	0.39223
β_{14}	-1.23	-0.29846
β_{21}	12	4.32295
β_{22}	0.03	-0.02109
β_{23}	5.54	2.05088
β_{24}	-0.3	-0.03011
σ_1	0.32	0.35637
σ_2	0.18	0.18479
π_1	0.32	0.31552
π_2	0.68	0.68448

We have created graphs that display the values and the trajectory shapes for all the groups.



We can obtain another set of parameter approximations for the model by multiplying each $\hat{\beta}_k$ by the average $average_k = \sum_{l=1}^{4} \frac{\beta_{kl}}{\hat{\beta_{kl}}}$, which is the sum of the quotients of the real parameters to their corresponding approximations.

$$average_1 \times \hat{\beta}_1 = (1.58506, 0.7065, 1.02693, -0.78143)$$

and

$$average_2 \times \hat{\beta}_2 = (15.15009, -0.07392, 7.18746, -0.10552)$$

CHAPTER 4

Identifiability of the model

4.1 Identifiability

4.1.1 Introduction

In longitudinal studies, we can view the individual observations Y_{it} for $1 \leq i \leq n$ and $1 \leq t \leq T$ as a vector $Y_i = (Y_{i1}, Y_{i2}, ..., Y_{iT})$. Therefore, under the assumption that the individual observations Y_{it} are independent and identically distributed (i.i.d.), we can extend the analysis from a one-dimensional distribution to a T-dimensional distribution.

$$f_k(y_i; \psi) = \sum_{k=1}^{K} \pi_k g_k(y_i; \theta_k) \tag{4.1}$$

Utilizing the linearity of both integration and differentiation, this is equivalent to the following expression:

$$F_k(y_i; \psi) = \sum_{k=1}^{K} \pi_k G_k(y_i; \theta_k) \tag{4.2}$$

where F_k is the cdf of f_k and G_k the cdf of g_k.
We rely on Teicher (1963a) to establish the identifiability of multiple distributions, with particular references to Yakowitz and Spragins (1968) who provided a proof for multidimensional distributions.

Consider a family denoted as $\mathcal{F} = \left\{F(x; \alpha),\ x \in \mathbb{R}^T,\ \alpha \in \mathbb{R}_1^n\right\}$ comprising T-dimensional cumulative distribution functions (cdf) indexed by a parameter α within a Borel subset $\mathbb{R}_1^n$ of the Euclidean n-space $\mathbb{R}^n$. Each $F(x; \alpha)$ is measurable in $\mathbb{R}^T \times \mathbb{R}_1^n$. The n-dimensional cdf $H(x)$ can be expressed as the image of a mapping, denoted as Q, acting on the n-dimensional cdf G (with the measure μ_g induced by G assigning a measure of 1 to $\mathbb{R}_1^n$). H is referred to as the mixture of $\mathcal{F}$, and G represents the mixing distribution.

Let $\mathcal{G}$ represent the collection of all n-dimensional cdf's G, and $\mathcal{H}$ denotes the corresponding set of induced mixtures H. We consider $\mathcal{H}$ to be identifiable if the mapping Q is a one-to-one map from $\mathcal{G}$ to $\mathcal{H}$. A finite mixture occurs when the mixing distribution, or more specifically, the associated measure μ_g, is discrete and assigns positive mass to only a finite number of points in $\mathbb{R}_1^n$.

In this scenario, we can represent the mixture as described in our paper. The set $\mathcal{H}$, consisting of all finite mixtures from the class $\mathcal{F}$ of distributions, corresponds to the convex hull of $\mathcal{F}$.

$$\mathcal{H} = \left\{ H(x) : H(x) = \sum_{i=1}^{N} c_i F(x, \alpha_i),\ c_i > 0, \sum_{i=1}^{N} c_i = 1,\ F(x, \alpha_i) \in \mathcal{F},\ N = 1, 2, \cdots \right\} \tag{4.3}$$

The α_i's are presumed distinct.
We remark that $\mathcal{H}$ is equivalent to

$$\mathcal{H}_f = \left\{ h(x) : h(x) = \sum_{i=1}^{N} c_i f(x, \alpha_i),\ c_i > 0, \sum_{i=1}^{N} c_i = 1,\ f(x, \alpha_i) \text{ the pdf of } F(x, \alpha_i) \in \mathcal{F},\ N = 1, 2, \cdots \right\} \tag{4.4}$$

In this context, the definition of identifiability means that $\mathcal{F}$ generates an identifiable finite mixture model if and only if

$$\sum_{i=1}^{N} c_i F_i = \sum_{i=1}^{M} c_i' F_i' \tag{4.5}$$

implies $N = M$ and for each i, $1 \leq i \leq N$ there is some j, $1 \leq j \leq N$, such that $c_i = c_j'$ and $F_i = F_j'$. F_i means $F(x, \alpha_i)$.

In the context of mixture models, there are two main types of indeterminates:

- Label Switch: Label Switch occurs when the labels or elements of a mixture model are swapped without change basic distribution. In other words, components can be relabeled in different ways, leading to different parameter estimates representing the same statistical model. This unrecognizability can make it difficult to interpret results and compare models because components can be rearranged arbitrarily.

- Parameter Redundancy : Parameter redundancy occurs when some parameters of a mixture model are not uniquely identifiable from the observed data. This can happen when there is not enough information in the data to estimate all the parameters separately. In such cases, it is not possible to determine the exact values of certain parameters, and different parameter values may lead to the same likelihood.

4.1.2 Characterizations of identifiability

Theorem 2. *A necessary and sufficient condition that the class $\mathcal{H}$ of all finite mixtures of the family $\mathcal{F}$ be identifiable is that the $\mathcal{F}$ be a linearly independent set over the field of real numbers.*

We note $< A >$ the span of A over the real numbers.

Proof. Necessity.
Suppose that the family $\mathcal{F}$ is not linearly independent. Thus, there exist a linear relation in $\mathcal{F}$, $\sum_{i=1}^{N} a_i F_i = 0,\ a_i \in \mathbb{R}^*$ such, at least, one value $a_i \neq 0$. Without lose on generality, we can suppose that a_i are subscript such as, for some $M \in \mathbb{N}$, $a_i < 0 \Leftrightarrow i \leq M$. Then $\sum_{i=1}^{M} |a_i| F_i = \sum_{i=M+1}^{N} |a_i| F_i$. Not all values of a_i can be exclusively negative or positive, as F_i represents cumulative distribution functions (CDFs) that are inherently positive.
Since F_i are cdf, we have $\lim\limits_{x \to (+\infty, \cdots, +\infty)} F_i(x) = 1$ and

$$\lim_{x \to (+\infty, \cdots, +\infty)} \sum_{i=1}^{M} |a_i| F_i = \lim_{x \to (+\infty, \cdots, +\infty)} \sum_{i=M+1}^{N} |a_i| F_i \tag{4.6}$$

$$\sum_{i=1}^{M} |a_i| = \sum_{i=M+1}^{N} |a_i| \tag{4.7}$$

By noting $b = \sum_{i=M+1}^{N} |a_i|$, we remark that $b > 0$ and if $c_i = \frac{|a_i|}{b}$ for all i we obtain the following relationship:

$$\sum_{i=1}^{M} c_i F_i = \sum_{i=M+1}^{N} c_i F_i$$

Clearly by definition $\sum_{i=1}^{M} c_i = \sum_{i=M+1}^{N} c_i = 1$ and thus we have two different distinct representation of the same mixture and therefore $\mathcal{H}$ is not identifiable.

Sufficiency.
If $\mathcal{F}$ is linearly independent there exist a basis of $< \mathcal{F} >$. If we suppose that $\mathcal{H}$ is non identifiable there exist two distinct representations of the same mixture. Therefore $\mathcal{H} \subset < \mathcal{F} >$ and they would contradict the uniqueness of representation property of bases. □

Theorem 3 (Teicher, 1963). *Let $\mathcal{F} = \{F\}$ be a family of cdf's of one dimensional with transforms $\phi(t)$ defined for $t \in S_\phi$, the domain of definition of ϕ, such that the mapping $M : F \to \phi$ is linear and one-to-one. Suppose that there exists a total ordering ($\preccurlyeq$) of $\mathcal{F}$ such that $F_1 \prec F_2$ implies*

(i) $S_{\phi_1} \subseteq S_{\phi_2}$

(ii) the existence of some $t_1 \in \overline{S}_{\phi_1}$ *(t_1 being independent of ϕ_2) such that* $\lim\limits_{t \to t_1} \phi_2(t)/\phi_1(t) = 0$.

Then the class $\mathcal{H}$ of all finite mixtures of $\mathcal{F}$ is identifiable.

Proof. Suppose that there are two finite sets of elements of $\mathcal{F}$, say $\mathcal{F}_1 = \{F_i,\ 1 \le i \le k\}$ and $\mathcal{F}_2 = \{\hat{F}_i,\ 1 \le i \le \hat{k}\}$ such that

$$\sum_{i=1}^{k} c_i F_i(x) = \sum_{i=1}^{\hat{k}} \hat{c}_j \hat{F}_j(x),\ 0 < c_i, \hat{c}_j \le 1,\ \sum_{i=1}^{k} c_i = \sum_{j=1}^{\hat{k}} \hat{c}_j = 1 \tag{4.8}$$

Without any loss of generality, we can arrange the indices of the cdf's such that $F_i \prec F_j$ and $\hat{F}_i \prec \hat{F}_j$ for all $i < j$.

If $F_1 \ne \hat{F}_1$, we can also assume, without loss of generality, $F_1 \prec \hat{F}_1$. Then $F_1 \prec \hat{F}_j,\ 1 \le j \le \hat{k}$ and from the modified version of equation (4.8), it follows that for $t \in T_1 = S_{\phi_1}[t : \phi_1(t) \ne 0]$,

$$c_1 + \sum_{i=2}^{k} c_i[\phi_i(t)/\phi_1(t)] = \sum_{j=1}^{\hat{k}} \hat{c}_j[\hat{\phi}_j(t)/\phi_1(t)] \tag{4.9}$$

Since $F_1 \prec F_i,\ 1 \le i \le k$ and $F_1 \prec \hat{F}_j,\ 1 \le j \le \hat{k}$, we have

- $S_{\phi_1} \subseteq S_{\phi_i},\ 1 \le i \le k$ and $S_{\phi_1} \subseteq S_{\hat{\phi}_j},\ 1 \le j \le \hat{k}$
- there exist some $t_1 \in \overline{S}_{\phi_1}$ (t_1 being independent of ϕ_i and $\hat{\phi}_j$) such that $\lim\limits_{t \to t_1} \phi_i(t)/\phi_1(t) = 0$ and $\lim\limits_{t \to t_1} \hat{\phi}_j(t)/\phi_1(t) = 0$ for $1 \le i \le k$ and $1 \le j \le \hat{k}$.

Thus, letting $t \to t_1$ through values in T_1 (this is possible because $t_1 \in \overline{S}_{\Phi_1}$), we have $c_1 = 0$ contradicting (4.8) that $c_1 > 0$. By consequence, $F_1 = \hat{F}_1$ and for $t \in T_1$

$$(c_1 - \hat{c}_1) + \sum_{i=2}^{k} c_i[\phi_i(t)/\phi_1(t)] = \sum_{j=2}^{\hat{k}} \hat{c}_j[\hat{\phi}_j(t)/\phi_1(t)] \tag{4.10}$$

Again letting $t \to t_1$ through values in T_1, $c_1 = \hat{c}_1$ whence

$$\sum_{i=2}^{k} c_i F_i(x) = \sum_{j=2}^{\hat{k}} \hat{c}_j \hat{F}_j(x) \tag{4.11}$$

Repeating the prior argument a finite number of times, we conclude that $F_i = \hat{F}_i$ and $c_i = \hat{c}_i$ for $i = 1, 2, \cdots \min(k, \hat{k})$.

Further, if $k \ne \hat{k}$, say $k > \hat{k}$ then $\sum_{i=\hat{k}+1}^{k} c_i F_i(x) = 0$ implying $c_i = 0,\ \hat{k}+1 \le i \le k$ in contradiction to (4.8).

Thus, $k = \hat{k}$, $c_i = \hat{c}_i$ and $F_i = \hat{F}_i,\ 1 \le i \le k$, implying $\mathcal{F}_1 = \mathcal{F}_2$ and identifiability of $\mathcal{H}$.

□

Teicher (1967) extended the concept of identifiability from mixtures to mixtures of products, whether finite or infinite.

Let $\mathcal{F}_{n,m} = \{F(x;\alpha) : \alpha \in R_1^m, x \in \mathbb{R}^n\}$ a family of n-dimensional cdf's indexed by α in a Borel subset R_1^m of $\mathbb{R}^m$ such that $F(x;\alpha)$ is measurable in $\mathbb{R}^n \times R_1^m$, $\mathcal{G}$ the class of m-dimensionnal cdf's G whose induce measures μ_G assign measure one to $\mathbb{R}^n$ and $\mathcal{H} = \left\{\int_{R_1^m} F(x;\alpha)dG(\alpha),\ G \in \mathcal{G}\right\}$. It can be observed that when the entire mass of a measure μ_G is concentrated on a finite set of points in R_1^m, the resulting mixture H is finite. In this case, we align with the definition of $\mathcal{H}$ introduced earlier in the context of finite mixture models.

For $n = 1$, we write $\mathcal{F}_{1,m} = \mathcal{F}$ and for $n \geq 1$,

$$\mathcal{F}^*_{n,mn} = \left\{F^*(x;\alpha) :\ F^*(x;\alpha) = \prod_{i=1}^{n} F(x_i;\alpha_i),\ F(x_i;\alpha_i) \in \mathcal{F}, 1 \leq i \leq n\right\}$$

Theorem 4 (Teicher)**.** *If the class of all mixtures of $\mathcal{F}_{1,m}$ is identifiable, then for every $n > 1$, the class of mixtures of $\mathcal{F}^*_{n,mn}$ is likewise identifiable.*

Conversely, if for some $n > 1$, the class of all mixtures of $\mathcal{F}^*_{n,mn}$ is identifiable, the same is true for $\mathcal{F}_{1,m}$.

Proof. The converse is trivial since taking $F(x;\alpha) \in \mathcal{F}_{1,m}$, if

$$\int F(x;\alpha)dG(\alpha) = \int F(x;\alpha)d\hat{G}(\alpha)$$

By multiplying both sides of the equation by $\prod_{i=1}^{n-1} F(x_i;\alpha_0)$ where $\alpha_0 \in R_1^m$,

$$\int F(x;\alpha)\prod_{i=1}^{n-1} F(x_i;\alpha_0)dG(\alpha)d\mathbb{1}_{\alpha_0}\cdots d\mathbb{1}_{\alpha_0} = \int F(x;\alpha)\prod_{i=1}^{n-1} F(x_i;\alpha_0)d\hat{G}(\alpha)dG(\alpha)d\mathbb{1}_{\alpha_0}\cdots d\mathbb{1}_{\alpha_0}$$

Since $\mathcal{F}^*_{n,mn}$ is identifiable, necessarily $\mathbb{1}_{\alpha_0}\cdots\mathbb{1}_{\alpha_0}\cdot G = \mathbb{1}_{\alpha_0}\cdots\mathbb{1}_{\alpha_0}\cdot\hat{G}$ and therefore $G = \hat{G}$.

To prove the first part of the theorem, we assume that it is true for some n, and we will show that it is also true for $n+1$. This means that if we assume the class of all mixtures of $F^* \in \mathcal{F}^*_{n,mn}$ is identifiable, the same is true for $F^* \in \mathcal{F}^*_{n+1,m(n+1)}$.

We assume that, for $F^* \in \mathcal{F}^*_{n,mn}$, $F \in \mathcal{F}_{1,m}$,

$$\int F^*(x;\alpha)F(y;\beta)dG(\alpha,\beta) = \int F^*(x;\alpha)F(y;\beta)d\hat{G}(\alpha,\beta) \tag{4.12}$$

Let $G_2(\beta)$ and $\hat{G}_2(\beta)$ denotes the marginal distributions of β corresponding to G and $\hat{G}$. Let

$G(\alpha|\beta)$ and $\hat{G}(\alpha|\beta)$ denotes the versions of the conditional probabilities, such that for each β there are distribution functions in the variable α and for each α, $G(\alpha|\beta)$ and $\hat{G}(\alpha|\beta)$ are equal almost everywhere to measurable functions of β.

Then equation (4.12) become

$$\int F(y;\beta)H(x;\alpha)dG_2(\beta) = \int F(y;\beta)\hat{H}(x;\alpha)d\hat{G}_2(\beta) \tag{4.13}$$

where

$$H(x;\alpha) = \int F^*(x;\alpha)d_\alpha G(\alpha,\beta) \tag{4.14}$$

$$\hat{H}(x;\alpha) = \int F^*(x;\alpha)d_\alpha \hat{G}(\alpha,\beta) \tag{4.15}$$

Remember that if for some function measurable on X f, $\varphi(E) = \int_E f d\mu$ we have $\int_X g d\varphi = \int_X g f d\mu$ for every g measurable on X (see for example Rudin 1.29), we can write equation (4.13) as

$$\int F(y;\beta)dJ_x(\beta) = \int F(y;\beta)d\hat{J}_x(\beta) \tag{4.16}$$

where for each β

$$J_x(\beta) = \int_{-\infty}^{\beta} H(x;\gamma)dG_2(\gamma) \tag{4.17}$$

$$\hat{J}_x(\beta) = \int_{-\infty}^{\beta} \hat{H}(x;\gamma)d\hat{G}_2(\gamma) \tag{4.18}$$

Since F^* is a cdf, $H(x;\alpha) \leq 1$ and $\hat{H}(x;\alpha) \leq 1$ therefore $J_x(\beta) \leq G_2(\beta)$ and $\hat{J}_x(\beta) \leq \hat{G}_2(\beta)$.

Since $G_2(\beta)$ and $\hat{G}_2(\beta)$ are cdf, the dominated convergence insures that for each x, $J_x(\infty) = \hat{J}_x(\infty)$ is a finite value.

Thus, from (4.16) and the initial hypothesis $J_x = \hat{J}_x$, that is, for each β,

$$\int_{-\infty}^{\beta} H(x;\gamma)dG_2(\gamma) = \int_{-\infty}^{\beta} \hat{H}(x;\gamma)d\hat{G}_2(\gamma) \tag{4.19}$$

Letting $x \to \infty$ in equations (4.14) and (4.13) yields

$$\int F(y;\beta)dG_2(\beta) = \int F(y;\beta)d\hat{G}_2(\beta) \tag{4.20}$$

implies as above that

$$G_2 = \hat{G}_2 \tag{4.21}$$

However, equation (4.20) with (4.19) necessitates $H(x;\beta) = \hat{H}(x;\beta)$ for almost all β.

Combines with equation (4.14) and the hypothesis $\mathcal{F}^*_{n,mn}$ is identifiable, entails

$$G(\cdot|\beta) = \hat{G}(\cdot|\beta), \text{ almost all } \beta \tag{4.22}$$

Finally, combining (4.20) and (4.22), $G(\cdot|\cdot) = \hat{G}(\cdot|\cdot)$ so that $\mathcal{F}^*_{n+1,m(n+1)}$ is identifiable. □

4.1.3 Some identifiable family of mixtures

Proposition 4 (Teicher, 1963)**.** *The class of all mixtures of one dimensional normal distributions is identifiable.*

Proof. Let $N = N(x;\mu;\sigma^2)$ denote the normal cdf with mean μ and variance σ^2. Its bilateral Laplace transform is given by $\phi(t;\mu;\sigma^2) = e^{\sigma^2 t^2/2-\mu t}$. Order the family lexicographically by : $N_1 = N(x;\mu_1;\sigma_1^2) \prec N(x;\mu_2;\sigma_2^2) = N_2$ if $\sigma_1 > \sigma_2$ or if $\sigma_1 = \sigma_2$ but $\mu_1 < \mu_2$. Then the theorem 3 applies with $S_\phi = (-\infty,\infty)$ and $t_1 = +\infty$. □

Proposition 5 (Yakowitz & Spragins, 1968)**.** *The family $\mathcal{F}$ of n-dimensional Gaussian cdf's generate identifiable finite mixtures.*

Proof. Let's assume that $\mathcal{F}$ is not identifiable, which means that $\mathcal{F}$ is not linearly independent. In this case, there exist non-zero real values c_i, $M > 0$, and distinct cdf F_i such that $\sum_{i=1}^M c_i F_i = 0$. We can express this as $\sum_{i=1}^M c_i f_i$, where f_i represents the pdf of F_i. Each f_i corresponds to the pdf of a Gaussian distribution with mean vector θ_i and covariance matrix Δ_i for a random variable X_i.

We should also note that the moment generative function (mgf) of a continuous random variable X is defined, if it exists, as the bilateral Laplace transform of its probability density function f_X.

$$M_X(t) = \mathrm{E}\left(e^{\mathbf{t}^{\mathrm{T}}\mathbf{X}}\right) = \mathcal{B}\{f_X\}(-t)$$

where $\mathcal{B}\{f_X\}(s) = \int_{-\infty}^{\infty} e^{-sx} f_X(x)\,dx$,

Since the Laplace transform is linear, we have

$$\sum_{i=1}^M c_i M_{X_i}(t) = \sum_{i=1}^M c_i e^{\frac{1}{2}t'\Delta_i t + t'\theta_i} = 0 \tag{4.23}$$

where the points (θ_i, Δ_i) are all distinct.

Setting $t = cu$ where c is a scalar and u a vector, (4.23) becomes

$$\sum_{i=1}^M c_i e^{\frac{c^2}{2}u'\Delta_i u + cu'\theta_i} = 0 \tag{4.24}$$

If we assume that all the covariance matrices Δ_i for $1 \leq i \leq M$ are identical, then all the vectors θ_i for $1 \leq i \leq M$ are distinct. For $1 \leq i \neq j \leq M$,

$$u'\theta_i = u'\theta_j \iff u \in H_i = \big\{u : u'(\theta_i - \theta_j) = 0\big\}$$

So, if

$$u \notin \cup_{1\leq i\neq j\leq M} \big\{u : u'(\theta_i - \theta_j) = 0\big\}$$

i.e., for a points u outside a finite number of hyperplanes, the inner products $u \cdot \theta_i$ are distinct for $1 \leq i \leq M$.

It's important to note that we can always find such a point u that satisfies these conditions. If we fail to find such a u, we would have to write $\mathbb{R}^n = \cup_{i=1}^{M'} H_i$, but this is not possible according to Baire's theorem, which implies that the countable union of closed sets with empty interior cannot cover the entire space.

Therefore, for values of u that do not lie on a finite number of hyperplanes, the pairs of real numbers $(u'\Delta_i u, u\theta_i)$, where $1 \leq i \leq M$, are distinct.

Suppose, without loss of generality, that $\Delta_1, \ldots, \Delta_k$ are the only distinct matrices among $\Delta_1, \ldots, \Delta_M$.

In the same way as before, for u not in a finite number of conics, the pairs of real numbers $(u'\Delta_i u, u\theta_i), 1 \leq i \leq k$ are distinct. Since the pairs (θ_i, Δ_i) are all distinct, the θ_i associated with the same Δ_i are different for $k < i \leq M$, and outside a finite number of conics, the corresponding numbers $u'\theta_i$ are distinct.

Consequently, for u not in a finite number of hyperplane and a finite number of conics, the pairs of real numbers $(u'\Delta_i u,\ u\theta_i),\ 1 \leq i \leq M$ are distinct.

For such a choice of u, Equation (4.24) asserts that the class of finite mixtures of one-dimensional normal distributions is not identifiable, which contradicts proposition 4. □

4.1.4 Elements of identifiability of mixtures of regression

4.1.4.1 Normal distributions

In the case of normal distribution we have

$$Y_{it} = f(a_{it}; \beta_k, \delta_k) + \epsilon_{itk} = \beta_k A_{it} + \delta_k W_{it} + \epsilon_{itk} \tag{4.25}$$

$$Y_i = \beta_k A_i + \delta_k W_i + \epsilon_{ik} \tag{4.26}$$

with $Y_i = (Y_{i1}, \cdots, Y_{iT})$, $A_i = (A_{i1}, \cdots, A_{iT})$, $W = (W_{i1}, \cdots, W_{iT})$ and $\epsilon_{ik} \sim \mathcal{N}(0; \sigma_k I_T)$. For a long description of these variables see 3.1.3.1.
Hence, we can express Y_i in the following manner for each cluster k:

$$Y_i \sim \mathcal{N}(\beta A_i + \delta W_i, \sigma I_T)$$

C. Hennig (2000) demonstrates the identifiability of clusterwise regression models when dealing with a one-dimensional normal distribution. We build upon this research by extending their findings to encompass multi-dimensional normal distributions. When considering fixed covariates, we denote

$$\mathcal{L}((Y_i)_{i\in I}) = \bigotimes_{i\in I} F_{A_i,W_i,J} \tag{4.27}$$

where $F_{A_i,W_i,J}(Y_i) = \int_{T_1} \Phi_{0,\Sigma}(Y_i - \beta A_i - \delta W_i)\, dJ\left(\beta, \delta, \sigma^2\right)$, $T_1 = \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \times \mathbb{R}_0^+$, $J \in \Omega_1 = \mathcal{J}(T_1)$, $\Sigma = \sigma I_T$.
$\mathcal{J}(T_1)$ denotes the set of mixing distributions with finite support on the parameter set T_1. $S(J)$ is the support set of $J \in \mathcal{J}(T_1)$. Thus, $s = |S(J)|$ is the number of mixture components. That is, the members of $\mathcal{J}(T_1)$ are distributions generating parameters values $(\beta_1, \delta_1, \sigma_1^2), \cdots, (\beta_s, \delta_s \sigma_s^2)$ for s clusters with probability $J(\beta_1, \delta_1, \sigma_1^2), \cdots, J(\beta_s, \delta_s, \sigma_s^2)$. I is some index set, here $I = 1, \cdots n$. $\bigotimes$ denotes the independent product of distributions.

4.1.4.1.1 Case $\beta_k A_i$

Let

$$\mathcal{C}_1 = \left(F_{A,J} \;:\; F_{A,J} = \bigotimes_{i\in I} F_{A_i,J}\right)_{J\in\Omega_1}$$

A denote the set of all $A_i,\ i \in I$.
In this context, identifiability means that given the data distributions $\mathcal{L}(Y_i)_{i\in I}$, we can uniquely determine the mixing distribution J, and there are no two distinct sets of parameters, such as $(\beta_{1i}, \sigma_{1i}^2, J(\beta_{1i}, \sigma_{1i}^2)), \cdots, (\beta_{si}, \sigma_{si}^2, J(\beta_{si}, \sigma_{si}^2)),\ i = 1, 2$, that result in the same data distributions.

Let $\mathcal{H}_{n-1}$ is the set of $(n-1)$-dimensional hyperplane.

Theorem 5. *Let* $h_j = \min\{q \;:\; \{A_{ij}, i \in I\} \subseteq \cup_{i=1}^{q} H_i \;\; Hi \in \mathcal{H}_{n-1}\}$.
If there exist some j *such that* $|S(J)| < h_j,\ \forall J$ *then* $\mathcal{C}_1$ *is identifiable.*

Proof. We need to show only $F_{A_i,J} = F_{A_i,\tilde{J}} \Rightarrow J = \tilde{J}$ because J contains all information to define the common distribution $F_{A_i,J}$ of $(Y_i)_{i\in I}$.

Let $F_{A_i,J} = F_{A_i,\tilde{J}}$ and $J \neq \tilde{J}$, without loss of generality we can suppose $|S(J)| \geq |S(\tilde{J}),|$. Thus there exist $(\beta_1, \sigma_1) \in S(\tilde{J})$ such as

$$J\{(\beta_1, \sigma_1^2)\} \neq \tilde{J}\{(\beta_1, \sigma_1^2)\} \tag{4.28}$$

$F_{A_i,J} = F_{A_i,\tilde{J}}$ implies the equality of the marginal Gaussian mixtures at all $A_i,\ i \in I$.

$$F_{A_i,J}(Y_i) = \int_{T_1} \phi_{\beta A_i,\Sigma}\left(Y_i\right) dJ\left(\beta, \sigma^2\right) \tag{4.29}$$

$$=F_{A_i,\tilde{J}}(Y_i) = \int_{T_1} \phi_{\beta A_i,\Sigma}\left(Y_i\right) d\tilde{J}\left(\beta, \sigma^2\right) \tag{4.30}$$

By identifiabilty of finite Gaussian mixtures, for $i \in I$:

$$J\left\{(\beta, \sigma^2) : \left(\beta A_i, \sigma^2\right) = \left(\beta_1 A_i, \sigma_1^2\right)\right\} = \tilde{J}\left\{(\tilde{\beta}, \tilde{\sigma}^2) : \left(\tilde{\beta} A_i, \tilde{\sigma}^2\right) = \left(\beta_1 A_i, \sigma_1^2\right)\right\} \tag{4.31}$$

The proof relies on the idea that the restriction to $|S(\tilde{J})|$ ensures the existence of a matrix, denoted as A_i, in such a way that the marginal mixture, $\mathcal{N}(\beta_1 A_i, \sigma_1^2)$, which is parameterized by J, cannot be accounted for by any combination of $(\tilde{\beta}, \tilde{\sigma}^2) \in S(\tilde{J})$ where $\tilde{\beta}$ is not equal to β_1. Consequently, it follows that $S(\tilde{J})$ must include the pair (β_1, σ_1^2).

Suppose that for all $(\beta, \sigma^2) \in S(J)$, and in particular for (β_1, σ_1^2), there exists $i(\beta) \in I$ such that

$$\forall (\tilde{\beta}, \tilde{\sigma}^2) \in S(\tilde{J})\ :\ \beta A_{i(\beta)} = \tilde{\beta} A_{i(\beta)} \Rightarrow \beta = \tilde{\beta} \tag{4.32}$$

For $i = i(\beta_1)$, tThe definition of $A_i = A_{i(\beta_1)}$ implies that

$$\forall S(\tilde{J}) \ni (\tilde{\beta}, \tilde{\sigma}^2) \neq (\beta_1, \sigma_1^2) :\ (\tilde{\beta} A_i, \tilde{\sigma}^2) \neq (\beta_1 A_i, \sigma_1^2) \tag{4.33}$$

Thus, using (4.30) and (4.31),

$$\tilde{J}\{(\beta_1, \sigma_1^2)\} = J\{(\beta, \sigma) :\ (\beta A_i, \sigma^2) = (\beta_1 A_i, \sigma_1^2)\} \tag{4.34}$$

$J\{(\beta, \sigma) :\ (\beta A_i, \sigma^2) = (\beta_1 A_i, \sigma_1^2)\} \neq 0$ because (β_1, σ_1^2) belong to it. Thus, $\tilde{J}\{(\beta_1, \sigma_1^2)\} \neq 0$ too. Therefore, equation (4.33) leads to the conclusion that $(\beta_1 A_i, \sigma_1^2) \in S(\tilde{J})$.

By (4.28), $\tilde{J}\{(\beta_1, \sigma_1^2)\} \neq J\{(\beta_1, \sigma_1^2)\}$ and remember (4.34), we can deduce

$$\exists S(J) \ni (\beta_2, \sigma_2^2) \neq (\beta_1, \sigma_1^2) :\ (\beta_2 A_i, \sigma_2^2) = (\beta_1 A_i, \sigma_1^2) \tag{4.35}$$

Consider $A_i = A_{i(\beta_2)}$ and apply the arguments above again to get $(\beta_2 A_i, \sigma_2^2) \in S(\tilde{J})$. This results leads a contradiction between (4.33) and (4.35) .
Indeed, $(\beta_2, \sigma_2^2) \in S(\tilde{J})$ and $(\beta_2, \sigma_2^2) \neq (\beta_1, \sigma_1^2)$. By (4.33), $(\beta_2 A_i, \sigma_2^2) \neq (\beta_1 A_i, \sigma_1^2)$ and by (4.35), $(\beta_2 A_i, \sigma_2^2) = (\beta_1 A_i, \sigma_1^2)$.
Thus there exist some $(\beta, \sigma) \in S(J)$ such that $\forall i \in I \; \forall (\tilde{\beta}, \tilde{\sigma}^2) \in S(\tilde{J}) \; : \; \beta A_i = \tilde{\beta} A_i \Rightarrow \beta \neq \tilde{\beta}$.
Thus,

$$\{A_{ij} \; : \; i \in I, \; j = 1 \cdots T\} \subset \cup_{(\tilde{\beta}, \tilde{\sigma}^2) : \tilde{\beta} \neq \beta} \{x : \; \beta x = \tilde{\beta} x\} \tag{4.36}$$

Therefore $\cup_{(\tilde{\beta}, \tilde{\sigma}^2) : \tilde{\beta} \neq \beta} \{x : \; \beta x = \tilde{\beta} x\}$ is composed by $|S(\tilde{J})|$ different hyperplanes.
So for $j = 1 \cdots T$, $h_j \leq |S(\tilde{J})| \leq |S(J)|$. □

4.1.4.1.2 Using trajectory for $\beta_k A_i$

In LCGA, we possess knowledge about the trajectory's shape, which is specifically linked to the passage of time. Consequently, the explanatory covariate and the trajectory's shape contain a greater amount of information than what is considered in the Hennig theorem.

Let

$$\mathcal{C}_2 = \left(F_{A,J} \; : \; F_{A,J} = \bigotimes_{i \in I} F_{A_i, W_i, J} \right)_{J \in \Omega_1} \tag{4.37}$$

$$\mathcal{C}_{2A} = \left(F_{A,J} \; : \; F_{A,J} = \bigotimes_{i \in I} F_{A_i, J} \right)_{J \in \Omega_1} \tag{4.38}$$

$$\mathcal{C}_{2W} = \left(F_{A,J} \; : \; F_{A,J} = \bigotimes_{i \in I} F_{W_i, J} \right)_{J \in \Omega_1} \tag{4.39}$$

Proposition 6. *$\mathcal{C}_{2A}$ is identifiable if and only if $d_k < T$ for all $1 \leq k \leq K$ and a_{it} are all distinct, for all $i \in I$ and $1 \leq t \leq T$.*

Proof. We need to show $F_{A_i,J} = F_{A_i,\tilde{J}} \Leftrightarrow J = \tilde{J}$.

$$F_{A_i,J}(Y_i) = \int_{T_1} \Phi_{\beta_k A_i, \Sigma}(Y_i) \, dJ\left(\beta, \sigma^2\right) \tag{4.40}$$

$$= F_{A_i,\tilde{J}}(Y_i) = \int_{T_1} \Phi_{\beta_k A_i, \Sigma}(Y_i) \, d\tilde{J}\left(\beta, \sigma^2\right) \tag{4.41}$$

By identifiabilty of finite Gaussian mixtures, the equality above is equivalent, for $i \in I$, to:

$$J\left\{(\beta, \sigma^2) : \left(\beta A_i, \sigma^2\right) = \left(\mu_1, \sigma_1^2\right)\right\} = \tilde{J}\left\{(\tilde{\beta}, \tilde{\sigma}^2) : \left(\tilde{\beta} A_i, \tilde{\sigma}^2\right) = \left(\mu_1, \sigma_1^2\right)\right\} \tag{4.42}$$

for some (μ_1, σ_1).

Assume there exist $\tilde{\beta}$ in $S(\tilde{J})$ such that $\beta A_i = \tilde{\beta} A_i$ for some $\beta \in S(J)$, that is for $1 \leq t \leq T$, $\beta A_{it} = \tilde{\beta} A_{it}$ but $\tilde{\beta} \neq \beta$ for all $\beta \in S(J)$.

If $d_k < T$ and a_{it} are all different, we have 2 different polynomials there are equal on T points. The unisolvence theorem states that when we have a set of $n+1$ distinct points, namely $x_0, x_1, \ldots, x_n$, along with their corresponding values $y_0, y_1, \ldots, y_n$, there is a single polynomial, of degree no greater than n, that can precisely pass through all these data points, forming an interpolation of the dataset $\{(x_0, y_0), \ldots, (x_n, y_n)\}$.

Consequently, for any set of T points, there exists a unique polynomial with a degree no greater than $T-1$ that can interpolate the given data.

Thus $\beta = \tilde{\beta}$. □

If we know the membership of cluster for each value Y_i, we can write .

$$Y_k = A_k \beta_k^t$$

where $Y_k = \begin{pmatrix} Y_{k11} \\ \vdots \\ Y_{kn_kT} \end{pmatrix}$ and $A_k = \begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_kT} & \cdots & a_{n_kT}^{d_k-1} \end{pmatrix}$.

If $d_k < T$, which is a requirement for the matrix to be invertible, then, given that the matrix is a Vandermonde matrix in the form of: $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_kT} & \cdots & a_{n_kT}^{d_k-1} \end{pmatrix}$ the condition for invertibility is ensured if all the values of a_{it} are distinct. We can write

$$\beta_k = Y_k \left(A_k^t A_k\right)^{-1} A_k \tag{4.43}$$

Proposition 7. *If for all $1 \leq t, t' \leq T$ and for $i, j \in I$, $a_{it} = a_{jt}$ and $a_{it} \neq a_{it'}$, $\mathcal{C}_{2A}$ is identifiable if and only if the degree of the polynomial shape is strictly inferior to the number of time values for all clusters.*

$\mathcal{C}_{2A}$ is identifiable if and only if $d_k < T$ for all $1 \leq k \leq K$.

Proof. In this case the matrix $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{nT} & \cdots & a_{nT}^{d_k-1} \end{pmatrix}$ become $\begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{1T} & \cdots & a_{1T}^{d_k-1} \end{pmatrix}$ and it is invertible if $d_k < T$ and $a_{it} \neq a_{it'}$, $1 \leq t, t' \leq T$. □

4.1.4.1.3 Numerical example

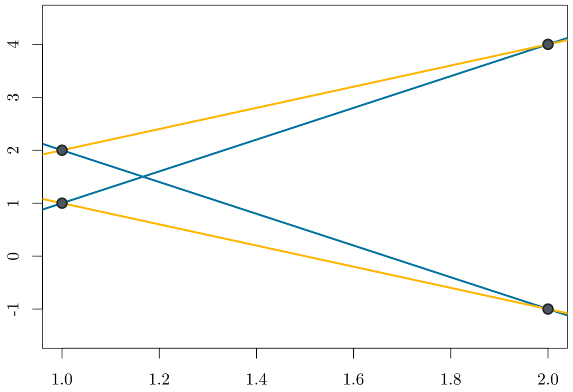


Figure 4.1: multiple line regression passing by two points.

To demonstrate the theorem mentioned above, we generated synthetic data with varying parameters. In all cases, we intentionally selected coefficients in such a way that the resulting clusters are well-separated to avoid issues related to separation.

In each group, we set $\sigma_k = 0.1$ and used two clusters with equal probabilities $(0.5$ each$)$. Each mixture distribution was defined with two data points, specifically $(1, 2)$, and we manipulated the degree of the polynomial that shapes the trajectory.

As per Teicher's results, we understand that the mixture distribution $\frac{1}{2}\mathcal{N}\left(\mu_{1t};\sigma_1\right)+\frac{1}{2}\mathcal{N}\left(\mu_{2t};\sigma_2\right)$ is identifiable, meaning we can determine unique values for (μ_{1t}, σ_1) and (μ_{2t}, σ_2).

Nonetheless, when attempting to express μ_1 and μ_2 in terms of a polynomial with time as the variable, it's important to note that the coefficients of this polynomial may not always be uniquely identifiable. For example, if we suppose that for the two clusters:

- $t = 1,\ a_{11} = 1, t = 2,\ a_{12} = -1$ and $\sigma_1 = 0.11$;
- $t = 1,\ a_{21} = 2, t = 2,\ a_{22} = 4$ and $\sigma_2 = 0.11$;

and if we want write $\mu_{kt} = \beta_k A_t$ for all k, t, we have two solutions

- $\beta_1 = (5, -3)$ or $\beta_2 = (-2, 3)$;
- $\beta_1 = (3, -2)$ or $\beta_2 = (0, 2)$.

The mixture is not identifiable.

However, if we have prior knowledge that the values follow a linear trajectory, we can distinguish which of the two solutions is correct. This is because two points uniquely determine a straight line, making the coefficients identifiable, and consequently, the mixture distribution as well.

On the other hand, if we desire a polynomial of a higher degree, for example, 2, the mixture distribution typically remains unidentifiable. In the right part of Figure 4.2, it's evident

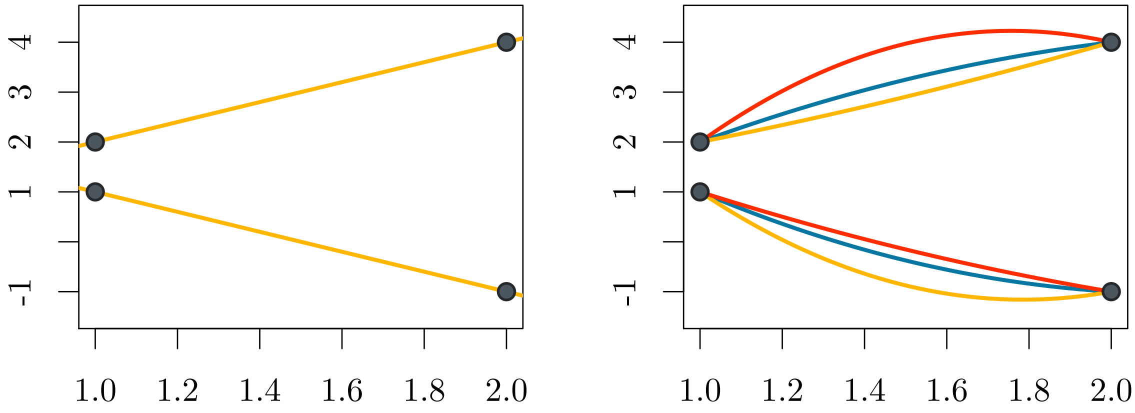


Figure 4.2: Line and 2 degree polynomials passing through 2 points.

that only one solution exists, while in the left part, multiple degree-2 polynomials can pass through the data points, rendering the mixture distribution non-identifiable. To demonstrate the algorithm's ability to illustrate this issue, we generated 50 samples, each containing 100 observations from the mixture models with parameters, in addition to the ones previously defined.

- $\beta_1 = (3, -2)$ or $\beta_2 = (0, 2)$ for the line version ;
- $\beta_1 = (10, -12.5, 3.5)$ or $\beta_2 = (-2, 5, -1)$ for the degree 2 polynomial version.

In each sample, a mixture with two clusters and equal standard deviation is fitted using the | trajeR| algorithm. To mitigate the risk of getting stuck in local maxima, we run the algorithm five times with different random initializations and report the best solution obtained. To prevent degeneracy, we remove the lowest and highest 5 % of values for each dimension.

To investigate the estimated parameters, we employ parallel coordinate plots (as described by Wegman (1990)). In the case of an identifiable two-mixture model, this type of graphic should reveal two distinct bundles of lines or data clusters. In the linear model (the left part), we observe two distinct bundles for β_0 and β_1, and a single bundle for π and σ. However, in the second-degree model, we only see one large bundle with a wide range of variability for the parameters β_1 and β_2.

4.1.4.1.4 Using trajectory for $\beta_k A_i + \delta_k W_i$

$$\text{Let } Y_k = \begin{pmatrix} Y_{k11} \\ \vdots \\ Y_{kn_kT} \end{pmatrix} \text{ and } A_k = \begin{pmatrix} 1 & a_{11} & \cdots & a_{11}^{d_k-1} \\ \vdots & & & \vdots \\ 1 & a_{n_kT} & \cdots & a_{n_kT}^{d_k-1} \end{pmatrix} \text{ and } W_k = \begin{pmatrix} w_{11} & \cdots & w_{11}^{n_\delta} & \\ \vdots & & & \vdots \\ w_{n_kT} & \cdots & w_{n_kT}^{n_\delta} & \end{pmatrix}.$$

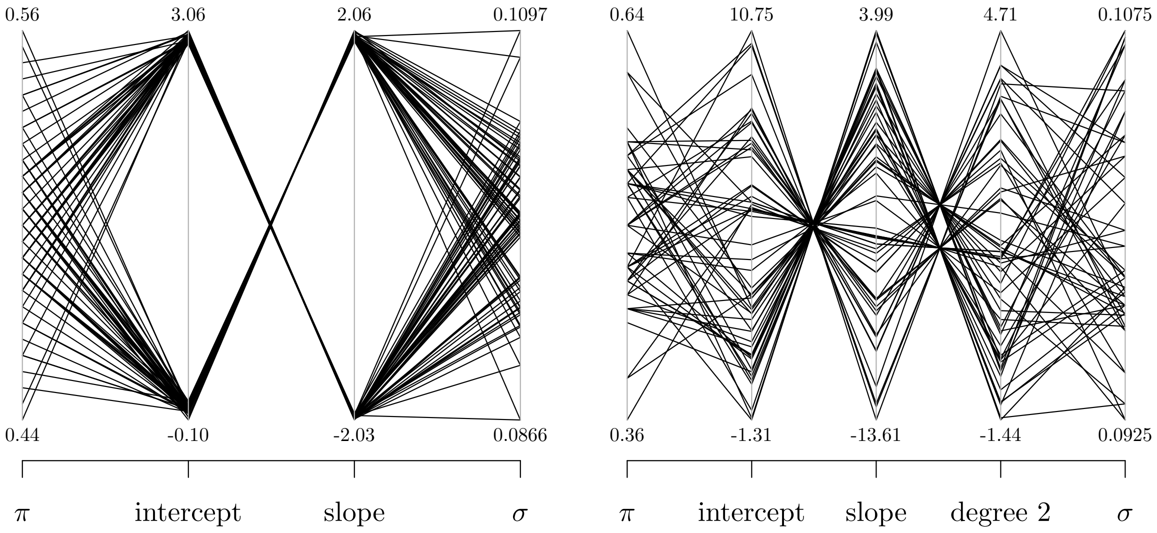


Figure 4.3: Parallel coordinate plots: CNORM
Parallel coordinate plots of the estimated parameters for 50 samples of the artificial example with different degree of the polynomial shape of trajectory.

Thus,

$$Y_k = A_k \beta_k^t + W_k \delta_k^t$$

Proposition 8. *$\mathcal{C}_2$ is identifiable if*

- *$d_k < T$ for all $1 \leq k \leq K$ and a_{it} are all distinct, for all $i \in I$ and $1 \leq t \leq T$*
- *there exist t such that W_t is full rank ;*
- *$rk(A_t, W_t) = rk(A_t) + rk(W_t)$ where $rk(\cdot)$ determines the rank of a matrix.*

Proof. The complete model is

$$Y_i = \beta_k A_i + \delta_k W_i + \epsilon_i$$

If, for all $i \in I$ and $1 \leq t \leq T$, $w_{it} = w_i$ is time-independent, the model simplifies to: $y_{it} = \beta_k A_{it} + \delta_k(w_i, \cdots, w_i) + \epsilon_{it}$.
In this case, the model represents a translation of the polynomial shape for all clusters k. Consequently, the parameters of the polynomial are identifiable, and since the rank of the matrix A_t, W is equal to the sum of the ranks of A_t and W (i.e., $rk(A_t, W) = rk(A_t) + rk(W)$), we can determine the value of δ_k as well.

Suppose that $d_k < nT$ for all $1 \leq k \leq K$. Thus for a integer c a mixture of c component of form $\sum_{k=1}^{c} \pi_k \mathcal{N}\left(\beta_k A_{it}, \sigma_k\right)$ is identifiable.
If we know the membership of each Y_i, we can write like in equation (4.43), $\beta_k = Y_k \left(A_k^t A_k\right)^{-1} A_k$.

We note $P_k = A_k^t \left(A_k A_k^t\right)^{-1} A_k$ and $R_k = I - P_k$.

$$\begin{aligned}
\beta_k A_k + \delta_k W_k &= \beta_k A_k + \delta_k W_k P_k + \delta_k W_k \left(I - P_k\right) && (4.44)\\
&= \beta_k A_k + \delta_k W_k A_k^t \left(A_k A_k^t\right)^{-1} A_k + \delta_k W_k \left(I - P_k\right) && (4.45)\\
&= \left(\beta_k + \delta_k W_k A_k^t \left(A_k A_k^t\right)^{-1}\right) A_k + \delta_k W_k R_k && (4.46)\\
&= \left(\beta_k + \delta_k W_k A_k^t \left(A_k A_k^t\right)^{-1} \quad \delta_k\right) \begin{pmatrix} A_k \\ W_k R_k \end{pmatrix} && (4.47)\\
&= \lambda_k V && (4.48)
\end{aligned}$$

Suppose $\lambda_k V = 0$ for some λ_k. This give $\beta_k A_k + \delta_k W_k = 0$, with both $\beta_k = \delta_k = 0$ by linear independence of the columns of A_k and W_k. Hence V is a full rank $T + rk(W)$ matrix .
Since $Y_k = \lambda_k V + \epsilon$, we have

$$\begin{aligned}
\hat{\lambda}_k &= Y_k \left(V V^t\right)^{-1} V^t && (4.49)\\
&= \left(A_k \quad W_k R_k\right) \begin{pmatrix} A_k A_k^t & A_k R_k^t W_k^t \\ W_k R_k A_k^t & W_k R_k R_k^t W_k^t \end{pmatrix}^{-1} && (4.50)\\
&= \left(A_k \quad W_k R_k\right) \begin{pmatrix} A_k A_k^t & A_k R_k W_k^t \\ W_k R_k A_k^t & W_k R_k R_k W_k^t \end{pmatrix}^{-1} && (4.51)
\end{aligned}$$

From the definition $R_k = I - A_k^t \left(A_k A_k^t\right)^{-1} A_k$, we have $A_k R_k = R_k A_k^t = 0$ and $P_k^2 = P_k$.

$$\begin{aligned}
\hat{\lambda}_k &= Y_k \left(V V^t\right)^{-1} V^t && (4.52)\\
&= Y_k \left(A_k \quad W_k R_k\right) \begin{pmatrix} A_k A_k^t & 0 \\ 0 & W_k R_k W_k^t \end{pmatrix}^{-1} && (4.53)\\
&= Y_k \left(A_k \left(A_k A_k^t\right)^{-1} \quad W_k R_k \left(W_k R_k W_k^t\right)^{-1}\right) && (4.54)\\
&= \left(Y_k A_k \left(A_k A_k^t\right)^{-1} \quad Y_k W_k R_k \left(W_k R_k W_k^t\right)^{-1}\right) && (4.55)
\end{aligned}$$

From the right part we have

$$\hat{\delta}_k = Y_k W_k R_k \left(W_k R_k W_k^t\right)^{-1}$$

and from the left one

$$\hat{\beta}_k = Y_k A_k \left(A_k A_k^t\right)^{-1} - \hat{\delta}_k W_k A_k^t \left(A_k A_k^t\right)^{-1}$$

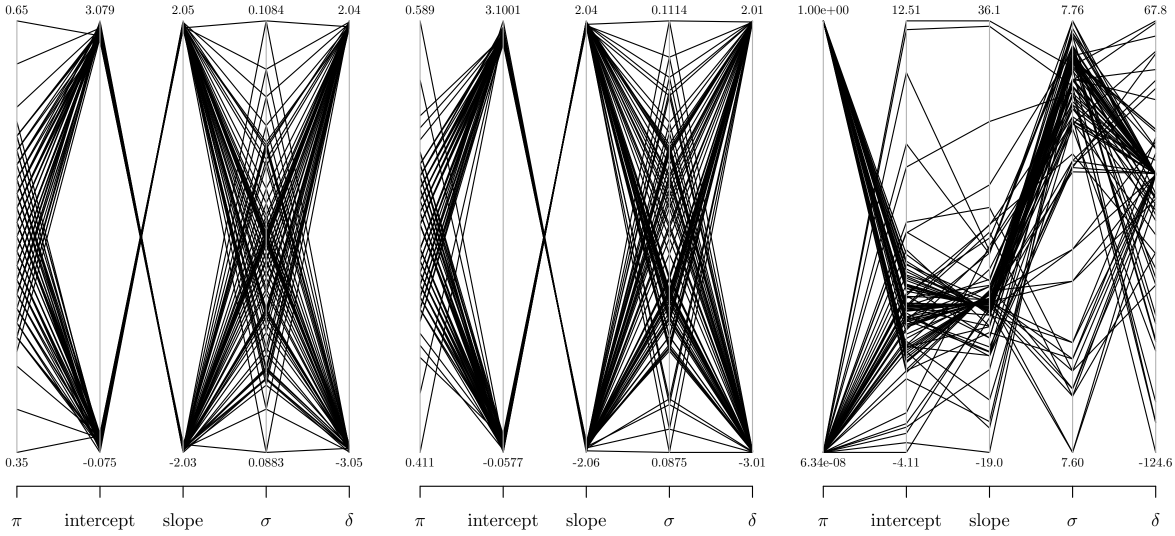


Figure 4.4: Parallel coordinate plots: CNORM.
Parallel coordinate plots of the estimated parameters for 50 samples of the artificial example with different degree of the polynomial shape of trajectory.

.

It proves the identifiability of the parameters. □

4.1.4.1.5 Numerical example

To illustrate the theorem mentioned above, we generated synthetic data with various parameters. In each case, we intentionally selected coefficients to create well-separated clusters and avoid separation problems.

In all scenarios, we maintained $\sigma_k = 0.1$ and used two clusters with equal probabilities (0.5 each). The parameter values were set as follows: $\beta_1 = (3, -2)$, $\beta_2 = (0, 2)$, $\delta_1 = 2$, and $\delta_2 = -3$. For each example, we generated 50 samples, each consisting of 100 observations from the mixture models with a time covariate and the specified parameters.

- time covariate is independent of time and it value is 1 or 0;

- time covariate is dependent of time but linearly independent;

- time covariate is dependent of time but linearly dependent;

As depicted in Figure 4.4, the two left graphs display two distinct bundles of parameters, indicating that the mixture is identifiable. However, in the right graph, the linear dependence on the time covariate leads to non-identifiability, with wide parameter value ranges for β_k, σ_k, and δ_k.

4.1.4.2 Logistic model

In the case of the logistic model, we introduce a latent variable, denoted as y_{it}^{*}, such that

$$y_{it}^{*} = \beta_k A_{it} + \delta_k W_{it} + \epsilon_{it} \tag{4.56}$$

where $A_{it} = (1, a_{it}, a_{it}^{2}, \cdots, a_{it}^{n_\beta - 1})^t$, $W_{it} = (w_{it}^{1}, \cdots, w_{it}^{n_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ and $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$.
Based on this assertion, it is typically assumed that the binary variable y_{it} takes the value 1 if $y_{it}^{*} > 0$ and 0 if $y_{it}^{*} \leq 0$.

Let $\rho_{ikt} = P(Y_{it} = 1 | W_i = w_i, C_i = k)$ the probability of $y_{it} = 1$ given membership in group k.

$$\rho_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \tag{4.57}$$

We call $\mathcal{H}_C$ the set of discrete distributions on $\mathcal{H}$ with at most C atoms.

$$\mathcal{H}_C = \left\{ H(x) : H(x) = \sum_{i=1}^{c} c_i f(x, \beta_i, \delta_i),\ c_i > 0, \sum_{i=1}^{c} c_i = 1,\ c_i \in [\![0; C]\!] \right\} \tag{4.58}$$

where f is cdf of a mutlivariate Bernoulli distribution.

The paper by Blischke (1964) demonstrates that a finite mixture of K binomial distributions, each constructed from n independent repetitions of a Bernoulli experiment, is identifiable if and only if $n \geq 2K - 1$. Notably, since a univariate Bernoulli distribution corresponds to a binomial distribution with $n = 1$, this finding implies that there are no identifiable mixtures of univariate Bernoulli distributions.

The study by Gyllenberg et al. (1994) establishes that there are no identifiable mixtures of multivariate Bernoulli distributions.

Consequently, this implies that there are no identifiable mixtures of multivariate Bernoulli distributions for regression, and as a result, $\mathcal{H}_C$ is not identifiable.

4.1.4.2.1 Numerical example

To illustrate the theorem mentioned above, we generated synthetic data with various parameters. In each case, we intentionally selected coefficients to create well-separated clusters and avoid separation problems.

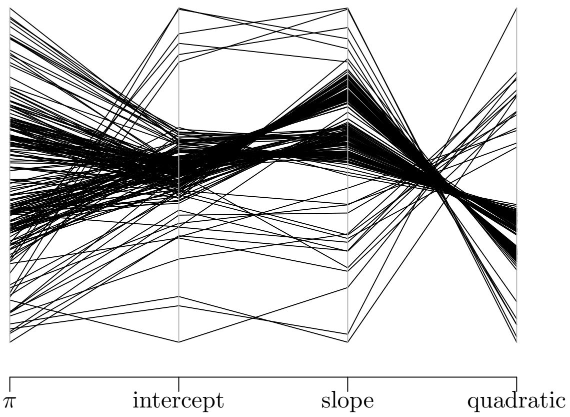


Figure 4.5: Parallel coordinate plots: logistic.
Parallel coordinate plots of the estimated parameters for 100 samples of the artificial example.

In all scenarios, we used two clusters with equal probabilities (0.5 each) and set the parameter values as follows: $\beta_1 = (-3.5, 5, -1.2)$ and $\beta_2 = (-3, 0.45, 0.2)$. For each example, we generated 100 samples, each consisting of 100 observations from the mixture models.

As shown in Figure 4.5, the covariate β_k is not identifiable, with a wide range of possible values, confirming the non-identifiability of the mixture.

4.1.4.3 ZIP model

In the case of ZIP model, we confider

$$P\left(Y_{it} = y_{it} | W_i = wi, C_i = ci\right) = \begin{cases} \rho_{ikt} + (1-\rho_{ikt})e^{-\lambda_{ikt}}, \ y_{it} = 0 \\ (1-\rho_{ikt})\frac{\lambda_{ikt}^{y_{it}} e^{-\lambda_{ikt}}}{y_{it}!}, \ y_{it} > 0 \end{cases} \tag{4.59}$$

where $\log\left(\lambda_{ikt}\right) = \beta_k A_{it} + \delta_k W_{it}$ and $\log\left(\frac{\rho_{ikt}}{1-\rho_{ikt}}\right) = \nu_k A_{it}$ with $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{n_\beta - 1})^t$, $W_{it} = (w_{it}^1, \cdots, w_{it}^{n_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kn_\beta})$ et $\delta_k = (\delta_{k1}, \cdots, \delta_{kn_\delta})$.

The paper by C.-S. Li (2012) demonstrates that, under certain conditions, the ZIP model is identifiable. If we consider the ZIP model as follows:

$$f(y; p(x), \lambda(x), x) = (1-p(x))\mathbb{1}_{\{y=0\}} + p(x)\frac{e^{-\lambda(x)}\lambda(x)^y}{y!} \tag{4.60}$$

for $y = 0, 1, \ldots$ and x a continuous covariate in the design space $[a_0, a_1]$ where $-\infty < a_0 < a_1 < +\infty$.

It has been proven that if the probability function $p(x)$ is linked via the logit function to a smooth function, and the Poisson parameter $\lambda(x)$ is linked via the natural logarithm to a smooth function, then the ZIP model is identifiable in the sense that: $f(y;p(x),\lambda(x),x) = f(y;p^*(x),\lambda^*(x),x)$ if and only if $p(x) = p^*(x)$ and $\lambda(x) = \lambda^*(x)$ for all x within the design space, and for all values of $y = 0, 1, \ldots$.

Theorem 6 (Li). *Let x be a continuous covariate in the design space $\chi = [a_0, a_1]$ here $-\infty < a_0 < a_1 < +\infty$. The ZIP model is identifiable if*

1. *$p(x)$ is specified as $logit[p(x)] = \log\left[p(x)/(1-p(x))\right] = \beta_0 + \beta_1 x$ and $\lambda(x)$ is specified as $\log[\lambda(x)] = \alpha_0 + \alpha_1 x$.*

2. *$p(x)$ is specified as $logit[p(x)] = \log\left[p(x)/(1-p(x))\right] = \beta_0 + \beta_1 x$ and $\log[\lambda(x)] = s(x)$, where $s(x)$ is an unspecified smooth function instead of the linear form $\alpha_0 + \alpha_1 x$.*

3. *$logit[p(x)] = r(x)$ where $r(x)$, is an unspecified smooth function instead of the linear form $\beta_0 + \beta_1 x$ and $\lambda(x)$ is specified as $\log[\lambda(x)] = \alpha_0 + \alpha_1 x$.*

4. *$logit[p(x)] = r(x)$, where $r(x)$ is an unspecified smooth function instead of the linear form $\beta_0 + \beta_1 x$ and $\log[\lambda(x)] = s(x)$, where $s(x)$ is an unspecified smooth function instead of the linear form $\alpha_0 + \alpha_1 x$.*

Proof. We need to show that $f(y;p(x),\lambda(x),x) = f(y;p^*(x),\lambda^*(x),x)$ if and only if $p(x) = p^*(x)$ and $\lambda(x) = \lambda^*(x)$ for $x \in \chi$, the design space and $y = 0, 1, \ldots$.

The logic from left to right is indeed self-evident; therefore, we must demonstrate the second part of the statement.

Assume that $f(y;p(x),\lambda(x),x) = f(y;p^*(x),\lambda^*(x),x)$. Using equation (4.60), with some algebra,

$$\frac{p(x)}{p^*(x)} = \frac{\mathbb{1}_{\{y=0\}} - e^{-\lambda^*(x)}\left[\lambda^*(x)\right]^y/y!}{\mathbb{1}_{\{y=0\}} - e^{-\lambda(x)}\left[\lambda(x)\right]^y/y!} \tag{4.61}$$

The left-hand side of the ratio depends on x, while the right-hand side depends on both x and y. Therefore, the ratio can be represented as a positive function of x, denoted by $c(x)$, for $x \in \chi$. The values of $p^*(x)$ and $\lambda^*(x)$ must satisfy the following equations:

$$p^*(x) = \frac{p(x)}{c(x)} \tag{4.62}$$

$$e^{-\lambda^*(x)}\left[\lambda^*(x)\right]^y = (1-c(x))y!\mathbb{1}_{\{y=0\}} + c(x)e^{-\lambda(x)}\left[\lambda(x)\right]^y, \quad y = 0, 1, \ldots \tag{4.63}$$

For $y = 1$, we have $e^{-\lambda^*(x)}\lambda^*(x) = c(x)e^{-\lambda(x)}\lambda(x)$

For $y = 2$, we have $e^{-\lambda^*(x)}\left[\lambda^*(x)\right]^2 = c(x)e^{-\lambda(x)}\left[\lambda(x)\right]^2$

Thus, $\lambda^*(x) = \lambda(x)$ for $x \in \chi$ and hence $c(x) = 1$ and $p^*(x) = p(x)$.

The other proof are the same. □

We call $\mathcal{H}_Z$ the set of discrete ZIP distributions on $\mathcal{H}$ with at most C atoms in which all Poisson parameters are different.

$$\mathcal{H}_Z = \left\{ H(x) : H(x) = \sum_{i=1}^{c} c_i F(x, \pi_i, \lambda_i),\ c_i > 0, \sum_{i=1}^{c} c_i = 1,\ c \in [\![0; C]\!], \lambda_i \neq \lambda_j\ \forall i \neq j \right\} \tag{4.64}$$

where F is cdf of a ZIP distribution with pdf

$$f(y; \pi_i, \lambda_i) = (1 - \pi_i)\mathbb{1}_{\{y=0\}} + \pi_i \frac{e^{-\lambda_i} \lambda_i^y}{y!}$$

Proposition 9. *Under the hypotheses of theorem 6, the class $\mathcal{H}_\mathcal{Z}$ of all finite mixtures of Zero Inflated Poisson distributions is identifiable.*

When we use mixtures of ZIP distributions with the same Poisson parameter λ, the model becomes non-identifiable. For instance, the following mixtures can produce the same distribution:

$$\mathcal{M}_1 :\ \{(1; f(\bullet; 1/2; \lambda)))\} \tag{4.65}$$

$$\mathcal{M}_2 :\ \{(1/2; f(\bullet; 1/4; \lambda)))\,,\ (1/2; f(\bullet; 3/4; \lambda)))\} \tag{4.66}$$

$$\mathcal{M}_3 :\ \{(1/2; f(\bullet; 1/6; \lambda)))\,,\ (1/2; f(\bullet; 5/6; \lambda)))\} \tag{4.67}$$

Lemma 1. *The moment generating function M_X of variable X of a ZIP distribution of parameters (π, λ) is*

$$M_X(t) = 1 - \pi + \pi e^{\lambda(e^t - 1)}$$

Proof. We write M_X the moment generating function of variable X of a ZIP distribution of parameters (π, λ).

$$M_X(t) = E\left(e^{tX}\right) \tag{4.68}$$

$$= \sum_{k \geq 0} P(X = k) e^{tk} \tag{4.69}$$

$$= P(X = 0) + \sum_{k > 0} P(X = k) e^{tk} \tag{4.70}$$

$$= 1 - \pi + \pi e^{-\lambda} + \sum_{k>0} \pi \frac{e^{-\lambda} \lambda^k}{k!} e^{tk} \tag{4.71}$$

$$= 1 - \pi + \pi e^{-\lambda} + \pi e^{-\lambda} \sum_{k>0} \frac{(\lambda e^t)^k}{k!} \tag{4.72}$$

$$= 1 - \pi + \pi e^{-\lambda} + \pi e^{-\lambda} \left(e^{\lambda e^t} - 1 \right) \tag{4.73}$$

$$= 1 - \pi + \pi e^{\lambda(e^t-1)} \tag{4.74}$$

□

Lemma 2. *Let* $\phi(t;\lambda,\pi) = 1 - \pi + \pi e^{\lambda(e^t-1)}$ *and* $\lambda_1, \lambda_2 \in \mathbb{N}^*$. *For* $\lambda_2 > \lambda_1$,

$$\lim_{t\to+\infty} \frac{\phi(t;\lambda_1,\pi_1)}{\phi(t;\lambda_2,\pi_2)} = 0$$

Proof.

$$\frac{\phi(t;\lambda_1,\pi_1)}{\phi(t;\lambda_2,\pi_2)} = \frac{1-\pi_1+\pi_1 e^{\lambda_1(e^t-1)}}{1-\pi_2+\pi_2 e^{\lambda_2(e^t-1)}} \tag{4.75}$$

$$= e^{(\lambda_1-\lambda_2)(e^t-1)} \frac{(1-\pi_1)e^{-\lambda_1(e^t-1)}+\pi_1}{(1-\pi_2)e^{-\lambda_2(e^t-1)}+\pi_2} \tag{4.76}$$

Since $\lambda_1, \lambda_2 > 0$ and $\lambda_1 - \lambda_2 < 0$, $\lim\limits_{t\to+\infty} e^{(\lambda_1-\lambda_2)(e^t-1)} = 0$ and $\lim\limits_{t\to+\infty} \frac{(1-\pi_1)e^{-\lambda_1(e^t-1)}+\pi_1}{(1-\pi_2)e^{-\lambda_2(e^t-1)}+\pi_2} = \frac{\pi_1}{\pi_2}$. □

Proof. (of proposition).

Let M be the mapping that transforms a ZIP distribution with cumulative distribution function F into its moment generating function, denoted as M_F. This function is defined as $\phi(t;\lambda,\pi) = 1-\pi+\pi e^{\lambda(e^t-1)}$. It's worth noting that M is a linear mapping.
We order the family lexicographically by : $F_1(\bullet,\pi_1,\lambda_1) < F_2(\bullet,\pi_2,\lambda_2)$ if $\lambda_1 > \lambda_2$. This induce a total order on the family $\mathcal{H}_Z$.
Let S_ϕ the support of ϕ, we have $S_\phi =]0,+\infty[$ for each ZIP distribution.
For $F_1(\bullet,\pi_1,\lambda_1) < F_2(\bullet,\pi_2,\lambda_2)$ we have

$$\lim_{t\to+\infty} \frac{\phi_2(t;\pi_2,\lambda_2)}{\phi_1(t;\pi_1,\lambda_1)} = 0 \tag{4.77}$$

Hence, for $t_1 = +\infty$ we can use the theorem 3 to prove the proposition. □

Remark: The identifiability of a Zero-Inflated Poisson (ZIP) distribution can be established by applying the proof of Teicher's theorem 3.

4.1.4.3.1 Numerical example

To illustrate the theorem mentioned above, we generated artificial data with various parameters. We considered two cases: the first case involved two clusters with different Poisson parameters, and the second case involved two clusters with the same Poisson parameter. In all scenarios, we used two clusters with equal probabilities (0.5), and

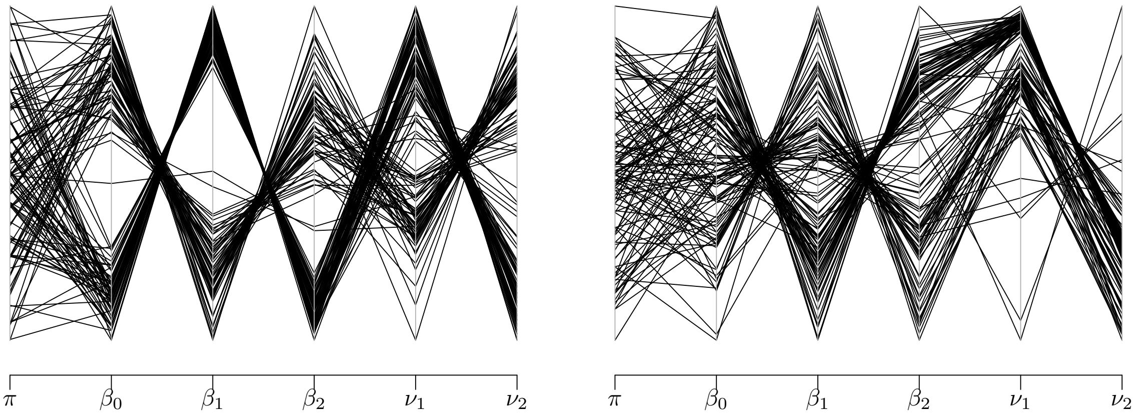


Figure 4.6: Parallel coordinate plots: ZIP
parallel coordinate plots of the estimated parameters for 100 samples of the artificial example.

- for the first
  - $\beta_1 = (2.331, -2.275, 0.4)$ and $\beta_2 = (-1.639, 3.461, -1.142)$;
  - $\nu_1 = (1, -0.75)$ and $\nu_2 = (-1, 0.25)$.
- for the second
  - $\beta_1 = \beta_2 = (2.331, -2.275, 0.4)$;
  - $\nu_1 = (1, -0.75)$ and $\nu_2 = (-11.6, 3.4)$.

For each example, we generated a sample of 500 observations for each mixture model and fitted the model using a polynomial of degree 2 for the polynomial part and 1 for the exceeded zero state. We randomly generated starting points and fit the model, excluding solutions with only one component.

As depicted in Figure 4.6, the two left parts reveal identifiable parameters, while the right part lacks identifiability, with only one bundle for the ν parameters.

4.1.4.3.2 Using trajectory for $\beta_k A_i$

Let

$$\mathcal{L}\left((Y_i)_{i\in I}\right) = \bigotimes_{i\in I} F_{A_i,J} \tag{4.78}$$

and

$$\mathcal{C}_A = \left(F_{A,J} \;:\; F_{A,J} = \bigotimes_{i\in I} F_{A_i,J}\right)_{J\in\Omega_1} \tag{4.79}$$

where $F_{A_i,J}(Y_i) = \int_{T_1} \prod_{t=1}^{T} F_{\lambda_{ikt},\rho_{ikt}}(Y_i)\, dJ(\nu,\beta)$, J define like in page 152 and $F_{\lambda_{ikt},\rho_{ikt}}$ is cdf of a vector of T variable of ZIP distribution with parameters such that $\log(\lambda_{ikt}) = \beta_k A_{it} + \delta_k W_{it}$ and $\log\left(\frac{\rho_{ikt}}{1-\rho_{ikt}}\right) = \nu_k A_{it}$ with $A_{it} = (1, a_{it}, a_{it}^2, \cdots, a_{it}^{d_{\beta_k}-1})^t$, $W_{it} = (w_{it}^1, \cdots, w_{it}^{n_\delta})^t$, $\beta_k = (\beta_{k1}, \cdots, \beta_{kd_{\beta_k}})$ et $\delta_k = (\delta_{k1}, \cdots, \delta_{kd_{\delta_k}})$.

Thus, $\mathcal{C}_A$ is the set of the mixture of product of ZIP distributions and a element $F_{A_i,J}$ can be rewrite like

$$\sum_{k=1}^{K} c_k \prod_{t=1}^{T} F_{\lambda_{ikt},\rho_{ikt}}(y_{it}; \nu_k, \beta_k),\ c_i > 0, \sum_{k=1}^{K} c_k = 1 \tag{4.80}$$

Proposition 10. *If $d_{\beta_k} < T$ and $d_{\nu_k} < T$ for all $1 \leq k \leq K$, a_{it} are all distinct, for all $i \in I, 1 \leq t \leq T$, and the hypotheses of theorem 6 hold then $\mathcal{C}_A$ is identifiable.*

Proof. By theorem 4,

$$\mathcal{H}_p = \left\{ H(x) : H(x) = \sum_{i=1}^{c} c_i \prod_{t=1}^{T} F(x_t; \lambda_{it}, \rho_{it}),\ c_i > 0, \sum_{i=1}^{c} c_i = 1, x = (x_1, \ldots, x_T),\ c \in [\![0; C]\!] \right\}$$

is identifiable if,

$$\mathcal{H}_Z = \left\{ H(x) : H(x) = \sum_{i=1}^{c} c_i F(x; \lambda_i, \rho_i),\ c_i > 0, \sum_{i=1}^{c} c_i = 1,\ c \in [\![0; C]\!] \right\}$$

is identifiable.

If λ_i and ρ_i fulfill the conditions on proposition 9, the set $\mathcal{H}$ is identifiable and $\mathcal{H}_p$ too regarding to λ_i and ρ_i.

Now we will show that $F_{A_i,J} = F_{A_i,\tilde{J}} \Leftrightarrow J = \tilde{J}$.

Building upon the results discussed earlier and considering that the parameters of the ZIP distribution are expressed as polynomials, we can infer that the equality mentioned above is equivalent to the following condition for $i \in I$:

$$J\{(\nu,\beta) : (\nu A_i, \beta A_i) = (\nu_1, \beta_1)\} = \tilde{J}\left\{(\tilde{\nu} A_i, \tilde{\beta} A_i) : \left(\tilde{\beta} A_i, \tilde{\sigma}^2\right) = (\nu_1, \beta_1)\right\}$$

for some (ν_1, β_1).

Suppose there exists a parameter vector $\tilde{\beta}$ in $S(\tilde{J})$ such that $\beta A_i = \tilde{\beta} A_i$ for some $\beta \in S(J)$, which means that for $1 \leq t \leq T$, $\beta A_{it} = \tilde{\beta} A_{it}$. However, $\tilde{\beta}$ is not equal to β for any β in $S(J)$.

If $d_\beta < T$ and all the values of a_{it} are distinct, this situation leads to two different polynomials that are equal at T distinct points.

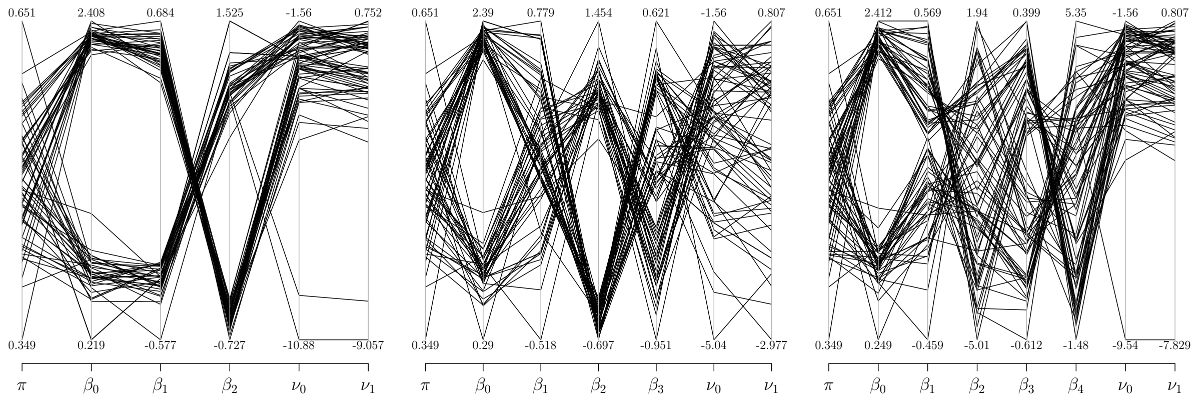


Figure 4.7: Parallel coordinate plots: ZIP
parallel coordinate plots of the estimated parameters for 50 samples of the artificial example. The left part with polynomial of degree 2 and the right one degree 4.

The unisolvence theorem implies $\beta = \tilde{\beta}$.

Similarly, if $d_\nu < T$ and all the values of a_{it} are distinct, it results in $\nu = \tilde{\nu}$.

□

4.1.4.3.3 Numerical example

To illustrate the theorem above we generate artificial data with different parameters. In all the case we have make the choice to take the coefficient in such way to have clusters there are very separate to avoid separation's problem.
Thus in all groups we use two clusters with probability 0.5 and

- $\beta_1 = (2.3025851, 0.5493061, -0.5493061)$ and $\beta_2 = (0.6931472, -0.3465736, 1.0397208)$;
- $\nu_1 = (-3, -1)$ and $\nu_2 = (-2, 0.2)$.

For each example, we generated 50 samples, each comprising 100 observations of the mixture models. We then fit the model using polynomials of degree 2, 3, and 4 for the polynomial part and 1 for the exceeded zero state.

As shown in Figure 4.7, the two left parts exhibit identifiable parameters, while the right part lacks identifiability, displaying only one bundle. Additionally, we can observe that the parameters ν are identifiable, but they have a wide range of values. This variability is due to the fact that in the zero-exceeded state, there are relatively few data points, around 10 %, which amounts to approximately 50 values.

4.1.4.3.4 Using trajectory for $\beta_k A_i + \delta_k W_i$

Proposition 11. *A REVOIR $\mathcal{C}_2$ is identifiable if*

- $d_{\beta_k} < T$ *and* $d_{\nu_k} < T$ *for all* $1 \leq k \leq K$ *and* a_{it} *are all distinct, for all* i, t
- *there exist* t *such that* W_t *is full rank ;*
- $rk(A_t, W) = rk(A_t) + rk(W)$ *where* $rk(\cdot)$ *determines the rank of a matrix.*

Proof. Since the covariates are just a linear addition to the model, the proof follows directly from proposition 10.

□

4.1.4.4 Beta model

In the case of the beta model, we assume that Y_{it} follows a beta distribution. The conditional density of this distribution, given the group k, is defined as:

$$h_k(y_{it}; \mu_{ikt}, \phi_{ikt}) = \frac{\Gamma(\phi_{ikt})}{\Gamma(\mu_{ikt}\phi_{ikt})\Gamma((1-\mu_{ikt})\phi_{ikt})} y_{it}^{\mu_{ikt}\phi_{ikt}-1}(1-y_{it})^{(1-\mu_{ikt})\phi_{ikt}-1}$$

where

$$\mu_{ikt} = \frac{e^{\beta_k A_{it} + \delta_k W_{it}}}{1 + e^{\beta_k A_{it} + \delta_k W_{it}}} \text{ and } \phi_{ikt} = \zeta_k A_{it}$$

It's important to note that mixtures of Beta distributions are generally not identifiable, as established by a characterization demonstrated in Yakowitz and Spragins (1968), and further remarked upon in Ahmad and Al-Hussaini (1982). This assumption serves as a crucial point in demonstrating the identifiability of our mixture model.

The non-identifiability of the Beta-GBTM may not be a significant issue, especially as the parameters of the shape are typically not directly interpretable in practical applications. The identifiability of the β parameters is primarily used to establish the identifiability of the δ parameters, which can have practical significance. While the β parameters lack a global interpretation, they still possess local meaning in relation to the β parameters.

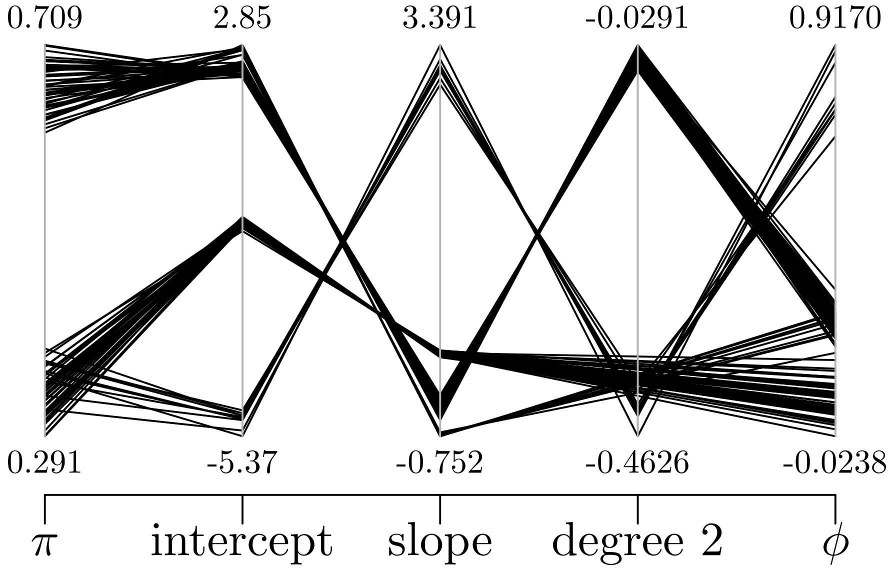


Figure 4.8: Parallel coordinate plots: Beta
parallel coordinate plots of the estimated parameters for 100 samples of the artificial example. We can see 2 differents model.

CHAPTER 5

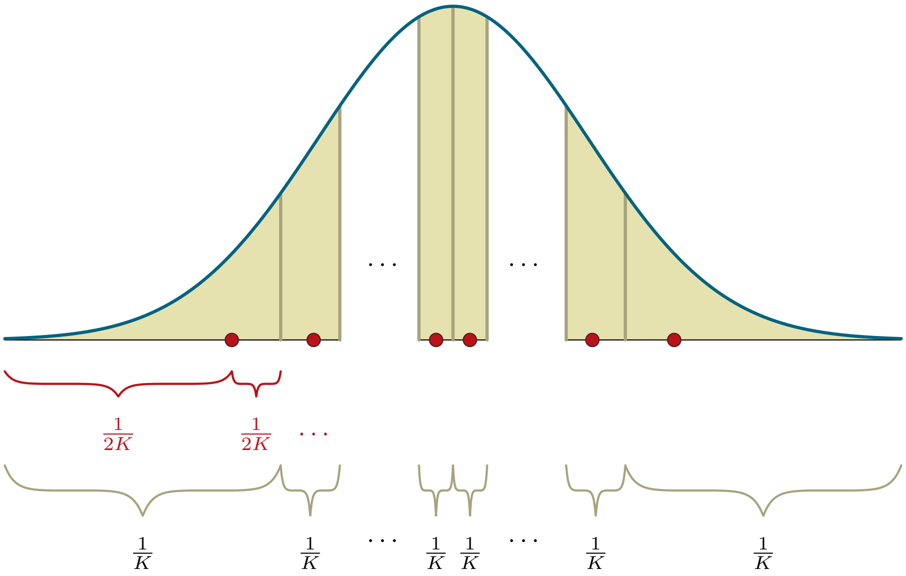
Starting values

5.1 Starting values

The selection of initial values is a critical aspect of the optimization process, especially when dealing with likelihood functions that may contain numerous local maxima. Starting values that are too distant from the true solution can impede the convergence of the optimization method. Unfortunately, there is no universally applicable method for determining the optimal initial values for every scenario. However, some heuristics can guide our choices.

In the context of regression with a normal distribution, it is commonly recommended to initialize the intercept parameter as the initial value. In our approach, we adhere to this principle and set the initial parameters to a vector of values, beginning with the intercept and followed by zeros for any additional coefficients. Considering our specific case with K groups, we partition the data from a normally distributed population into these K groups. This partitioning ensures that, for an observation following the normal distribution, each group has an equal probability of containing it. To determine the initial intercept value, we look for the 'center' of the divided data, a point that splits the area into two regions of equal probability. This 'center' corresponds to a probability mass of $\frac{1}{2K}$.

Concretely, if the distribution is characterized by a general probability density function (PDF) denoted as $\mathcal{D}$ with parameters θ, we divide this PDF into K subgroups. The initial intercept value is then estimated based on the group that encompasses the point representing the 'midllde' of the PDF. By 'middle,' we mean the value that separates the area into two regions of equal probability, each accounting for a probability mass of $\frac{1}{2K}$.



Therefore, the different value for the intercepts with K groups and a $\mathcal{D}(\theta)$ distribution are

$$\text{intercept}_k = q_{\frac{2(k-1)+1}{2K},\mathcal{D}(\theta)}, \quad k = 1, \cdots K$$

where $q_{\frac{2(k-1)+1}{2K},\mathcal{D}(\theta)}$ is the quantile of order $\frac{2(k-1)+1}{2K}$ for a distribution $\mathcal{D}(\theta)$ and $k = 1, \cdots K$.

5.1.1 Censored normal

In the Censored Normal model, we assume that the data Y follows a normal distribution with a mean equal to the sample mean $\overline{Y}$ and a standard deviation equivalent to the sample standard deviation σ_Y. The initial estimation of the intercept is determined as follows:

$$\text{intercept}_k = q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)}, \quad k = 1, 2, \ldots, K.$$

In this context, $q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)}$ represents the $\frac{2(k-1)+1}{2K}$-th quantile of the normal distribution with a mean of $\overline{Y}$ and a standard deviation of σ_Y, with k ranging from 1 to K.

$$\beta_k^{(0)} = \left(q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)}, 0, \cdots, 0\right) \tag{5.1}$$

For σ we consider the whole variability of sample.

$$\sigma_k^{(0)} = \sigma_Y \tag{5.2}$$

5.1.2 Logit

In the Logit model, we postulate that the values Y_{it} for an individual i are linked to predictor variables X_i through a latent variable Y_{it}^*, defined as:

$$Y_{it}^* = \beta_k X_i + \epsilon_i$$

Here, $\epsilon_i \sim G(\cdot)$, and $Y_{it} = 1_{Y_{it}^* > 0}$.

In the Logit model, $G(\cdot)$ corresponds to the logistic distribution, while in the Probit model, $G(\cdot)$ represents the normal distribution.

$$\begin{aligned} p_i = P\left(Y_{it} = 1 | X_i\right) &= P\left(Y_{it}^* > 0 | X_i\right) && (5.3) \\ &= P\left(\beta_k X_i + \epsilon_i > 0 | X_i\right) && (5.4) \\ &= P\left(\epsilon_i > -\beta_k X_i\right) && (5.5) \\ &= P\left(\epsilon_i < \beta_k X_i\right) && (5.6) \\ &= G^{-1}\left(\beta_k X_i\right) && (5.7) \end{aligned}$$

We can consider that the p_i are distributed according to $G(\cdot)$, and the quantities of interest are $\mu \approx \overline{Y}$ and $\sigma \approx s_Y$.

The choice between the Probit or Logit model depends on the specific situation, and the distinction between them is not particularly significant. This is because the logistic distribution and the normal distribution exhibit close similarities.

The Logit model allows for straightforward interpretation of the coefficient β_k, but the logistic distribution tends to assign more weight to extreme values compared to the normal distribution. It is for this reason that we prefer to use the normal distribution rather than the logistic distribution to determine the intercept.

By employing the method described above, we calculate the probability for each group as follows:

$$p_k = q_{\frac{2(k-1)+1}{2K}, \mathcal{N}(\overline{Y}, \sigma_Y)}, \quad k = 1, 2, \ldots, K$$

We then determine the intercept using the logistic distribution:

$$\text{intercept}_k = \log\left(\frac{p_k}{1 - p_k}\right), \quad k = 1, 2, \ldots, K$$

In these expressions, it's possible for p_k to be negative. In such cases, to ensure mathematical consistency, we replace intercept_k with a value that results in a probability close to zero, such as intercept $= -5$.

Finally, the initial values are established as follows for $1 \leq k \leq K$:

$$\beta_k^{(0)} = \begin{cases} \left(\log\left(\frac{q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)}}{1-q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)}}\right), 0\cdots, 0\right) & \text{if } q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)} \geq 0 \\ (-5, 0, \cdots, 0) & \text{if } q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)} < 0 \end{cases}$$

5.1.3 ZIP

In the Zero Inflated Poisson model, we assume that the values Y_{it} for an individual i follow a distribution given by:

$$Y_{it} \sim \rho_{ikt} + (1-\rho_{ikt})\mathcal{P}(\lambda_{ikt})$$

To begin, we make an arbitrary choice to allocate approximately 5 % of individuals to the zero excess state for each group. Consequently, we initialize the intercept for the parameter ρ_{ikt} as -3 ($e^{-3} \approx 0.0498$).

Next, we exclude the count of individuals in the zero excess state, and consider that the remaining values follow a Poisson distribution with the parameter $\lambda = e^{\text{intercept}}$.

Following the methodology outlined by Resear et al. (2016), we assume that the distribution of λ is of Pearson Type III, with rate $\hat{k} = \overline{Y}$ and shape $\hat{m} = \frac{\overline{Y}^2}{s_Y^2 - \overline{Y}}$.

The initial value, for $1 \leq k \leq K$, is set as follows:

$$\beta_k^{(0)} = \left(\log\left(q_{\frac{2(k-1)+1}{2K},\Gamma(\hat{k},\hat{m})}\right), 0, \cdots, 0\right) \tag{5.8}$$

$$\nu_k^{(0)} = (-3, 0, \cdots, 0) \tag{5.9}$$

5.1.4 Non linear model

In the case of non linear model, we suppose that the values Y_{it} of an individual i is defined by

$$y_{it} = f(a_{it}; \beta_k, \delta_k) + \epsilon_{it}$$

where $\epsilon_{it} \sim \mathcal{N}(0;\ \sigma_k)$.

Thus we can use the same method as Normal distribution. If $\eta : \beta_k \mapsto f(a_{it}; \beta_k, \delta_k)$ is reversible we consider

$$\beta_k^{(0)} = \left(\eta^{-1}\left(q_{\frac{2(k-1)+1}{2K},\mathcal{N}(\overline{Y},\sigma_Y)}\right), 0, \cdots, 0\right) \tag{5.10}$$

For σ we consider the whole variability of sample.

$$\sigma_k^{(0)} = \sigma_Y \tag{5.11}$$

5.1.5 Beta

The beta distribution is unique due to its dependence on two parameters. In the parametrization proposed by Ferrari Ferrari and Cribari-Neto (2004), the first parameter represents the mean, while the second parameter characterizes the degree of variability. When the variability is too small, it may lead to situations where a group does not receive all the individuals. Conversely, excessive variability can result in numerical errors due to individual values.

We explore two distinct approaches for determining initial values. First, we utilize the same method as previously, involving a beta sample. Second, we employ a normal approximation.

In the Beta model, we assume that the values Y_{it} for an individual i follow a beta distribution with parameters μ_{ikt} and ϕ_{ikt}.

5.1.5.1 One beta sample - obs

In this section, we assume that all values adhere to a Beta distribution and result from the summation of K Beta distributions, each carrying a weight of $\frac{1}{K}$. An approximation to the linear combination of independent Beta distributions is presented by Johannesson Jóhannesson and Giri (1995).

Suppose

$$Y = \sum_{k=1}^{K} \frac{1}{K} B_k \tag{5.12}$$

Y can be approximate by using $\gamma^{-1} Y^*$ where Y^* has a central beta distribution with parameters (g, h). These three parameters are determined by equating the first three central moments of the variable Y.

Let,

$$k_1 = \overline{Y} \tag{5.13}$$

$$k_2 = \frac{\overline{Y^2}}{\overline{Y}} \tag{5.14}$$

$$k_3 = \frac{\overline{Y^3}}{\overline{Y^2}} \tag{5.15}$$

Equating the first three centrals moments we obtained

$$A = k_3^2 k_2 + k_3 k_2 k_1 - 2k_3^2 k_1 \tag{5.16}$$

$$B = k_3^2 - 3k_3k_2 + 3k_3k_1 - k_2k_1 \tag{5.17}$$

$$C = 2k_2 - k_1 - k_3 \tag{5.18}$$

By substitution

$$\gamma = \frac{-B - \sqrt{B^2 - 4AC}}{2A} \tag{5.19}$$

$$g = \frac{(\gamma k_3 - 1)(\gamma k_2 - 1)}{\gamma (k_3 - k_2)} \tag{5.20}$$

$$h = \frac{\gamma k_3 g + 2\gamma k_3 - 2}{1 - \gamma k_3} \tag{5.21}$$

Finally we obtained the mean parameters (μ^*, ϕ^*) of Y^*

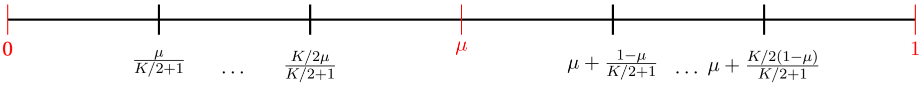
$$\mu^* = \frac{g}{g+h} \tag{5.22}$$

$$\phi^* = g + h \tag{5.23}$$

For the starting values we use this parameter ϕ and for the parameter β we follow this method :
Let $\mu = \mu^*/\gamma$,

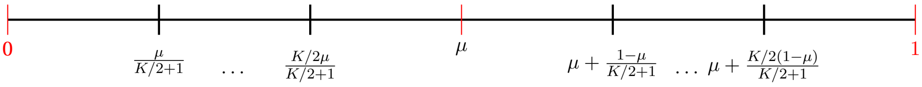
- if the number of groups is peer let

$$m_0 = \left(\frac{\mu}{K/2+1}, \cdots, \frac{K/2\mu}{K/2+1}, \mu + \frac{1-\mu}{K/2+1}, \cdots, \mu + \frac{K/2\,(1-\mu)}{K/2+1} \right) \tag{5.24}$$



- if the number of groups is odd let

$$m_0 = \left(\frac{\mu}{K/2+1}, \cdots, \frac{K/2\mu}{K/2+1}, \mu, \mu + \frac{1-\mu}{K/2+1}, \cdots, \mu + \frac{K/2\,(1-\mu)}{K/2+1} \right) \tag{5.25}$$



and finally, for $1 \leq k \leq K$,

$$\beta_k^{(0)} = \left(\log \left(\frac{m_{0k}}{1 - m_{0k}} \right), 0, \cdots, 0 \right) \tag{5.26}$$

Inverting the equation 3.4.1, $\phi = \frac{\mu(1-\mu)}{V(Y)} - 1$ and

$$\phi_k^{(0)} = \left(\frac{\overline{Y}(1-\overline{Y})}{V(Y)} - 1, 0, \cdots, 0) \right) \tag{5.27}$$

The $\log$ above is due to numerical transformation.

5.1.5.2 Normal approximation - na

A Taylor series expansion of the Beta distribution probability density function shows that the $Beta(a,b)$ distribution can be approximated by the Normal distribution when a and b are sufficiently large. More specifically, the conditions are:

$$\frac{a+1}{a-1} \simeq 1 \text{ and } \frac{b+1}{b-1} \simeq 1 \tag{5.28}$$

In such case,

$$Beta(a,b) \simeq \mathcal{N}\left(\frac{a}{a+b}, \sqrt{\frac{ab}{(a+b)^2(a+b+1)}} \right) \tag{5.29}$$

We use this approximation, even if it is not necessarily accurate, to find starting point of the algorithm. We remark that the parameters of the normal distribution is the mean and the standard deviation of the beta distribution too. Thus, with the mean characterization, see page 115, we have

$$Beta(\mu,\phi) \simeq \mathcal{N}\left(\mu, \sqrt{\frac{V(\mu)}{1+\phi}} \right) \tag{5.30}$$

For the β_k parameter and for $1 \leq k \leq K$, we first calculate the intercept as for normal distribution for $1 \leq k \leq K$, as $q_{\frac{2(k-1)+1}{2K}, \mathcal{N}(\overline{Y}, \sigma_Y)}$ and then we have,

$$\beta_k^{(0)} = \left(\log \left(\frac{q_{\frac{2(k-1)+1}{2K}, \mathcal{N}(\overline{Y}, \sigma_Y)}}{1 - q_{\frac{2(k-1)+1}{2K}, \mathcal{N}(\overline{Y}, \sigma_Y)}} \right), 0, \cdots, 0 \right) \tag{5.31}$$

Inverting the equation 3.4.1, $\phi = \frac{\mu(1-\mu)}{V(Y)} - 1$ and

$$\phi_k^{(0)} = \left(\frac{\overline{Y}(1-\overline{Y})}{V(Y)} - 1, 0, \cdots, 0) \right) \tag{5.32}$$

5.1.6 Comparison with Traj SAS

To compare the selection of initial values, we utilize the dataset provided within the R package `trajeR`. For all the examples, we employ initial values comprising only the non-zero intercept, i.e., $(\text{intercept}, 0, 0, \ldots)$. We then compare these values with those obtained using SAS, along with the likelihood.

Furthermore, it's worth noting that the probability of membership for each group remains consistent across the comparisons.

5.1.6.1 CNORM model

We use the data `CNORM_data01.csv`. The model is fit with 2 groups and the polynomial trajectory degrees are both 3.

	β		π		**Likelihood**
	Intercept 1	Intercept 2	Group 1	Group 2	
SAS	-2.760426	11.255411	0.5	0.5	**9292.7890572**
trajeR	-2.84408	11.33906	0.5	0.5	**9294.883100**

We use the data `CNORM_data07.csv`. The model is fit with 3 groups and the polynomial trajectory degrees are 1, 3 and 4.

	β			π			**Likelihood**
	Intercept 1	Intercept 2	Intercept 3	Group 1	Group 2	Group 3	
SAS	3.136829	9.61599	6.47916	0.33333	0.33333	0.33333	**16510.2557917**
trajeR	3.347283	9.615996	15.88471	0.3333333	0.3333333	0.3333333	**16492.294126**

5.1.6.2 LOGIT model

We use the data `LOGIT_2gr.csv`. The model is fit with 2 groups and the polynomial trajectory degrees are both 3.

	β		π		**Likelihood**
	Intercept 1	Intercept 2	Group 1	Group 2	
SAS	-1.372136	1.877511	0.5	0.5	**2140.571368**
trajeR	-1.396918	1.912452	0.5	0.5	**2156.367378**

We use the data `logitSASV2.csv`. The model is fit with 2 groups and the polynomial trajectory degrees are both 3.

	β		π		**Likelihood**
	Intercept 1	Intercept 2	Group 1	Group 2	
SAS	-5	-1.37390	0.5	0.5	**857.0584407**
trajeR	-5	-1.362981	0.5	0.5	**857.705652**

We use the data `dataLOGIT.csv`. The model is fit with 3 groups and the polynomial trajectory degrees are 1, 3 and 4.

	β			π			**Likelihood**
	Intercept 1	intercept 2	Intercept 3	Group 1	Group 2	Group 3	
SAS	-5	-0.725423	1.355464	0.333333	0.33333	0.33333	**3235.7175153**
trajeR	-5	-0.7254226	1.264463	0.3333333	0.3333333	0.3333333	**3223.413976**

5.1.6.3 ZIP model

We use the data `ZIPdata01.csv`. The model is fit with 2 groups and the polynomial trajectory degrees are both 3.

	β		ν				π		**Likelihood**
	Intercept 1	Intercept 2	Group 1		Group 2		Group 1	Group 2	
SAS	-0.738409	0.489316	-3	0	-3	0	0.5	0.5	**9034.0838866**
trajeR	0.2625239	1.55318	-3	0	-3	0	0.5	0.5	**6411.139523**

5.1.6.4 BETA model

We use the data `BETAdata01.csv`. The model is fit with 2 groups and the polynomial trajectory degrees are both 2.

	β		ϕ		π		**Likelihood**
	Intercept 1	Intercept 2	Group 1	Group 2	Group 1	Group 2	
SAS	-0.69315	0.69315	0.30125	0.30125	0.5	0.5	**-10874.38183**
trajeR	-0.17575	3.08273	0.39929	0.39929	0.5	0.5	**-10591.376253**
(obs)			(ζ_1 =-0.91806)	(ζ_1 =-0.91806)	0.5	0.5	
trajeR	-1.26929	1.45496	0.29963	0.29963	0.5	0.5	**-10061.386189**
(na)			(ζ_1 =-1.2052)	(ζ_2 =-1.2052)	0.5	0.5	

We use the data `BETAdata02.csv`. The model is fit with 3 groups and the polynomial trajectory degrees are 2 for the first two and 1 for the last one.

	β			ϕ			π			**Likelihood**
	Intercept 1	Intercept 2	Intercept 3	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	
SAS	-1.09861	0	1.09861	1.42472	1.42472	1.42472	0.33333	0.33333	0.33333	**-611.293760**
trajeR	-1.34602	0.06882	1.57065	1.42125	1.42125	1.42125	0.33333	0.33333	0.33333	**-440.763465**
(obs)				(ζ_1 =0.35153)	(ζ_1 =0.35153)	(ζ_1 =0.35153)				
trajeR	-1.738337	-0.2064379	0.5424494	1.250157	1.250157	1.250157	0.33333	0.33333	0.33333	**-468.840693**
(na)				($\zeta_1 = 0.2232689$)	($\zeta_1 = 0.2232689$)	($\zeta_1 = 0.2232689$)				

CHAPTER 6

Model selection and performance

6.1 Model selection for Mixture Models

Model selection is a major challenge in mixture models, especially in GBTM. It is necessary to determine the number of components G of finite mixtures and the degree of polynomial shape of the trajectory, and Nagin explains that there are two steps to proceed:

- the first step is to find the optimal number of clusters, the number of clusters,
- the last step is to determine the correct degrees of each trajectory.

Celeux et al. (2019) or Van Der Nest et al. (2020) make a review of different tools to determine the "best" number of clusters while Daniel S. Nagin (2005) use BIC and AIC method to select model.

6.1.1 Likelihood Ratio test

From a frequentist perspective, a commonly employed approach for determining the appropriate order of a statistical model is to employ the likelihood ratio test denoted as λ_{LR}. This test facilitates a comparison between two hypotheses: (H_0), representing a simpler model labeled as $model_0$, and (H_1), denoting a more complex model referred to as $model_1$. The formula for λ_{LR} is expressed as:

$$\lambda_{\text{LR}} = -2 \left[\ell(\hat{\theta}_0) - \ell(\hat{\theta}_1) \right]$$

Here, $\hat{\theta}_1$ signifies the maximum likelihood estimator under (H_1), while $\hat{\theta}_0$ represents the maximum likelihood estimator under (H_0).

Nevertheless, the proposed tests are numerically difficult to implement and slow. Furthermore, the likelihood-ratio test mandates that the models exhibit a nested structure. In other words, the more complex model must be transformable into the simpler model by imposing constraints on the parameters of the former. Consequently, the likelihood ratio test can prove valuable for exploring the degree of a polynomial but is unsuitable for determining the number of distinct groups within the data.

6.1.2 Information criteria

The information criterion is a method that involves penalizing the logarithm of the likelihood function based on a set of parameters. This concept is mathematically expressed as: $L(\Theta) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} g_k(y_{i_t}); \Theta_k \right)$. In this equation, Ω represents the set of parameters for the model, denoted as $\{K, \pi_1, \ldots, \pi_K, \Theta^1, \ldots, \Theta_K\}$, and Θ belongs to the space defined by Ω.
The key feature of the information criterion is that it introduces a penalty term that is directly proportional to the number of free parameters within the model, which is denoted as v_Ω. The choice of this penalty term plays a critical role in shaping the criteria. Essentially, the information criterion serves as a tool for quantifying the complexity of the model. It aids in balancing the goodness of fit of the model with its complexity, helping researchers make informed decisions when selecting models for their data analysis.

6.1.2.1 AIC and BIC

In the realm of classical model selection criteria, two widely recognized methods are the Akaike Information Criterion (AIC), introduced by Akaike (1974), and the Bayesian Information Criterion (BIC), formulated by Schwarz (1978). These criteria provide valuable tools for assessing the goodness of fit of statistical models.

$$AIC = -2 \log L(\hat{\Theta}) + 2v_\Omega \tag{6.1}$$

whereas BIC is defined as

$$BIC = -2 \log L(\hat{\Theta}) + v_\Omega \log(N) \tag{6.2}$$

6.1.2.2 The Slope Heuristics

Within the framework of Slope Heuristics, the penalty function is designed to be directly proportional to the number of free parameters denoted as v_Ω. To determine the appropriate penalty, a data-driven slope estimation procedure, known as Data-Driven Slope Estimation (DDSE), is employed. This method operates under the assumption of a linear relationship between the likelihood of observed data and the penalty term.
The primary goal of the DDSE procedure is to estimate the slope, represented as κ, within

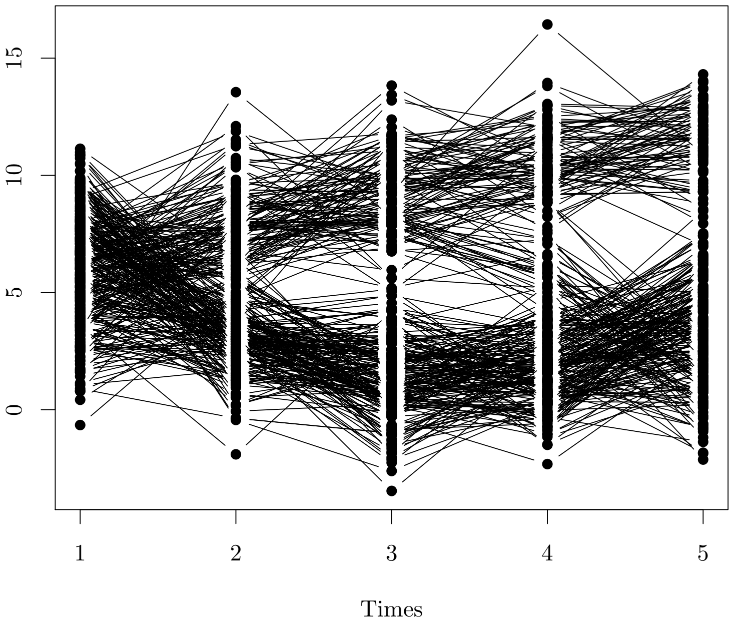


Figure 6.1: Artificial dataset

this linear relationship. This slope defines the minimum penalty, κv_Ω, below which smaller penalties would favor the selection of more complex models. Conversely, higher penalties are meant to favor the selection of models with a reasonable level of complexity. This concept has been advocated by Birgé and Massart (2007) and Baudry et al. (2012), who suggest setting the penalty to be $2\kappa v_\Omega$. As a result, the Slope Heuristics criterion (SH) is defined as:

$$SH = -\log L(\hat{\Theta}) + 2\kappa v_\Omega$$

To estimate the slope κ, one can utilize the `capushe` package in the R programming environment, as demonstrated by Baudry et al. (2012), for instance.

6.1.3 Numerical example

For illustrative purposes in this section, we employ synthetic data. This dataset is generated to emulate a distribution that consists of a mixture of normal distributions.

In the initial step, our objective is to determine the optimal number of groups.
We set the polynomial degree to 2 for all models and employ the `trajeR` package to fit various models, systematically varying the number of groups from 2 to 11. Subsequently, we compute and evaluate three different information criteria—namely, AIC, BIC, and SH—for each group configuration.

Group	AIC	BIC
2	6972.187	7005.521
3	6548.822	6600.675
4	6285.395	6355.767
5	6293.933	6382.823
6	6300.2	6407.61
7	6307.999	6433.927
8	6309.845	6454.293
9	6323.433	6486.4
10	6327.453	6508.939
11	6333.31	6533.314

When applying the SH method, we rely on the `capushe` package to calculate the Slope Heuristics (SH) score. Interestingly, each of these methods consistently identifies a model with four groups as the optimal choice.

In the second step of our analysis, our focus shifts to determining the appropriate polynomial degree.
Recalling the previous step where we opted for a model with four groups and a polynomial degree of 2, we now visually explore the data to assess if an alternative choice of polynomial degree could be more suitable.

Upon examining the plot, it becomes evident that the segments marked in pink and blue exhibit a nearly flat trend. As a result, it seems reasonable to consider maintaining a polynomial degree of 2 for the yellow and orange segments, while potentially using a lower degree, such as 0 or 1, for the pink and blue sections.
To explore these possibilities, we proceed to fit various models with differing polynomial degree configurations using the AIC, BIC, and SH methods.

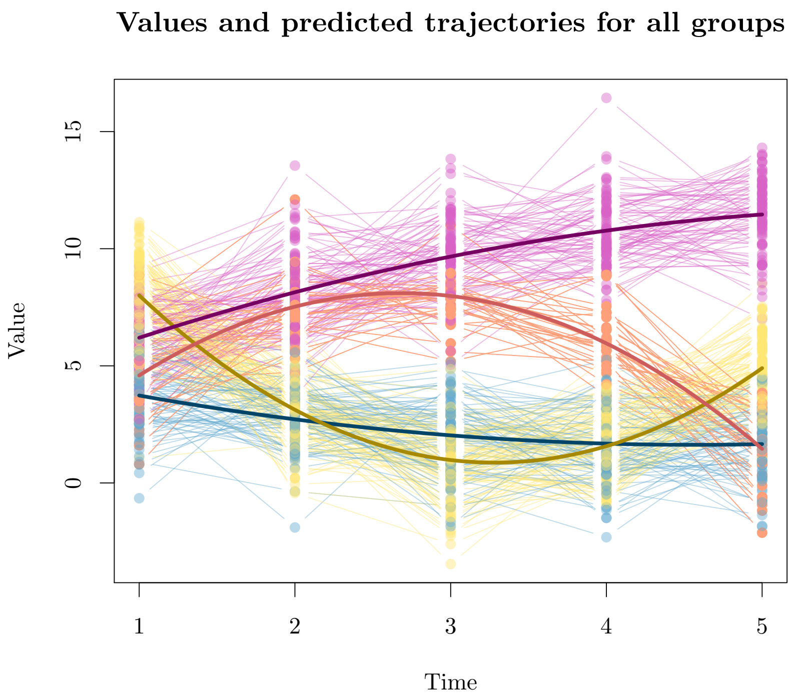


Figure 6.2: Model fitted for the data.

Group	AIC	BIC
(**2**,2,2,**2**)	6285.395	6355.767
(**0**,2,2,**0**)	6873.03	6928.587
(**0**,2,2,**2**)	6358.474	6421.439
(**2**,2,2,**0**)	6829.348	6892.313
(**2**,2,2,**1**)	6305.923	6372.591
(**1**,2,2,**2**)	6293.308	6359.976
(**1**,2,2,**1**)	6313.667	6376.631
(**1**,2,2,**0**)	6833.981	6893.241
(**0**,2,2,**1**)	6377.591	6436.852
(**1**,2,**3**,**1**)	6315.057	6381.725

The necessity for the last change arises from the fact that the SH method requires a minimum of 10 data points. Following the guidance provided by AIC and BIC, we opt to maintain a polynomial degree of 2 for all segments. Interestingly, the scores obtained for degree **1**, 2, 2, and **2** are quite similar according to AIC and BIC. In contrast, the SH method recommends a slightly different configuration, leading us to select degree **2**, 2, 2, and **0**.

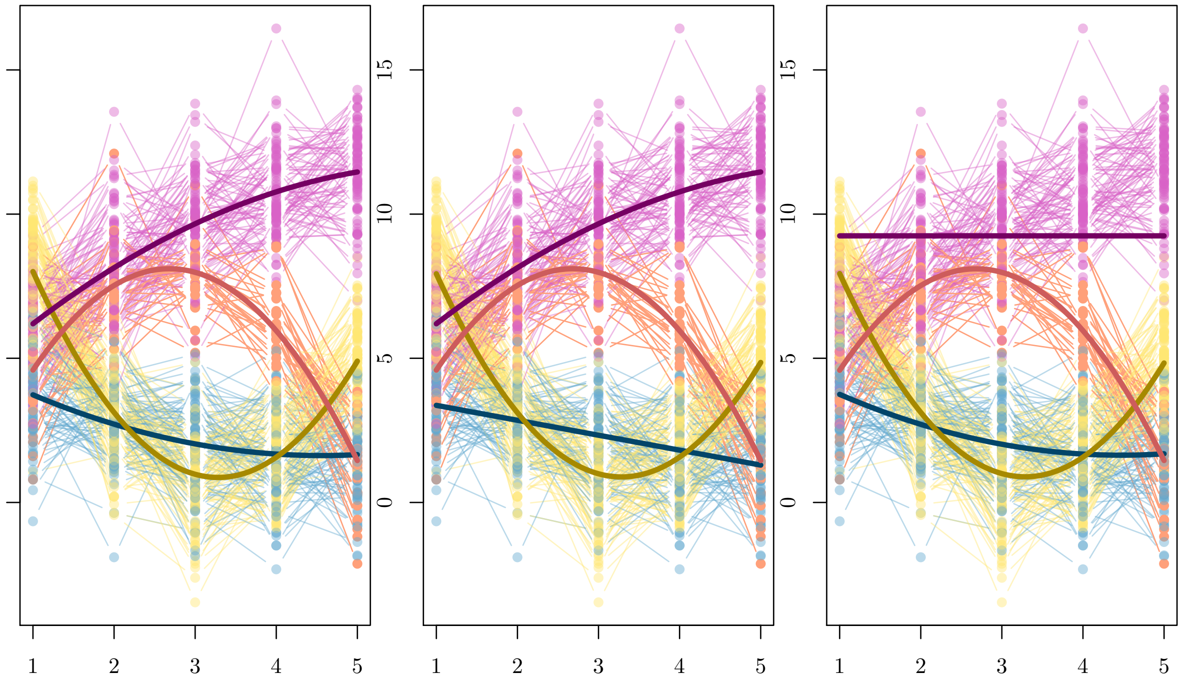


Figure 6.3: Comparison of three models.
Different models fitted. The left one is the model preconized by AIC and BIC. The middle one is the model near the preconized one. The right one is the model preconized by SH.

6.1.4 Non parametric indices for the number of clusters

One limitation of using likelihood-based methods to determine the number of clusters is their sensitivity to the initial choice of the polynomial's degree. Different initial degree selections can lead to varying cluster numbers. To mitigate this challenge, alternative indices can be employed to evaluate clustering quality. For instance, B. Desgraupes and M. B. Desgraupes (2018) has compiled a list of such indices for optimizing clustering solutions. Additionally, studies by Shim et al. (2005) and Milligan and Cooper (n.d.) have compared the efficiency of several indices, highlighting the Calinski and Harabasz (CH) criterion as one of the top performers.

The CH criterion is based on a combination of within-group variance and between-group variance. To define this, we introduce two essential terms: $BGSS$ (Between-Group Sum of Squares), which quantifies the dispersion between clusters, and $WGSS$ (Within-Group Sum of Squares), representing the cumulative dispersion within each cluster. The $WGSS$ term is computed as the sum of squared distances between individual observations and the barycenter (center of mass) of their respective clusters. On the other hand, the $BGSS$ term is calculated as the weighted sum of squared distances between the barycenter of each cluster and the overall barycenter of the entire dataset.

The CH index is defined as follows:

$$CH = \frac{BGSS/(K-1)}{WGSS/(N-K)} \tag{6.3}$$

Here, N denotes the total number of elements in the dataset, and K represents the number of clusters. The optimal number of clusters is determined by finding the value of K that maximizes the CH index. In order to calculate the CH index, it's necessary to define a measure for the trajectories. This measure is crucial for evaluating and selecting the optimal number of clusters based on the CH criterion.

6.1.4.1 Euclidean distance

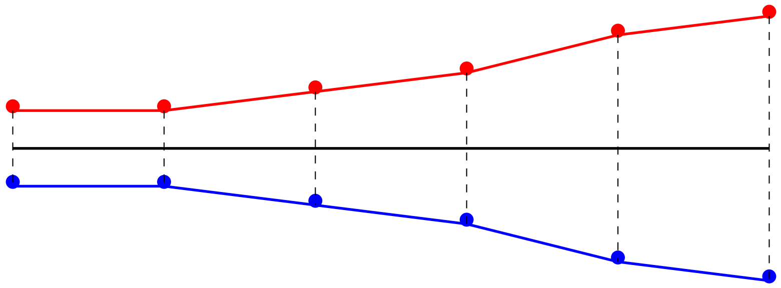
To gauge the distance between two trajectories, the Euclidean distance can be employed by comparing values at the same time points. Consequently, the mean trajectory is defined as $\overline{y} = (\overline{y_{11}}, \ldots, \overline{y_{1T}})$, where $\overline{y_{it}}$ is calculated as $\frac{1}{n}\sum_{i=1}^{n} y_{it}$, capturing the average trajectory values across all data points.

With this measure in place, the CH index is then expressed as follows:

$$CH = \frac{\displaystyle\sum_{k=1}^{K} n_k \left(\overline{y_k} - \overline{y}\right)\left(\overline{y_k} - \overline{y}\right)'}{\displaystyle\sum_{k=1}^{K} \sum_{\substack{i=1 \\ y_i \in C_k}}^{n} \left(y_i - \overline{y_k}\right)\left(y_i - \overline{y_k}\right)'} \tag{6.4}$$

where $\overline{y_k}$ is the mean trajectory of the cluster k, $y_i = (y_{i1}, \ldots, y_{iT})$ and C_k the cluster k.

While the CH index offers the advantage of straightforward computation, it has a limitation in that it does not consider the shape or pattern of the trajectories. For instance, when employing the Euclidean distance, the index treats the distance between the red path and the black path as equivalent to the distance between the blue path and the black path. This limitation arises from the fact that it solely relies on point-to-point distance measures, failing to capture the potential differences in trajectory shape.



6.1.4.2 Discrete Fréchet distance

Several distance metrics are available that consider the shape of trajectories, such as Dynamic Time Warping (DTW) and the Fréchet distance.
In the case of the Fréchet distance, it is often illustrated using the scenario of a man and his dog walking together. Both the man and the dog are permitted to move along curves at different speeds, but they are not allowed to backtrack. The Fréchet distance is defined as the length of the shortest leash required for both the man and the dog to traverse their respective curves without violating these constraints.

Definition 6.1.1. *Fréchet distance Let P and Q be two metric curves in a metric space S. Let α and β be two reparametrization mappings, i.e. continuous, non-decreasing, surjective function from $[0,1]$ to $[0,1]$. Let d be a distance function int he metric space S. The Fréchet distance δ_F is defined as :*

$$\delta_F = \inf_{\alpha,\beta} \max_{t\in[0,1]} d\left(P(\alpha(t)), Q(\beta(t))\right)$$

In the context of our study, the trajectories represent polygonal curves, and we can effectively utilize the Fréchet distance. Specifically, this distance is referred to as the discrete Fréchet distance, and it serves as a valuable measure of similarity between two polygonal curves.

In this context, the trajectories can be considered as P and Q, corresponding to the man and the dog's paths, with the mappings α and β representing their respective speeds. The discrete Fréchet distance is a measure that quantifies the resemblance between these polygonal curves, taking into account their specific characteristics.

Definition 6.1.2. *Coupling Let $P = (u_1, u_2, \ldots, u_p)$ and $Q = (v_1, v_2, \ldots, v_q)$ be two polygonal curves. A coupling L between P and Q is defined as a sequence:*

$$(u_{a_1}, v_{b_1}), \ldots, (u_{a_m}, v_{b_m})$$

Such that:

- $a_1 = b_1 = 1$, $a_m = p$ *and* $b_m = q$;
- $\forall i = 1, \cdots m \begin{cases} a_{i+1} = a_i \text{ or } a_{i+1} = a_i + 1 \\ b_{i+1} = b_i \text{ or } b_{i+1} = b_i + 1 \end{cases}$

This definition avoid back-step and allow different speed on each curve.

Definition 6.1.3. *Length of coupling L The length of a coupling L, denoted $||L||$ is the length of*

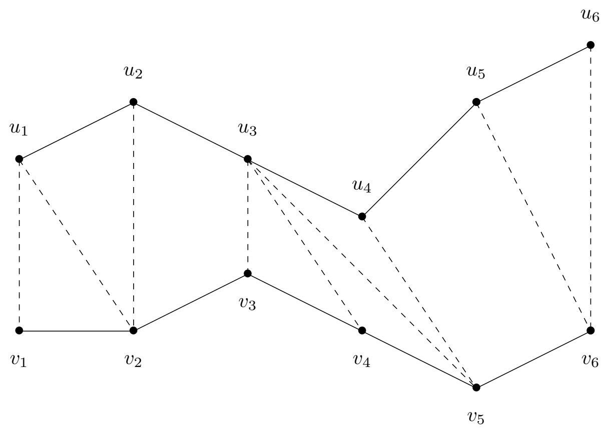


Figure 6.4: Fréchet distance
Example of coupling between two curves. A coupling is
$(u_1, v_1), (u_1, v_2), (u_2, v_2), (u_3, v_3), (u_3, v_4), (u_3, v_5), (u_4, v_5), (u_5, v_6), (u_6, v_6)$.

the longest link in L,

$$||L|| = \max_{i=1,\dots,m} d\left(u_{a_i}, v_{b_i}\right)$$

Definition 6.1.4. *discrete Fréchet distance The discrete Fréchet distance d_{dF} between two polygonal curve P and Q is the minimum length of all the coupling between P and Q.*

$$d_F = \min_{\textit{coupling}} \max_{i=1,\dots,m} d\left(u_{a_i}, v_{b_i}\right)$$

To calculate the discrete Fréchet distance, we have the option of using optimized algorithms, such as the ones developed by Devogele et al. (2017), or the original method proposed by Eiter and Mannila (n.d.). The distance is computed via dynamic programming and involves filling two matrices.

Consider two polygonal curves, denoted as $P = (u_1, u_2, \dots, u_p)$ and $Q = (v_1, v_2, \dots, v_q)$. The process involves the following steps:

1. First, we compute the distance matrix, labeled as DM, where each element (i, j) is defined as:

$$DM_{ij} = d(u_i, v_j)$$

2. Next, we calculate the Fréchet matrix, referred to as FM, with each element (i, j) expressed as:

$$FM_{ij} = \max\left(DM_{ij}, \min\left[FM_{(i-1)j}, FM_{i(j-1)}, FM_{(i-1)(j-1)}\right]\right)$$

The discrete Fréchet distance is ultimately found in the last cell of the FM matrix. This distance measure effectively captures the similarity between the two polygonal curves, accounting for their distinctive characteristics.

For instance, in order to calculate the discrete Fréchet distance between the two curves depicted in the previous example which values are

x	1	2	3	4	5	6
P	7	8	7	6	8	9
Q	4	4	5	4	3	4

1. First we compute DM

3.00	3.16	2.83	4.24	5.66	5.83
4.12	4.00	3.16	4.47	5.83	5.66
3.61	3.16	2.00	3.16	4.47	4.24
3.61	2.83	1.41	2.00	3.16	2.83
5.66	5.00	3.61	4.12	5.00	4.12
7.07	6.40	5.00	5.38	6.08	5.00

2. Next we compute FM

3.00	3.16	2.83	4.24	5.66	5.83
4.12	4.00	3.16	4.47	5.83	5.66
3.61	3.61	3.16	3.16	4.47	4.47
3.61	3.61	3.16	3.16	3.16	3.16
5.66	5.00	3.61	4.12	5.00	4.12
7.07	6.40	5.00	5.38	6.08	5.00

Thus, the discrete Fréchet distance is 5.

We introduce the Calinski-Harabasz-Fréchet (CHF) criterion, which combines the concepts of the Calinski and Harabasz (CH) criterion with the Fréchet distance to evaluate the quality of clustering solutions.

$$CHF = \frac{\sum\limits_{k=1}^{K} n_k d_F^2\left(\overline{y_k}, \overline{y}\right)/(K-1)}{\sum\limits_{\substack{i=1\\ y_i \in C_k}}^{n} d_F^2\left(y_i, \overline{y_k}\right)/(N-K)}$$

where $\overline{y}$ is the Fréchet mean of all values, $\overline{y_k}$ is the Fréchet mean of the values within cluster k, denoted as C_k.

To calculate the Calinski-Harabasz-Fréchet criterion, it is necessary to find an average curve. There are several strategies to determine a central curve when dealing with sets of polygonal curves, denoted as S, containing n elements $\{p_1, \ldots, p_n\}$.

1. One approach is to use the equations of the polynomial curves fitted by the clustering algorithm for each cluster.

2. Another method involves approximating the median using a marginal median, as described in works such as Puri and Sen (1971), or utilizing algorithms presented in Petitjean et al. (2011).

3. Alternatively, an approximation of the mean can be employed. Some methods for this include:

   - Genolini, Ecochard, et al. (2016) define the mean between two polygonal curves P and Q by coupling their vertices and averaging them:

     $$meanFr\acute{e}chet\,(P, Q) = \left(\frac{p_{a_1} + q_{b_1}}{2}, \ldots, \frac{p_{a_m} + q_{b_m}}{2}\right)$$

     They expand this method to multiple curves by calculating the mean two by two and propose several ways to pair the curves.

   - Ahn et al. (2020) offer an algorithm to compute the mean curve by selecting each vertex from all the curves.

   - Buchin, Driemel, Gudmundsson, et al. (2018) use the (k, l)-center method to find a mean curve. By setting $k = 1$ and $l = T$, we obtain a mean curve.

However, it is important to note that these methods have certain drawbacks. They can be computationally expensive, and differences in the y-values and x-values, potentially expressed in different units, may affect the Fréchet distance. For example, in Figure 6.5, we have illustrated three curves p_1 in black, p_2 in red, p_3 in blue, and p_4 in green, but with different x-value scales. From left to right, the scales are 1, 0.1, and 2. If we compute the discrete Fréchet distance in all of these cases, we find:

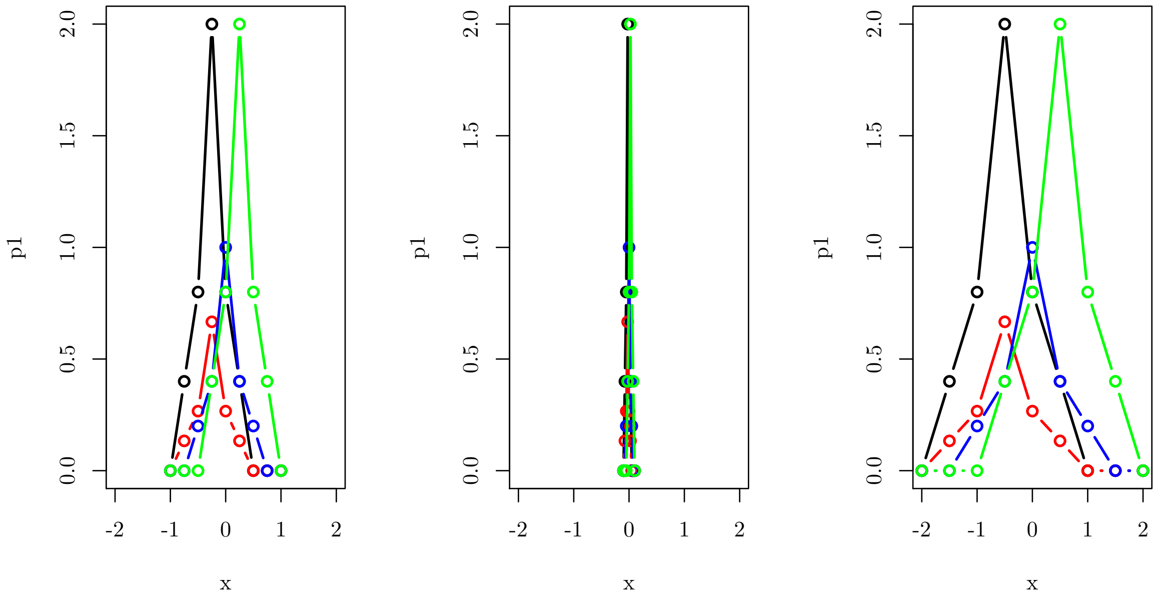


Figure 6.5: Scale effect of Fréchet distance.
The impact of scaling the x value ont he distance betweenn trajectories.

Discrete Fréchet distance

	Left case	Middle case	Right case
scale	0.1	1	2
$d_F(p_1, p_2)$	2.147	2.933	2.933
$d_F(p_1, p_3)$	1.883	2.473	3.000
$d_F(p_1, p_4)$	0.400	3.887	5.600

In the case on the left, the curves p_1 and p_4 exhibit the closest proximity, indicating a lower Fréchet distance. In the middle case, it's the curves p_1 and p_3 that are most closely aligned. Conversely, in the rightmost scenario, it's the curves p_1 and p_2 that show the least separation, resulting in a smaller Fréchet distance.

It's worth noting that as the scale tends toward positive infinity, the discrete Fréchet distance converges toward the Euclidean distance, while as the scale approaches zero, the discrete Fréchet distance tends to resemble DTW. To mitigate this challenge, Genolini, Ecochard, et al. (2016) introduced a time scale index that allows for the adjustment of the cost associated with horizontal shifts. The user is required to define the time scale according to the specific context and needs of their analysis.

6.1.4.3 Dynamic time Wraping (DTW)

Another classical method for comparing trajectories is the DTW approach. DTW seeks to find the optimal alignment between two curves by minimizing a cost function, allowing for flexible comparisons by matching the coordinates within both curves. In this method, the cost of an association is represented by the distance metric d between two elements. More precisely, DTW is defined through a recursive process:

$$d_{DTW}\left(A_i, B_j\right) = d(a_i, b_j) + \min \begin{cases} d_{DTW}\left(A_{i-1}, B_{j-1}\right) \\ d_{DTW}\left(A_i, B_{j-1}\right) \\ d_{DTW}\left(A_{i-1}, B_j\right) \end{cases}$$

where A_i is the subsequence $(a_1, \ldots, a_i)$ and B_i is the subsequence $(b_1, \ldots, b_i)$. The final measure of overall similarity is computed as the last term, $d_{DTW}\left(A_{|A|}, B_{|B|}\right)$.

It's important to note that d_{DTW} is not a distance metric in the traditional sense.

In this context as well, a challenge lies in finding a 'middle' curve. Petitjean et al. (2011) introduce a heuristic strategy for obtaining an averaging curve, known as DTW Barycenter Averaging (DBA). This approach involves iteratively refining an initially computed average sequence to minimize its square DTW distance from the average sequences. One drawback of this method is that the choice of the initial average can influence the final outcome. In our approach, we opt to begin with a curve whose elements represent the arithmetic mean of all elements with the same time value in the set of curves.
We define the Calinski-Harabasz-DTW (CHDTW) criterion:

$$CHDTW = \frac{\sum_{k=1}^{K} n_k d^2_{DTW}\left(\overline{y_k}, \overline{y}\right)/(K-1)}{\sum_{\substack{i=1 \\ y_i \in C_k}}^{n} d^2_{DTW}\left(y_i, \overline{y_k}\right)/(N-K)}$$

Here, $\overline{y}$ represents the Fréchet mean of all data points, while $\overline{y_k}$ signifies the Fréchet mean of values within cluster k, noted as C_k.

6.1.4.4 Numerical example

In the example above, we calculated several indices to determine the number of clusters, including AIC, BIC, CH, CHF (using polynomial shape), CHF (using Fréchet mean with time scales of 0.1 and 4), and CHDTW for various degrees of the polynomial shape in the model. We present only the number of clusters selected by each method.

Degree of polynomial	AIC	BIC	CH	CHF polynomial	CHF scale 0.1	CHF scale 4	CHDTW
1	3	3	2	2	3	3	3
2	4	4	3	3	3	4	3
3	4	4	3	3	3	4	3
4	4	4	3	3	3	6	3
5	3	3	3	2	3	3	3

Another approach to assess the accuracy of a model is by evaluating its 'performance,' a concept we will define in the following sections.

6.2 Using posterior Group-Membership probabilities

6.2.1 Posterior probability

The posterior probability of group membership, denoted as $P(k|Y_i)$, represents the likelihood of individual i belonging to group k based on their observed behavior in each period. Using Bayes' formula, we can express this probability as:

$$P(k|Y_i) = \frac{P(Y_i|k)\pi_k}{\sum_{k=1}^{K} P(Y_i|k)\pi_k}$$

With this probability in hand, we have a method for assigning individuals to specific groups. Specifically, we assign individual i to the group for which the posterior probability is highest. It's important to note that this is the method used to color the data points in each graph throughout this text.

To illustrate this concept, we utilize artificial data comprising four groups, with each group characterized by a 2-degree polynomial function. The time periods span from 1 to 5, and the probabilities are influenced by five covariates, represented by the variable X.

X1	X2	X3	X4	X5
3.68	-9.37	9.01	1	1
5.87	-2.38	19.51	0	1
9.45	9.42	93.78	1	0
3.28	-5.62	12.92	1	1
2.54	5.32	58.66	0	0
8.92	-2.23	30.15	1	0

X_1, X_2, and X_3 may represent continuous covariates, such as salary or hemoglobin levels in the blood, whereas X_4 and X_5 may denote binary characteristics, like gender or health status (e.g., male/female or sick/not sick).
For instance, in the artificial data mentioned earlier, let's examine the membership probabilities for the first 17 individuals:

Group	**Probability - individual**																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0	0	0	0.03	0	0	0	0	0	0	0	**0.738**	0	0	**0.902**	0
2	0	0	0	0	0.004	0	0	0	0	0	0	0	0.238	0	0	0.043	0
3	**1**	**1**	0	**1**	0	**1**	0	**1**	**1**	0	**1**	**1**	0	**1**	**1**	0	0
4	0	0	**1**	0	**0.966**	0	**1**	0	0	**1**	0	0	0.024	0	0	0.056	**1**

In each plot, the most likely group for each individual is highlighted in gray. This practical application offers valuable insights.

6.2.2 Profiles of group

An essential question for researchers is whether certain characteristics of individuals in trajectory groups can distinguish them from their counterparts in other trajectory groups. To address this question, the posterior probabilities serve as a straightforward foundation for creating such profiles. By performing a cross-tabulation of individual-level trajectory group assignments with individual-level characteristics that could be associated with trajectory group membership, valuable insights can be gained.

	Gr 1	**Gr 2**	**Gr 3**	**Gr 4**
X1	1.8467	3.2709	6.0562	6.5621
X2	7.7916	-0.9212	-4.9915	5.3515
X3	56.1620	74.0841	41.0704	32.8957
X4	0.3636	0.2895	0.4972	0.3333
X5	0.2545	0.3289	0.4525	0.3421

Since X_1, X_2, and X_3 are continuous variables, the values in the table represent the means of each variable within each group. On the other hand, X_4 and X_5 can be interpreted as percentages. In the figure shown below, it's evident that Group 3 exhibits an increasing trend over time. Further analysis of group profiles reveals that this group tends to have lower values for the covariate X_2 and a higher proportion of individuals possessing the covariates X_4 and X_5. These findings strongly suggest that these covariates may influence an individual's membership in a particular group.

6.2.3 Model adequacy

In practice, a crucial question arises: Is our model adequate? In other words, does the model accurately represent the data, and can we trust it as a reliable description, or is it merely the 'least worse' model available? This section serves as a guide to help address this fundamental question.

6.2.3.1 Average Posterior Probability - diagnostic 1

We refer to the Average Posterior Probability (AvePP) as the average posterior probability of membership for individuals assigned to each group. In an ideal scenario, the assignment probability for each individual would be 1, and AvePP would also equal 1. When the posterior probability assignments decrease, AvePP decreases as well. The question arises: from what value can we consider AvePP to be sufficiently close to 1?
Daniel S. Nagin (2005) suggests, based on experience, that AvePP should be at least 0.7. In the example above, we calculated AvePP for each group while considering the covariate:

AvePP - with covariates			
Gr 1	Gr 2	Gr 3	Gr 4
0.9999	0.9983	1	1

AvePP is close to 1, the model classifies correctly the individuals.

Calculating AvePP without considering covariates results in:

AvePP - without covariates			
Gr 1	Gr 2	Gr 3	Gr 4
0.8984	0.9396	0.9843	0.9974

It's worth noting that the values are lower when no covariates are considered compared to when covariates are included. This observation suggests that the addition of covariates enhances accuracy by correctly assigning more individuals to the appropriate groups. For instance, in the case of Group 1, the model without covariates correctly classifies approximately 89.84 % of individuals. However, when covariates are included in the model, the accuracy rate significantly improves to 99.99 %.

6.2.3.2 Odds of Correct Classification - diagnostic 2, no covariate

The Odds of Correct Classification (OCC) for group k (OCC$_{\text{j}}$) is a measure obtained by comparing the odds of a correct classification into group j using the posterior probability rule with the odds of correct assignment based on random assignments, where the probability of assignment to group j is represented by $\hat{\pi_k}$ which is the probability estimated by the model.

$$OCC_k = \frac{AvePP_k/(1 - AvePP_k)}{\hat{\pi_k}/(1 - \hat{\pi_k})} \tag{6.5}$$

This statistic can be interpreted as follows: Imagine that the maximum probability assignment rule has no predictive capacity compared to random chance, meaning $\hat{\pi_k} = AvePP_k$. In this scenario, OCC_k would equal 1. Conversely, if the maximum probability assignment rule demonstrates good predictive capacity, causing $AvePP_k$ to approach the ideal value of 1, OCC_k increases. Therefore, larger values of OCC_k indicate better assignment accuracy.
It's important to note that this method is applicable only when we have no covariates influencing the membership probability.

In the example above, we fitted the model without covariates

Probability				**AvePP**			
π_1	π_2	π_3	π_4	Gr 1	Gr 2	Gr 3	Gr 4
0.1023	0.3118	0.2247	0.3612	0.8984	0.9396	0.9843	0.9974

and then OCC_k

OCC_1	OCC_2	OCC_3	OCC_4
77.59455	34.3318	216.7626	676.5347

All this values are big, the model allow better assignment accuracy.

6.2.3.3 Estimated group probabilities versus proportion of the sample assigned to the group - diagnostic 3, no covariate

The probability of group membership can be estimated in two ways: firstly, by using $\hat{\pi_k}$, and secondly, by using the proportion P_k of the sample assigned to group k. When assignments within each group are perfectly done (i.e., AvePP is 1), $\hat{\pi_k} = P_k$. Therefore, differences between these two quantities indicate assignment errors.

	Gr 1	**Gr 2**	**Gr 3**	**Gr 4**
Probability	0.0980	0.3180	0.2220	0.3620
Proportion	0.1023	0.3118	0.2247	0.3612

6.2.3.4 Confidence interval - diagnostic 4, no covariate

A narrow confidence interval of $\hat{\pi_k}$ indicates that the probability is accurately estimated. However, this diagnostic has two limitations. First, there are no formal criteria for determining what constitutes a sufficiently narrow confidence interval to consider it accurate. Second, calculating the interval directly from the model's parameter estimates can be challenging due to the use of nonlinear functions.

The second challenge can be overcome by employing statistical methods like bootstrap. After fitting the model, we obtain the distribution of θ, which is the vector of all θ_k. As θ results from maximum likelihood estimation, it asymptotically follows a normal distribution with a mean of $\hat{\theta}$ and a variance-covariance matrix of $V(\hat{\theta})$. We can utilize these values to generate a large sample of θ values and calculate associated probabilities. Then, using these computed probabilities, we can determine confidence intervals with a 98 % degree of confidence.

	Gr 1	**Gr 2**	**Gr 3**	**Gr 4**
1%	0.0819	0.2720	0.1916	0.3166
99%	0.1273	0.3521	0.2620	0.4101

6.2.3.5 Summary

Let's summarize the methods discussed above in a single table.

	Gr 1	**Gr 2**	**Gr 3**	**Gr 4**
Prob. est.	0.1023	0.3118	0.2247	0.3612
CI inf.	0.0819	0.2728	0.1909	0.3170
CI sup.	0.1269	0.3522	0.2611	0.4087
Prop.	0.0980	0.3180	0.2220	0.3620
AvePP	0.8984	0.9396	0.9843	0.9974
OCC	77.5945	34.3318	216.7626	676.5347

CHAPTER 7

Group-based multi-trajectory modeling

7.1 Definitions

In practical situations, it's common to analyze various indicators of interest. Rather than studying each indicator separately, a more intriguing approach is to examine them jointly. Daniel S. Nagin (2005) and Daniel S Nagin, Bobby L Jones, et al. (2018) introduce a method to address this situation, which is an extension of the GBTM model.

The central concept of interest is the outcomes conditional on age. In the GBTM, the likelihood for each individual i conditional on group k is as follows:

$$P(Y_i) = \sum_{k=1}^{K} \pi_k P^k(Y_i|A_i, W_i, \Theta_k) = \sum_{k=1}^{K} \pi_k P^k(Y_i|\Theta_k) \tag{7.1}$$

Where: Y_i represents an individual's longitudinal measurements, A_i is a vector of time measurements, π_k denotes the probability of belonging to group k, W_i is a vector of covariates, Θ_k is an unknown vector, among other things, determining the shape of the group-specific trajectory. For a more comprehensive explanation of this notation, you can refer to page 6.

For a given group, we assume the independence of the elements y_{it} in Y_i. Therefore, the probability for an individual belonging to group k and their measurements is represented as follows:

$$P^k(Y_i|\Theta_k) = \prod_{t=1}^{T} p^k(y_{it}|\Theta_k) \tag{7.2}$$

Here, $p^k(\cdot)$ is the distribution of y_{it} conditional on group membership k.

In the multi-trajectory framework, each group contains a set of trajectories for multiple outcomes, denoted as Y^j, where $1 \leq j \leq J$. More specifically, Y_i^j represents the time measurements for the ith individual and the jth outcome. Let $P^k(Y_i^j|\Theta_k^j)$ be the distribution of that vector conditional on group k and the parameters Θ_k^j.

7.1.1 Constrained model

In this model, we posit the existence of K distinct groups to represent the combined developmental trajectories of Y^j, where $1 \leq j \leq J$. As in the model with a single outcome, we assume that, within a given group, Y_i^j are independently distributed. This assumption implies that the behavior of the outcomes is determined by certain intrinsic characteristics specific to each group. It's important to note that at the population level, the outcomes are not independent. However, the conditional independence of the time measurements for each outcome within a given group is preserved.

Therefore, the joint likelihood for each individual conditional on group k is as follows:

$$P^{(}Y_i^1, \ldots, Y_i^J) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} P^k(Y_i^j|\Theta_{k_j}^j) \tag{7.3}$$

$$= \sum_{k=1}^{K} \pi_k \prod_{j=1}^{J} \prod_{t=1}^{T^j} p^k(y_{it}^j|\Theta_{k_j}^j) \tag{7.4}$$

7.1.2 Unconstrained model

In Nagin's book, he initially worked with two groups, but we can extend his method to accommodate several groups. In this model, we assume that the trajectories of an outcome can be related to the trajectories of other outcomes. Specifically, we consider a scenario where there are J outcomes, and each outcome Y^j can be assigned to one of the K_j groups.

Similar to previous models, we maintain the assumption that, within a given set of groups, the individual measurements Y_i^j are independently distributed. In other words, conditional on $(k_1, \ldots, k_J) \in [\![1; K_1]\!] \times \cdots \times [\![1; K_J]\!]$, the joint probability is expressed as $P^{k_1 \ldots k_J}(Y_i^1, \ldots, Y_i^J) = \prod_{j=1}^J P^{k_j}(Y_i^j)$. Additionally, the conditional independence of time measurements for each outcome within a given group is preserved.

It's important to note that in this model, the joint probability is not simply the product of individual probabilities. This is because the outcomes are interconnected. Therefore, the joint likelihood for each individual, conditional on group k, is as follows:

$$P(Y_i^1, \ldots, Y_i^J) = \sum_{(k_1,\ldots,k_J) \in K_1 \times \cdots \times K_J} \pi_{k_1 \ldots k_J} \prod_{j=1}^{J} P^{k_j}(Y_i^j|\Theta_{k_j}^j) \tag{7.5}$$

$$= \sum_{(k_1,\ldots,k_J) \in K_1 \times \cdots \times K_J} \pi_{k_1 \ldots k_J} \prod_{j=1}^{J} \prod_{t=1}^{T^j} p^{k_j}(y_{it}^j|\Theta_{k_j}^j) \tag{7.6}$$

that we can rewrite with conditional probabilities

$$P(Y_i^1,\dots,Y_i^J) = \sum_{(k_1,\dots,k_J)\in K_1\times\cdots\times K_J} \pi_{k_J|k_1\dots k_{J-1}} \times \cdots \times \pi_{k_2|k_1} \times \pi_{k_1} \prod_{j=1}^{J}\prod_{t=1}^{T^j} p^{k_j}(y_{it}^j|\Theta_{k_j}^j) \quad (7.7)$$

where $\pi_{k_j|k_1\dots k_{j-1}}$ is the membership probability to belonging to the group j conditional to the membership to groups 1 to $j-1$.

7.1.3 Membership probability

In the constrained model, the model for membership probability remains the same as in the single outcome model. We employ multinomial logit functions, either with or without covariates. However, in the unconstrained model, we must use multinomial logit functions as well, both with or without covariates. The key difference is that we need to introduce multiple probabilities, one for each conditional probability.

For a model without covariates, we must calculate:

$$\pi_{k_1} = \frac{e^{\theta_{k_1}}}{\sum\limits_{k_1=1}^{K_1} e^{\theta_{k_1}}}, \quad \pi_{k_2|k_1} = \frac{e^{\theta_{k_2}^{k_1}}}{\sum\limits_{k_2=1}^{K_2} e^{\theta_{k_2}^{k_1}}}, \quad \dots \quad , \quad \pi_{k_J|k_1\dots k_{J-1}} = \frac{e^{\theta_{k_J}^{k_1\dots k_{J-1}}}}{\sum\limits_{k_J=1}^{K_J} e^{\theta_{k_J}^{k_1\dots k_{J-1}}}} \quad (7.8)$$

where $\theta_{k_1}, \theta_{k_2}^{k_1}, \dots, \theta_{k_1\dots k_{J-1}}^{k_J} \in \mathbb{R}$.

When calculating the probabilities π_{k_1}, we need to determine $K_1 - 1$ parameters, one for each group, as the last one can be derived by subtracting the sum of the others from 1. For $\pi_{k_2|k_1}$, we must calculate $(K_2 - 1) \times K_1$ parameters, where we need one for each group of K_2 and for each group of the K_1 groups associated with the outcome Y_1. The same principle applies to the other membership probabilities. For instance, for $\pi_{k_J|k_1\dots k_{J-1}}$, we have to compute $(K_J - 1) \times K_{J-1} \times \cdots \times K_1$ parameters.

The number of probabilities to compute then is $\prod_{j=1}^{J} K_j - 1$.

In the model with covariates, we assume the existence of variables such as x_i that influence the probability of group k_1, $x_i^{k_1}$ that influence the probability of group k_2 given group k_1, and so on, up to $x_i^{k_1\dots k_{J-1}}$ that influence the probability of group k_J given groups $k_1,\dots,k_{J-1}$. Consequently, equations (7.8) transform into:

$$\pi_{k_1} = \frac{e^{\theta_{k_1} x_i}}{\sum\limits_{k_1=1}^{K_1} e^{\theta_{k_1} x_i}}, \quad \pi_{k_2|k_1} = \frac{e^{\theta_{k_1}^{k_1} x_i^{k_2}}}{\sum\limits_{k_2=1}^{K_2} e^{\theta_{k_2}^{k_1} x_i^{k_1}}}, \quad \ldots \quad , \quad \pi_{k_J|k_1 \ldots k_{J-1}} = \frac{e_{k_J}^{\theta^{k_1 \ldots k_{J-1}} x_i^{k_1 \ldots k_{J-1}}}}{\sum\limits_{k_J=1}^{K_J} e^{\theta_{k_J}^{k_1 \ldots k_{J-1}} x_i^{k_1 \ldots k_{J-1}}}} \tag{7.9}$$

where $\theta_{k_1} \in \mathbb{R}^{n_{\theta^1}}, \theta_{k_2}^{k_1} \in \mathbb{R}^{n_{\theta^2}}, \ldots, \theta_{k_1 \ldots k_{J-1}}^{k_J} \in \mathbb{R}^{n_{\theta^J}}$ for some integers $n_{\theta^1}, \cdots, n_{\theta^J}$.
When calculating the probabilities π_{k_1}, we need to determine $n_{\theta^1}(K_1 - 1)$ parameters, the number of covariate and intercept for each group, as the parameters of the last one can are set to zero. For $\pi_{k_2|k_1}$, we must calculate $n_{\theta^2}(K_2 - 1) \times K_1$ parameters, where we need n_{θ^2} for each group of K_2 and for each group of the K_1 groups associated with the outcome Y_1. The same principle applies to the other membership probabilities. For instance, for $\pi_{k_J|k_1 \ldots k_{J-1}}$, we have to compute $n_{\theta^J}(K_J - 1) \times K_{J-1} \times \cdots \times K_1$ parameters.
The number of parameters to compute then is $n_{\theta^1}(K_1 - 1) + \sum_{j=2}^{J} \left(n_{\theta^j}(K_j - 1) \prod_{i=1}^{j-1} K_i \right)$.

A limitation of this approach is the substantial increase in the number of parameters, making it challenging to interpret these parameters effectively.

For example, if we consider three outcomes, each containing three groups, and two covariates that influence the probabilities of group membership, we have to compute 78 parameters in total.

7.1.4 Multivariate Logit Group Based Trajectory Modeling

In this section, we utilize the Multivariate Logit (MVL) model introduced by Cox (1972) and further specified by Russell and A. Petersen (2000), Ben-Akiva and Lerman (1985) and K. Bel and Paap (2014) to calculate the probabilities, particularly using the formulation refined in those works.

We denote the covariate as $Z_i = \left(Z_i^1, \ldots, Z_i^J\right)$, which contains the group membership of individual i for each outcome $Y^1, \ldots, Y^J$, and S represents the set of all possible realizations of Z_i. Therefore, $Z_i \in [\![1; K_1]\!] \times \cdots \times [\![1; K_J]\!]$.
As previously, we consider the conditional membership probabilities for the j-th outcome given all other choices z_i^h, for $1 \leq k \leq K_j$, represented as:

$$P\left(Z_i^j = k | z_i^h \text{ for } h \neq j, X_i^j\right) = \frac{e^{B_{ik}^j}}{\sum_{l=1}^{K_j} e^{B_{il}^j}} \tag{7.10}$$

Where:

- $B_{ik}^{j} = \theta_{k}^{j} X_{i}^{j} + \sum_{h \neq j} \psi_{kz_i^h}^{jh}$
- θ_{k0}^{j} is a choice-specific intercept.
- θ_{k}^{j} is a vector corresponding to the covariate X_{i}^{j}.
- z_{i}^{h} is the group membership of individual i for the h-th outcome.
- ψ_{kl}^{jh} represents association parameters between belonging to group k for the j-th outcome and belonging to group l for the l-th outcome.

It's worth noting that if all parameters $\psi_{kl}^{jh} = 0$, we obtain the definition of π_k when the outcomes are independent. Therefore, to achieve parameter identification, we impose standard identification restrictions by selecting one group and setting the parameter values to zero. For this purpose, we choose $\theta_{1}^{j} = 0$ for all j. Additionally, as we aim to use the parameters ψ_{kl}^{jh} to describe correlations, we enforce the symmetry condition $\psi_{kl}^{jh} = \psi_{lk}^{hj}$ for all k and l. For parameter identification, we set these parameters to zero for some groups. Specifically, we impose $\psi_{1l}^{jh} = \psi_{l1}^{hj} = 0$ for all k and l. This choice aligns with the aforementioned selection and allows for a straightforward interpretation of the association parameters with odds ratios.

Theorem 7. *The joint probability is characterize as:*

$$P\left(Z_i = z_i | X_i\right) = \frac{e^{\mu_{z_i}}}{\sum_{z_i \in \mathcal{S}} e^{\mu_{z_i}}} \tag{7.11}$$

where, $\mu_{z_i} = \sum_{j=1}^{J} \left(\theta_{z_i^j}^{j} X_i^j + \sum_{h<j} \psi_{z_i^h z_i^j}^{hj} \right)$, *$X_i$ represents the J covariates X_i^j and $\mathcal{S}$ is the set of all possible realizations of Z_i.*

Proof. The probability $P(Z_i = z_i)$ is denoted as $P(z_i)$, and $P(Z_i = (1, \cdots, 1))$ is denoted as $P(\mathbb{1})$.

According to the theorem by Besag (1974a), we have:

$$\frac{P(z_i)}{P(\mathbb{1})} = \prod_{j=1}^{J} \frac{P(z_i^j | z_i^1, \ldots, z_i^{j-1}, 1, \ldots, 1)}{P(1 | z_i^j, \ldots, z_i^{j-1}, 1, \ldots, 1)} \tag{7.12}$$

In the conditional probabilities below, the value 1 indicates that all groups for the outcomes from j to J are set to the first one. The denominator in the conditional probability in equation

7.10 is the same as both the numerator and denominator in the equation above, and we can simplify the ratio as follows:

$$\frac{P(z_i^j|z_i^1,\dots,z_i^{j-1},1,\dots,1)}{P(1|z_i^j,\dots,z_i^{j-1},1,\dots,1)} = \exp\left(\theta^j_{z_i^j}X_i^j + \sum_{h<j}\psi^{hj}_{z_i^h z_i^j} + \sum_{h>j}\psi^{hj}_{1z_i^j}\right) \tag{7.13}$$

With the identification relations:

$$\psi^{hj}_{1z_j} = 0$$

we have

$$\frac{P(z_i)}{P(\mathbb{1})} = \prod_{j=1}^{J}\exp\left(\theta^j_{z_i^j}X_i^j + \sum_{h<j}\psi^{hj}_{z_i^h z_i^j}\right) \tag{7.14}$$

Furthermore

$$P(z_i) = \frac{P(z_i)/P(\mathbb{1})}{\sum\limits_{s\in\mathcal{S}} P(s)/P(\mathbb{1})} \tag{7.15}$$

where $\mathcal{S}$ is the set of all possible choice combinations. Thus,

$$P(z_i) = \frac{e^{\mu_{z_i}}}{\sum\limits_{s\in\mathcal{S}} e^{\mu_s}} \tag{7.16}$$

where $\mu_{z_i} = \sum_{j=1}^{J}\left(\theta^j_{z_i^j}X_i^j + \sum_{h<j}\psi^{hj}_{z_i^h z_i^j}\right)$ □

With the chosen identification, $z_i = (1,\dots,1) = \mathbb{1}$, we have $\mu_{\mathbb{1}} = 0$. This allows us to view equation 7.11 as a generalization of equation 2.8. Moreover, we can provide a simple interpretation of the coefficients ψ.

The log odds ratios 7.14 are given by:

$$\log\left(\frac{P(Z_i = z_i|X_i)}{P(Z_i = \mathbb{1}|X_i)}\right) = \sum_{j=1}^{J}\left(\theta^j_{z_i^j}X_i^j + \sum_{h<j}\psi^{hj}_{z_i^h z_i^j}\right) \tag{7.17}$$

Applying these result, we can write:

$$\log\left(\frac{P\left(Z_i = (1,\dots,1,z_i^j,1,\dots,1,z_i^h,1,\dots,1)\right)}{P(Z_i = \mathbb{1}|X_i)}\right) = \theta^j_{z_i^j}X_i^j + \theta^j_{z_i^h}X_i^j + \psi^{jh}_{z_i^j z_i^h} \tag{7.18}$$

$$\log\left(\frac{P\left(Z_i = (1,\dots,1,z_i^j,1,\dots,1)\right)}{P(Z_i = \mathbb{1}|X_i)}\right) = \theta^j_{z_i^j}X_i^j \tag{7.19}$$

$$\log\left(\frac{P\left(Z_i=(1,\ldots,1,z_i^h,1,\ldots,1)\right)}{P(Z_i=\mathbb{1}|X_i)}\right)=\theta^j_{z_i^h}X_i^j \tag{7.20}$$

Hence, by subtracting the equations above:

$$\psi^{jh}_{z_i^j z_i^h}=\log\left(\frac{P\left(Z_i=(1,\ldots,1,z_i^j,1,\ldots,1,z_i^h,1,\ldots,1)\right)P\left(Z_i=\mathbb{1}|X_i\right)}{P\left(Z_i=(1,\ldots,1,z_i^j,1,\ldots,1)\right)P\left(Z_i=(1,\ldots,1,z_i^h,1,\ldots,1)\right)}\right) \tag{7.21}$$

This interpretation allows us to understand the parameter $\psi^{jh}_{z_i^j z_i^h}$. If the probability of z_i^j and z_i^h moving together is greater than the probability of them moving apart, $\psi^{jh}_{z_i^j z_i^h}>0$. Conversely, if the probability of z_i^j and z_i^h moving together is lower than the probability of them moving apart, $\psi^{jh}_{z_i^j z_i^h}<0$. If z_i^j and z_i^h move apart, $\psi^{jh}_{z_i^j z_i^h}=0$.

The model can also be extended to cases where some covariate U_i influences the link between two outcomes, by replacing ψ^{jh}_{kl} in (7.10) with:

$$\psi^{jh}_{kl}=\gamma^{jh}_{kl}U_i \tag{7.22}$$

Proposition 12. *The numbers of parameters is*

$$\sum_{j=1}^{J}\left(n_{\theta^j}(K_j-1)+\sum_{1\leq j'<j}(K_j-1)(K_{j'}-1)\right) \tag{7.23}$$

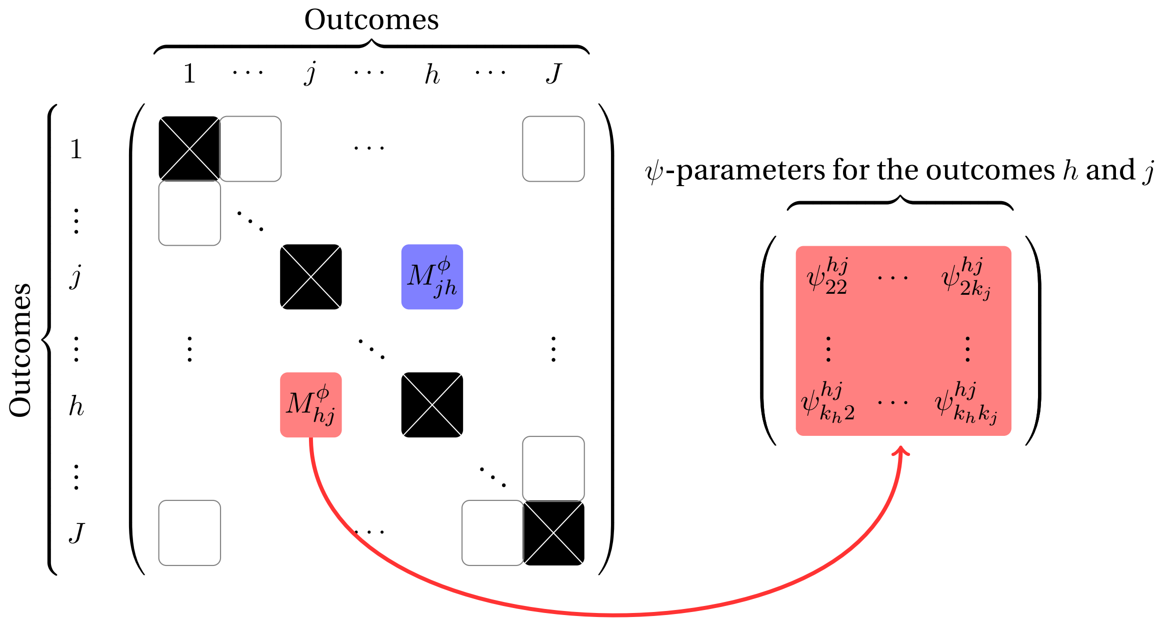
The term n_{θ^j} represents the number of covariates included for outcome j, in addition to the intercept term.

Proof. The left sum represents the total number of θ-parameters, considering the base group, while the right sum represents the number of ψ-parameters.
Let $\#\psi$ the number of ψ-parameters. We can construct a block matrix M^ψ as follows:

- A diagonal block M^ψ_{jj}, with a size of $(K_j-1)^2$, containing arbitrary values, such as 0.

- A block M^ψ_{jh}, with a size of $(K_j-1)\times(K_h-1)$, containing elements ψ^{jh}_{kl} for k and l.

We take into account the fact that the first group of each outcome is considered the base group. The assumption $\psi^{jh}_{kl}=\psi^{hj}_{lk}$ for all k and l implies that the block matrix at row j and column h is the transpose of the block matrix at row h and column j, i.e., $M^\psi_{jh}=\left(M^\psi_{hj}\right)'$. Therefore, the matrix M^ψ is symmetric.


Outcomes
1 ⋯ j ⋯ h ⋯ J
Outcomes
1
j
h
J
M^{ϕ}_{jh}
M^{ϕ}_{hj}
ψ-parameters for the outcomes h and j
ψ^{hj}_{22} ⋯ $\psi^{hj}_{2k_j}$
$\psi^{hj}_{k_h 2}$ ⋯ $\psi^{hj}_{k_h k_j}$

The number of values in M^{ϕ} is $\left(\sum_{j=1}^{J}(K_j-1)\right)^2$. Each diagonal block contains $(K_j-1)^2$ elements for $j \in [\![1;J]\!]$. Therefore, the total number of ψ-parameters is given by

$$\#\psi = \frac{\left(\sum_{j=1}^{J}(K_j-1)\right)^2 - \sum_{j=1}^{J}(K_j-1)^2}{2} \tag{7.24}$$

$$= \sum_{1\leq j\neq j'\leq J}(K_j-1)(K_{j'}-1) \tag{7.25}$$

□

7.1.5 Numerical example

We consider three outcomes, denoted as Y^1, Y^2, and Y^3, each consisting of three distinct groups. Additionally, there are two covariates that influence the probabilities of group membership. Following 7.11, the membership probabilities for these outcomes are given by:

$$\pi_{\breve{k}} = \pi_{k_1\ldots k_J} = \frac{e^{\mu_{\breve{k}}}}{\sum_{\breve{l}\in S} e^{\mu_{\breve{l}}}} \tag{7.26}$$

where $\mu_{\breve{k}} = \sum_{j=1}^{J}\left(\theta^{j}_{k_j}X^{j}_{i} + \sum_{h<j}\psi^{hj}_{k_h k_j}\right)$.

In the constrained case, where outcomes are independent of each other and the covariates are the same for all outcomes, we have $\psi^{jh}_{kl} = 0$ for all values. Consequently, the vector $\mu_{\breve{k}}$ can be expressed as:

$$\mu_{\check{k}} = \sum_{j=1}^{J} \left(\theta^{j}_{k_j} X_i\right) = \left(\sum_{j=1}^{J} \theta^{j}_{k_j}\right) X_i = \theta_k X_i \tag{7.27}$$

This results in a definition that aligns with the one presented in Chapter 1.

In the unconstrained case, where the values ψ^{jh}_{kl} are not necessarily null, we assume the first group to be the base group. According to the identification restrictions, $\psi^{jh}_{kl} = \psi^{hj}_{lk}$ and $\psi^{jh}_{1l} = \psi^{hj}_{l1} = 0$ for all k and l. In this scenario, the conditional probabilities are defined as shown in equation (7.10):

$$\begin{aligned}
&P(Z^1_i = 1|z^2_i, z^3_i, X^1_i) \propto 1\\
&P(Z^1_i = 2|z^2_i, z^3_i, X^1_i) \propto \exp\left(\theta^1_2 X^1_i + \psi^{12}_{2z^2_i} + \psi^{13}_{2z^3_i}\right)\\
&P(Z^1_i = 3|z^2_i, z^3_i, X^1_i) \propto \exp\left(\theta^1_3 X^1_i + \psi^{12}_{3z^2_i} + \psi^{13}_{3z^3_i}\right)\\
&P(Z^2_i = 1|z^1_i, z^3_i, X^2_i) \propto 1\\
&P(Z^2_i = 2|z^1_i, z^3_i, X^2_i) \propto \exp\left(\theta^2_2 X^2_i + \psi^{12}_{z1_i2} + \psi^{23}_{2z^3_i}\right)\\
&P(Z^2_i = 3|z^1_i, z^3_i, X^2_i) \propto \exp\left(\theta^2_3 X^2_i + \psi^{12}_{z^1_i2} + \psi^{23}_{3z^3_i}\right)\\
&P(Z^3_i = 1|z^1_i, z^2_i, X^3_i) \propto 1\\
&P(Z^3_i = 2|z^1_{i1}, z^2_i, X^3_i) \propto \exp\left(\theta^3_2 X^3_i + \psi^{13}_{z^1_i2} + \psi^{23}_{z^2_i2}\right)\\
&P(Z^3_i = 3|z^1_i, z^2_i, X^3_i) \propto \exp\left(\theta^3_3 X^3_i + \psi^{13}_{z^1_i3} + \psi^{23}_{z^2_i3}\right)
\end{aligned}$$

Using formula (7.23), we need to estimate 18 θ-parameters and 12 ψ-parameters, as evident from the equations above. In contrast, the Nagin model requires the estimation of 78 parameters.

To define $\pi_{\check{k}} = \pi_{k_1k_2k_3}$, we must calculate $P(Z_i = \left(z^1_i, z^2_i, z^3_i\right))$ for $z^1_i \in [\![1;3]\!]$, $z^2_i \in [\![1;3]\!]$, and $z^3_i \in [\![1;3]\!]$. This entails determining 27 joint probabilities.

Following equation (7.26), we can calculate each joint probability.

$$\begin{aligned}
&P(Z_i = (1,1,1)) \propto 1\\
&P(Z_i = (2,1,1)) \propto exp(\theta^1_2 X^1_i)\\
&P(Z_i = (1,2,1)) \propto exp(\theta^2_2 X^2_i)\\
&P(Z_i = (1,1,2)) \propto exp(\theta^3_2 X^3_i)\\
&P(Z_i = (1,1,3)) \propto exp(\theta^3_3 X^3_i)
\end{aligned}$$

$$
\begin{aligned}
&P(Z_i = (2,2,1)) \propto exp(\theta_2^1 X_i^1 + \theta_2^2 X_i^2 + \psi_{22}^{12})\\
&P(Z_i = (2,1,2)) \propto exp(\theta_2^1 X_i^1 + \theta_2^3 X_i^3 + \psi_{22}^{13})\\
&P(Z_i = (1,2,2)) \propto exp(\theta_2^2 X_i^1 + \theta_2^3 X_i^3 + \psi_{22}^{23})\\
&P(Z_i = (2,2,2)) \propto exp(\theta_2^1 X_i^1 + \theta_2^2 X_i^2 + \theta_2^3 X_i^3 + \psi_{22}^{12} + \psi_{22}^{13} + \psi_{22}^{23})\\
&\qquad\vdots
\end{aligned}
$$

We can summarize all the probabilities in the table below, reporting only the parameters.

Gr.	Param.	Gr.	Param.	Gr.	Param.
(1,1,1)	/	(2,1,1)	θ_2^1	(3,1,1)	θ_3^1
(1,1,2)	θ_2^3	(2,1,2)	$\theta_2^1, \theta_2^3, \psi_{22}^{13}$	(3,1,2)	$\theta_3^1, \theta_2^3, \psi_{32}^{13}$
(1,1,3)	θ_3^3	(2,1,3)	$\theta_2^1, \theta_3^3, \psi_{23}^{13}$	(3,1,3)	$\theta_3^1, \theta_3^3, \psi_{33}^{13}$
(1,2,1)	θ_2^2	(2,2,1)	$\theta_2^1, \theta_2^2, \psi_{22}^{12}$	(3,2,1)	$\theta_3^1, \theta_2^2, \psi_{32}^{12}$
(1,2,2)	$\theta_2^2, \theta_2^3, \psi_{22}^{23}$	(2,2,2)	$\theta_2^1, \theta_2^2, \theta_2^3, \psi_{22}^{12}, \psi_{22}^{13}, \psi_{22}^{23}$	(3,2,2)	$\theta_3^1, \theta_2^2, \theta_2^3, \psi_{32}^{12}, \psi_{32}^{13}, \psi_{22}^{23}$
(1,2,3)	$\theta_2^2, \theta_3^3, \psi_{23}^{23}$	(2,2,3)	$\theta_2^1, \theta_2^2, \theta_3^3, \psi_{22}^{12}, \psi_{23}^{13}, \psi_{23}^{23}$	(3,2,3)	$\theta_3^1, \theta_2^2, \theta_3^3, \psi_{32}^{12}, \psi_{33}^{13}, \psi_{23}^{23}$
(1,3,1)	θ_3^2	(2,3,1)	$\theta_2^1, \theta_3^2, \psi_{23}^{12}$	(3,3,1)	$\theta_3^1, \theta_3^2, \psi_{33}^{12}$
(1,3,2)	$\theta_3^2, \theta_2^3, \psi_{32}^{23}$	(2,3,2)	$\theta_2^1, \theta_3^2, \theta_2^3, \psi_{23}^{12}, \psi_{22}^{13}, \psi_{32}^{23}$	(3,3,2)	$\theta_3^1, \theta_3^2, \theta_2^3, \psi_{33}^{12}, \psi_{32}^{13}, \psi_{32}^{23}$
(1,3,3)	$\theta_3^2, \theta_3^3, \psi_{33}^{23}$	(2,3,3)	$\theta_2^1, \theta_3^2, \theta_3^3, \psi_{23}^{12}, \psi_{23}^{13}, \psi_{33}^{23}$	(3,3,3)	$\theta_3^1, \theta_3^2, \theta_3^3, \psi_{33}^{12}, \psi_{33}^{13}, \psi_{33}^{23}$

7.1.6 Differential of Likelihood

The joint likelihood for each individual is given by:

$$
L(\psi; Y_i^1, \dots, Y_i^J) = \sum_{(k_1,\dots,k_J)\in K_1\times\cdots\times K_J} \pi_{k_1\dots k_J} \prod_{j=1}^{J} P^{k_j}(Y_i^j | A_i^j, W_i^j) \tag{7.28}
$$

Since each outcome is assumed to be independent given time, in terms of density, we have:

$$
L(\psi; Y_i^1, \dots, Y_i^J) = \sum_{(k_1,\dots,k_J)\in K_1\times\cdots\times K_J} \pi_{k_1\dots k_J} \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_{k_j}^j) \tag{7.29}
$$

where g^k is the density of the outcomes j and the groups k_j. The log likelihood is then:

$$l(Y_i^1, \dots, Y_i^J) = \sum_{i=1}^{n} \log \left(\sum_{(k_1,\dots,k_J) \in K_1 \times \dots \times K_J} \pi_{k_1 \dots k_J} \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_{k_j}^j) \right) \tag{7.30}$$

Therefore, equation (2.15) provides the differential of the log likelihood of the entire dataset with respect to each parameter. For $1 \leq j \leq J$ and $1 \leq k \leq K^j$,

$$\frac{\partial l(\psi; y)}{\partial \theta_k^j} = \sum_{i=1}^{n} \frac{\sum_{(k_1,\dots,k_J) \in K_1 \times \dots \times K_J} \frac{\partial \pi_{k_1 \dots k_J}}{\partial \theta_k^j} \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_{k_j}^j)}{\sum_{(k_1,\dots,k_J) \in K_1 \times \dots \times K_J} \pi_{k_1 \dots k_J} \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_{k_j}^j)} \tag{7.31}$$

$$\frac{\partial l(\psi; y)}{\partial \Theta_k^j} = \sum_{i=1}^{n} \frac{\sum_{(k_1,\dots,k_J) \in K_1 \times \dots \times K_J} \pi_{k_1 \dots k_J} \frac{\partial}{\partial \Theta_k^j} \left[\prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_k^j) \right]}{\sum_{(k_1,\dots,k_J) \in K_1 \times \dots \times K_J} \pi_{k_1 \dots k_J} \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_{k_j}^j)} \tag{7.32}$$

In the following, we will denote

$$d_{\check{k}} = \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_k^j) \tag{7.33}$$

and

$$d_{\check{k} \setminus k_j} = \prod_{\substack{j'=1 \\ j' \neq j}}^{J} \prod_{t=1}^{T^{j'}} g^{k_{j'}}(y_{it}^{j'}; \Theta_k^{j'}) \tag{7.34}$$

7.1.6.1 Fist part

Just as in equation (7.11), we have

$$\pi_{\check{k}} = \pi_{k_1 \dots k_J} = \frac{e^{\mu_{\check{k}}}}{\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}} \tag{7.35}$$

where $\mu_{\check{k}} = \sum_{j=1}^{J} \left(\theta_{k_j}^j X_i^j + \sum_{h<j} \psi_{k_h k_j}^{hj} \right)$. Thus we have the partial differential for $l = 1, \dots, J$

$$\frac{\partial \pi_{\check{k}}}{\partial \theta_{k_j l}^j} = x_{il}^j \pi_{\check{k}} \left(\mathbb{1}_{(k_j \subset \check{k})} - \sum_{k_j \subset \check{k}} \pi_{\check{k}} \right) \tag{7.36}$$

$$\frac{\partial \pi_{\check{k}}}{\partial \psi^{jh}_{k_j k_h}} = \pi_{\check{k}} \left(\mathbb{1}_{(k_j,k_l \subset \check{k})} - \sum_{k_j,k_l \subset \check{k}} \pi_{\check{k}} \right) \tag{7.37}$$

Proof.

$$\begin{aligned}
\frac{\partial \pi_{\check{k}}}{\partial \theta^{j}_{k_j l}} &= \frac{\mathbb{1}_{(k_j \subset \check{k})} x^j_{il} e^{\mu_{\check{k}}} \sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}} - e^{\mu_{\check{k}}} \sum_{k_j \subset \check{k}} x^j_{il} e^{\mu_{\check{k}}}}{\left(\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}\right)^2} && (7.38)\\
&= x^j_{il} \left(\mathbb{1}_{(k_j \subset \check{k})} \frac{e^{\mu_{\check{k}}}}{\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}} - \frac{e^{\mu_{\check{k}}}}{\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}} \sum_{k_j \subset \check{k}} \frac{e^{\mu_{\check{k}}}}{\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}} \right) && (7.39)\\
&= x^j_{il} \left(\mathbb{1}_{(k_j \subset \check{k})} \pi_{\check{k}} - \pi_{\check{k}} \sum_{k_j \subset \check{k}} \pi_{\check{k}} \right) && (7.40)\\
\frac{\partial \pi_{\check{k}}}{\partial \psi_{jh,k_j k_h}} &= \frac{\mathbb{1}_{(k_j,k_h \subset \check{k})} e^{\mu_{\check{k}}} \sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}} - e^{\mu_{\check{k}}} \sum_{k_j,k_h \subset \check{k}} e^{\mu_{\check{k}}}}{\left(\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}\right)^2} && (7.41)\\
&= \left(\mathbb{1}_{(k_j,k_h \subset \check{k})} \frac{e^{\mu_{\check{k}}}}{\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}} - \frac{e^{\mu_{\check{k}}}}{\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}} \sum_{k_j,k_h \subset \check{k}} \frac{e^{\mu_{\check{k}}}}{\sum_{\check{l} \in \mathcal{S}} e^{\mu_{\check{l}}}} \right) && (7.42)\\
&= \left(\mathbb{1}_{(k_j,k_h \subset \check{k})} \pi_{\check{k}} - \pi_{\check{k}} \sum_{k_j,k_h \subset \check{k}} \pi_{\check{k}} \right) && (7.43)
\end{aligned}$$

□

Now, to complete the numerator in equation (7.31), let's consider a θ-parameter and introduce.

$$\begin{aligned}
\sum_{\check{k} \in \mathcal{S}} x^j_{il} \left(\mathbb{1}_{(k_j \subset \check{k})} \pi_{\check{k}} - \pi_{\check{k}} \sum_{k_j \subset \check{k}} \pi_{\check{k}} \right) d_{\check{k}} &= \sum_{k_j \subset \check{k}} x^j_{il} \pi_{\check{k}} d_{\check{k}} - \sum_{\check{k}} x^j_{il} \pi_{\check{k}} \left(\sum_{k_j \subset \check{k}} \pi_{\check{k}} \right) d_{\check{k}} && (7.44)\\
&= x^j_{il} \left(\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k}} - \left(\sum_{k_j \subset \check{k}} \pi_{\check{k}} \right) \sum_{\check{k}} \pi_{\check{k}} d_{\check{k}} \right) && (7.45)\\
&= x^j_{il} \left(\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k}} - \left(\sum_{k_j \subset \check{k}} \pi_{\check{k}} \right) \sum_{\check{k}} \pi_{\check{k}} d_{\check{k}} \right) && (7.46)
\end{aligned}$$

Similarly, for a ψ-parameter,

$$\sum_{\check{k} \in \mathcal{S}} \left(\mathbb{1}_{(k_j,k_h \subset \check{k})} \pi_{\check{k}} - \pi_{\check{k}} \sum_{k_j,k_h \subset \check{k}} \pi_{\check{k}} \right) d_{\check{k}} = \left(\sum_{k_j,k_h \subset \check{k}} \pi_{\check{k}} d_{\check{k}} - \left(\sum_{k_j,k_h \subset \check{k}} \pi_{\check{k}} \right) \sum_{\check{k}} \pi_{\check{k}} d_{\check{k}} \right) \tag{7.47}$$

In conclusion, the derivatives of the likelihood with respect to the θ-parameters or the ψ-

parameters, for $1 \leq j, h \leq J$, $1 \leq k_j \leq K^j$, $1 \leq k_h \leq K^h$ and $1 \leq l \leq n_{\theta^j}$, are

$$\frac{\partial l(\psi; y)}{\partial \theta^j_{k_j l}} = \sum_{i=1}^{n} x^j_{il} \left(\frac{\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k}}}{\sum_{\check{k} \in \mathcal{S}} \pi_{\check{k}} d_{\check{k}}} - \sum_{k_j \subset \check{k}} \pi_{\check{k}} \right) \tag{7.48}$$

$$\frac{\partial l(\psi; y)}{\partial \psi_{jh,k_j k_h}} = \sum_{i=1}^{n} \left(\frac{\sum_{k_j,k_h \subset \check{k}} \pi_{\check{k}} d_{\check{k}}}{\sum_{\check{k} \in \mathcal{S}} \pi_{\check{k}} d_{\check{k}}} - \sum_{k_j,k_h \subset \check{k}} \pi_{\check{k}} \right) \tag{7.49}$$

7.1.6.2 Second part

In each situation, the parameters are separable given outcomes J. Thus, by omitting the conditional condition to simplify the expression, we have:

$$\frac{\partial}{\partial \Theta^j_k} \left[\prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y^j_{it}; \Theta^j_{k_j}) \right] = \left[\prod_{\substack{j'=1 \\ j' \neq j}}^{J} \prod_{t=1}^{T^{j'}} g^{k_{j'}}(y^{j'}_{it}; \Theta^{j'}_{k_{j'}}) \right] \times \underbrace{\left[\frac{\partial}{\partial \Theta^j_k} \left(\prod_{t=1}^{T^j} g^{k_j}(y^j_{it}; \Theta^j_{k_j}) \right) \right]}_{\text{①}} \tag{7.50}$$

where ① is calculated as in the chapter 3 for $k_j \in \check{k}$, otherwise, the derivative is null. Thus, for all distributions, for $1 \leq j \leq J$ and $1 \leq k_j \leq K^j$,

$$\frac{\partial l(\psi; y)}{\partial \beta^j_{k_j l}} = \sum_{i=1}^{n} \frac{\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k} \backslash k_j} \frac{\partial}{\partial \beta^j_{k_j l}} \left(g^{k_j}(y^j_{it}; \Theta^j_{k_j}) \right)}{\sum_{\check{k} \in \mathcal{S}} \pi_{\check{k}} d_{\check{k}}} \tag{7.51}$$

$$\frac{\partial l(\psi; y)}{\partial \delta^j_{k_j l}} = \sum_{i=1}^{n} \frac{\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k} \backslash k_j} \frac{\partial}{\partial \delta^j_{k_j l}} \left(g^{k_j}(y^j_{it}; \Theta^j_{k_j}) \right)}{\sum_{\check{k} \in \mathcal{S}} \pi_{\check{k}} d_{\check{k}}} \tag{7.52}$$

For the Censored Normal distribution

$$\frac{\partial l(\psi; y)}{\partial \sigma^j_{k_j}} = \sum_{i=1}^{n} \frac{\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k} \backslash k_j} \frac{\partial}{\partial \sigma^j_{k_j}} \left(g^{k_j}(y^j_{it}; \Theta^j_{k_j}) \right)}{\sum_{\check{k} \in \mathcal{S}} \pi_{\check{k}} d_{\check{k}}} \tag{7.53}$$

and if we want same sigma for each group,

$$\frac{\partial l(\psi; y)}{\partial \sigma^j} = \sum_{i=1}^{n} \frac{\sum_{k_j=1}^{K^j} \left(\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k} \backslash k_j} \frac{\partial}{\partial \sigma^j} \left(g^{k_j}(y^j_{it}; \Theta^j_{k_j}) \right) \right)}{\sum_{\check{k} \in \mathcal{S}} \pi_{\check{k}} d_{\check{k}}} \tag{7.54}$$

For the ZIP distribution

$$\frac{\partial l(\psi;y)}{\partial \nu^j_{k_j l}} = \sum_{i=1}^{n} \frac{\displaystyle\sum_{k_j \subset \check{k}} \pi_{\check{k}} d_{\check{k}\backslash k_j} \frac{\partial}{\partial \nu^j_{k_j l}} \left(g^{k_j}(y^j_{it};\Theta^j_{k_j}) \right)}{\sum_{\check{k}\in\mathcal{S}} \pi_{\check{k}} d_{\check{k}}} \tag{7.55}$$

The function g^{k_j} and its derivative with respect to a parameter have the form as shown on section 3.1 for the censored normal distribution, section 3.2 for the logistic distribution, and section 3.3 for the ZIP distribution.

7.1.7 EM algorithm

Let $\boldsymbol{Y}_i = \left(Y_i^1,\dots,Y_i^J\right)$, $\boldsymbol{Z}_i = \left\{(z_{i\boldsymbol{k}} = z_{i(k_1,\dots k_J)}); k_1 \in [\![1;K_1]\!] \times \cdots \times k_J \in [\![1;K_J]\!]\right\}$ a sequence of zero values except for one value equal to 1, $\boldsymbol{S}_i = \left(S_i^1,\dots,S_i^J\right)$ a covariate with values of 0 or 1, and in the case of the ZIP distribution, it indicates if the individual is in the Poisson state or the excess zero state for all time values t. Thus, $\boldsymbol{S}_i = \left(S_i^1,\dots,S_i^J\right)$, where $\tilde{S}_i^j = \begin{cases} S_i^j & \text{if the density of the outcome } j \text{ is ZIP} \\ Y_i^j & \text{else} \end{cases}$. Since $\left(Y_i^j, S_i^j\right)$ are independent given Z_i^j, $(\boldsymbol{Y}_i, \boldsymbol{S}_i)$ are independent given $\boldsymbol{Z}_i$. The complete data associated with $\boldsymbol{Y}_i$ is $(\boldsymbol{Z}_i, \tilde{\boldsymbol{S}}_i, \boldsymbol{Y}_i)$, and the complete likelihood is:

$$L_C = \prod_{i=1}^{n} P\left(\tilde{\boldsymbol{S}}_i, \boldsymbol{Z}_i, \boldsymbol{Y}_i\right) = \prod_{i=1}^{n} P\left(\tilde{\boldsymbol{S}}_i, \boldsymbol{Y}_i|\boldsymbol{Z}_i\right) P\left(\boldsymbol{Z}_i\right) \tag{7.56}$$

$$= \prod_{i=1}^{n} \left[\prod_{\substack{j=1 \\ LOGIT \\ CNORM}}^{J} P\left(Y_i^j|\boldsymbol{Z}_i\right) \prod_{\substack{j=1 \\ ZIP}}^{J} P\left(Y_i^j, S_i^j|Z_i^j\right) \right] P\left(\boldsymbol{Z}_i\right) \tag{7.57}$$

$$\tag{7.58}$$

Here, the variables Z_i^j are not dependent each other but the Y_i^j are. We have :

$$P\left(\boldsymbol{Z}_i\right) = P\left(Z_i^1 = k_1,\dots,Z_i^J = k_J\right) = \prod_{\check{k}\in\mathcal{S}} \pi_{\check{k}}^{z_{ik}} \tag{7.59}$$

Thus

$$L_C = \prod_{i=1}^{n} \prod_{\substack{j=1 \\ LOGIT \\ CNORM}}^{J} P\left(Y_i^j|\boldsymbol{Z}_i\right) \prod_{\substack{j=1 \\ ZIP}}^{J} P\left(Y_i^j, S_i^j|Z_i^j\right) \prod_{\check{k}\in\mathcal{S}} \pi_{\check{k}}^{z_{i\check{k}}} \tag{7.60}$$

$$= \prod_{i=1}^{n} \prod_{\check{k} \in \mathcal{S}} \left(\prod_{\substack{j=1 \\ LOGIT \\ CNORM}}^{J} P\left(Y_i^j | \boldsymbol{Z}_i\right) \prod_{\substack{j=1 \\ ZIP}}^{J} P\left(Y_i^j, S_i^j | Z_i^j\right) \pi_{\check{k}} \right)^{z_{i\check{k}}} \tag{7.61}$$

and the complete log likelihood is

$$l_C = \sum_{i=1}^{n} \sum_{\check{k} \in \mathcal{S}} z_{i\check{k}} \log \left(\pi_{\check{k}} \prod_{\substack{j=1 \\ LOGIT \\ CNORM}}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_k^j) \prod_{\substack{j=1 \\ ZIP}}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j, s_{it}^j; \Theta_k^j) \right) \tag{7.62}$$

$$= \sum_{i=1}^{n} \sum_{\check{k} \in \mathcal{S}} z_{i\check{k}} \left[\log\left(\pi_{\check{k}}\right) + \sum_{\substack{j=1 \\ LOGIT \\ CNORM}}^{J} \sum_{t=1}^{T^j} \log\left(g^{k_j}(y_{it}^j; \Theta_k^j)\right) + \sum_{\substack{j=1 \\ ZIP}}^{J} \sum_{t=1}^{T^j} \log\left(g^{k_j}(y_{it}^j, s_{it}^j; \Theta_k^j)\right) \right] \tag{7.63}$$

As in formulas (2.44) and (2.45), for each step in the EM algorithm, we have:

$$E(z_{i\check{k}} | \boldsymbol{Y}_i) = \tau_{i\check{k}} = \frac{\pi_{\check{k}} \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_k^j)}{\sum_{\check{k} \in \mathcal{S}} \pi_{\check{k}} \prod_{j=1}^{J} \prod_{t=1}^{T^j} g^{k_j}(y_{it}^j; \Theta_k^j)} \tag{7.64}$$

and

$$Q = \sum_{i=1}^{n} \sum_{\check{k} \in \mathcal{S}} \tau_{i\check{k}} \left[\log\left(\pi_{\check{k}}\right) + \sum_{\substack{j=1 \\ LOGIT \\ CNORM}}^{J} \sum_{t=1}^{T^j} \log\left(g^{k_j}(y_{it}^j; \Theta_k^j)\right) + \sum_{\substack{j=1 \\ ZIP}}^{J} \sum_{t=1}^{T^j} \log\left(g^{k_j}(y_{it}^j, s_{it}^j; \Theta_k^j)\right) \right] \tag{7.65}$$

Given that the parameters Θ_k^j in the densities g^{k_j} are independent, to fit them, we need to solve:

$$\arg\min_{\Theta_k^j} \sum_{i=1}^{n} \sum_{\check{k} \in \mathcal{S}} \sum_{t=1}^{T^j} \tau_{i\check{k}} \log\left(g^{k_j}(y_{it}^j; \Theta_k^j)\right) \tag{7.66}$$

$$= \arg\min_{\Theta_k^j} \sum_{i=1}^{n} \left[\sum_{t=1}^{T^j} \log\left(g^{k_j}(y_{it}^j; \Theta_k^j)\right) \right] \left[\sum_{\substack{\check{k} \in \mathcal{S} \\ k_j \subset \check{k}}} \tau_{i\check{k}} \right] \tag{7.67}$$

$\sum_{\substack{\check{k} \in \mathcal{S} \\ k_j \subset \check{k}}} \tau_{i\check{k}}$ can be seen as a marginal rate of k_j. In the case of ZIP, we replace $g^{k_j}(y_{it}^j; \Theta_k^j)$ with $g^{k_j}(y_{it}^j, s_{it}^j; \Theta_k^j)$. We also need to solve $E(z_{i\check{k}} S_{it}^j | \boldsymbol{Y}_i) = \tau_{ikt}^j$ as in the previous step.

$$\tau_{\check{i k} t}^{j^{(t)}} = E_{\psi^{(t)}}(Z_{i\check{k}} S_{it}^{j} | \boldsymbol{Y}_i) = \begin{cases} 0 \text{ if } y_{it}^{j} > 0 \\ \frac{\tau_{i\check{k}}^{(t)}}{1+e^{-\nu_{j,k}^{(t)} A_{it}^{j} - \lambda_{j,ikt}^{(t)}}} \text{ if } y_{it}^{j} = 0 \end{cases} = s_{ik_j t} \tau_{i\check{k}}^{(t)} \tag{7.68}$$

Let for $1 \leq j, h \leq J$, $1 \leq k_j \leq K^j$ and $1 \leq k_h \leq K^h$.

For the parameters θ_k^j

$$\begin{aligned} &\arg\min_{\theta_k^j} \sum_{i=1}^{n} \sum_{\check{k} \in \mathcal{S}} \tau_{i\check{k}} \log\left(\pi_{\check{k}}\right) && (7.69) \\ &= \arg\min_{\theta_k^j} \sum_{i=1}^{n} \sum_{\check{k} \in \mathcal{S}} \tau_{i\check{k}} \left(\mu_{\check{k}} - \log\left(\sum_{\check{k} \in \mathcal{S}} e^{\mu_{\check{k}}} \right) \right) && (7.70) \end{aligned}$$

and $\psi_{k_j k_l}^{jh}$

$$\arg\min_{\psi_{k_j k_l}^{jh}} \sum_{i=1}^{n} \sum_{\check{k} \in \mathcal{S}} \tau_{i\check{k}} \left(\mu_{\check{k}} - \log\left(\sum_{\check{k} \in \mathcal{S}} e^{\mu_{\check{k}}} \right) \right) \tag{7.71}$$

As in the section 2.3.3 we have, for

$$\begin{aligned} &\theta_{kl}^{j} \text{ is a root of } \theta_{kl}^{j} :\mapsto \sum_{i=1}^{n} \sum_{\substack{\check{k} \in \mathcal{S} \\ k \in \check{k}}} x_{il}^{j} \left(\tau_{i\check{k}} - \pi_{i\check{k}} \right) && (7.72) \\ &\psi_{k_j k_l}^{jh} \text{ is a root of } \psi_{k_j k_l}^{jh} :\mapsto \sum_{i=1}^{n} \sum_{\substack{\check{k} \in \mathcal{S} \\ k_j, k_l \in \check{k}}} \left(\tau_{i\check{k}} - \pi_{i\check{k}} \right) && (7.73) \end{aligned}$$

Furthermore, by denoting $\tau_{\Sigma} = \sum_{\substack{\check{k} \in \mathcal{S} \\ k_j \in \check{k}}} \tau_{i\check{k}}$, we can express the parameters in the specific case of:

- normal distribution $\mathcal{N}\left(\beta_k^j A_{it}^j + \delta_k^j W_{it}^j; \sigma_k^j\right)$:

$$\beta_k^{j^{(t+1)}} = \left[\sum_{i=1}^{n} \tau_{\Sigma}^{(t)} \left(Y_i^j \left(A_i^j\right)^t - \delta_k^{j^{(t)}} W_i^j \left(A_i^j\right)^t \right) \right] \left(\sum_{i=1}^{n} \tau_{\Sigma}^{(t)} \left(A_i^j \left(A_i^j\right)^t \right) \right)^{-1} \tag{7.74}$$

$$\delta_k^{j^{(t+1)}} = \left[\sum_{i=1}^{n} \tau_{\Sigma}^{(t)} \left(Y_i^j \left(W_i^j\right)^t - \beta_k^{j^{(t)}} A_i^j \left(W_i^j\right)^t \right) \right] \left(\sum_{i=1}^{n} \tau_{\Sigma}^{(t)} \left(W_i^j \left(W_i^j\right)^t \right) \right)^{-1} \tag{7.75}$$

$$\sigma_k^{j^{(t+1)}} = \sqrt{\frac{\sum_{i=1}^{n} \tau_{\Sigma}^{(t)} \left(Y_i^j - \left(\beta_k^{j^{(t)}} A_i^j + \delta_k^{j^{(t)}} W_i^j \right) \right)^t \left(Y_i^j - \left(\beta_k^{j^{(t)}} A_i^j + \delta_k^{j^{(t)}} W_i^j \right) \right)}{T \sum_{i=1}^{n} \tau_{\Sigma}^{(t)}}} \tag{7.76}$$

In the case of the same value of σ within each cluster

$$\sigma^{j^{(t+1)}} = \sqrt{\frac{\sum_{k=1}^{K_j}\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\left(Y_i^j - \left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right)\right)^t\left(Y_i^j - \left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right)\right)}{T\sum_{k=1}^{K_j}\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}}} \tag{7.77}$$

- censored normal distribution $\mathcal{N}\left(\beta_k^j A_{it}^j + \delta_k^j W_{it}^j; \sigma_k^j\right)$:

$$\beta_k^{j^{(t+1)}} = \left[\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\left(\tilde{Y}_i^j\left(A_i^j\right)^t - \delta_k^{j^{(t)}}W_i^j\left(A_i^j\right)^t\right)\right]\left(\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\left(A_i^j\left(A_i^j\right)^t\right)\right)^{-1} \tag{7.78}$$

$$\delta_k^{j^{(t+1)}} = \left[\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\left(\tilde{Y}_i^j\left(W_i^j\right)^t - \beta_k^{j^{(t)}}A_i^j\left(W_i^j\right)^t\right)\right]\left(\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\left(W_i^j\left(W_i^j\right)^t\right)\right)^{-1} \tag{7.79}$$

$$\sigma_k^{j^{(t+1)}} = \sqrt{\frac{\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\left(\tilde{Y_{i,2}^{j}}^{t}\mathbb{1}_T - 2\tilde{Y_i^j}^{t}\left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right) + \left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right)^t\left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right)\right)}{T\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}}} \tag{7.80}$$

In the case of the same value of σ within each cluster

$$\sigma^{j^{(t+1)}} = \sqrt{\frac{\sum_{k=1}^{K_j}\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\left(\tilde{Y_{i,2}^{j}}^{t}\mathbb{1}_T - 2\tilde{Y_i^j}^{t}\left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right) + \left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right)^t\left(\beta_k^{j^{(t)}}A_i^j + \delta_k^{j^{(t)}}W_i^j\right)\right)}{T\sum_{k=1}^{K_j}\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}}} \tag{7.81}$$

where $\tilde{Y_i}$, $\tilde{Y_{i,2}}$ and $\mathbb{1}_T$ are respectively defined pages 38 and 39.

- logistic distribution

$$\beta_{kl}^j \text{ root of } \sum_{i=1}^{n}\sum_{t=1}^{T}\tau_{\Sigma}^{(t)}a_{it}^{j,l-1}\left(y_{it}^j - \frac{e^{\beta_k^{j^{(t)}}A_{it}^j + \delta_k^{j^{(t)}}W_{it}^j}}{1 + e^{\beta_k^{j^{(t)}}A_{it}^j + \delta_k^{j^{(t)}}W_{it}^j}}\right) \tag{7.82}$$

$$\delta_{kl}^j \text{ root of } \sum_{i=1}^{n}\sum_{t=1}^{T}\tau_{\Sigma}^{(t)}w_{it}^{j,l}\left(y_{it}^j - \frac{e^{\beta_k^{j^{(t)}}A_{it}^j + \delta_k^{j^{(t)}}W_{it}^j}}{1 + e^{\beta_k^{j^{(t)}}A_{it}^j + \delta_k^{j^{(t)}}W_{it}^j}}\right) \tag{7.83}$$

- Zero Inflated Poisson distribution

$$\beta_k^{j^{(t+1)}} = \arg\max_{\beta_k^j}\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\sum_{t=1}^{T}(1 - s_{ik_jt}^{(t)})\left(Y_{it}^j\left(\beta_k^j A_{it}^j + \delta_k^j W_t^j\right) - e^{\beta_k^j A_{it}^j + \delta_k^j W_t^j}\right) \tag{7.84}$$

$$\delta_k^{j^{(t+1)}} = \arg\max_{\delta_k^j}\sum_{i=1}^{n}\tau_{\Sigma}^{(t)}\sum_{t=1}^{T}(1 - s_{ik_jt}^{(t)})\left(Y_{it}^j\left(\beta_k^j A_{it}^j + \delta_k^j W_t^j\right) - e^{\beta_k^j A_{it}^j + \delta_k^j W_t^j}\right) \tag{7.85}$$

$$\nu_k^{j(t+1)} = \arg\max_{\nu_k^j} \sum_{i=1}^{n} \tau_{\Sigma}^{(t)} \sum_{t=1}^{T} s_{ik_jt}^{(t)} \nu_k^j A_{it}^j - \tau_{ik}^{(t)} \log\left(1 + e^{\nu_k^j A_{it}^j}\right) \qquad (7.86)$$

These formulas are derived from the equations presented in Chapter 3.

7.1.8 Numerical example

7.1.8.1 Normal - normal - normal

We use artificial data to illustrate the behavior of our algorithm. We consider 3 outcomes, denoted as Y^1, Y^2, and Y^3, each divided into 3 groups. These outcomes follow a normal distribution with means $\beta_{k_j}^j X_i^j$ and standard deviations $\sigma_{k_j}^j$ for each individual i. Here, X_i^j represents a column vector of ones. In this context, the β-parameters exclusively represent intercepts. The model parameters are given by:

- $Y^1 : \beta_1^1 = (3.53, -2.25, 0.47)$, $\beta_2^1 = (-1.62, 3.9, -0.65)$, $\beta_3^1 = (0.263, 0.036, 0.01)$, $\sigma_1^1 = \sigma_2^1 = \sigma_3^1 = 0.5$;
- $Y^2 : \beta_1^2 = (0.015, -0.11, 0.2)$, $\beta_2^2 = (3.9, 3, -0.8)$, $\beta_3^2 = (4.7, 0.1, -0.1)$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 0.5$;
- $Y^3 : \beta_1^3 = (0, 0, 0)$, $\beta_2^3 = (2.5, 0, 0)$, $\beta_3^3 = (5, 0, 0)$, $\sigma_1^3 = \sigma_2^3 = \sigma_3^3 = 0.5$.

Furthermore, we set all θ_k^j values to 0 and establish a connection between outcomes 1 and 2 by adjusting the parameter $\psi = (1.5, -2, -2, 1.5, -2, -2, -2, -2, -2, -2, -2, -2)$. As a result, the probabilities of group membership are defined as follows:

Group 1	Group 2	Group 3	probability
1	1	1	0.05664
2	1	1	0.05664
3	1	1	0.05664
1	2	1	0.05664
2	2	1	0.25385
3	2	1	0.00767
1	3	1	0.05664
2	3	1	0.00767
3	3	1	0.25385
1	1	2	0.05664
2	1	2	0.00767
3	1	2	0.00767
1	2	2	0.00767
2	2	2	0.00465
3	2	2	0.00014
1	3	2	0.00767
2	3	2	0.00014
3	3	2	0.00465
1	1	3	0.05664
2	1	3	0.00767
3	1	3	0.00767
1	2	3	0.00767
2	2	3	0.00465
3	2	3	0.00014
1	3	3	0.00767
2	3	3	0.00014
3	3	3	0.00465

We can observe that group 1 in outcome 1 is linked to group 1 in outcome 2 ($P\left(Y_i^1=1,Y_i^2=1\right)=0.16992$), group 2 in outcome 1 is linked to group 2 in outcome 2 ($P\left(Y_i^1=2,Y_i^2=2\right)=0.26315$), and group 3 in outcome 1 is linked to group 3 in outcome 2 ($P\left(Y_i^1=3,Y_i^2=3\right)=0.26315$). We have simulated 200 individuals for each outcome. In the first step, we use `trajeR` to fit the model for each outcome independently and use the results as a starting point for the algorithms.

```
        Outcome 1             |         Outcome 2             |         Outcome 3
-------------------------------------------------------------------------------------------
group   Parameter   Estimate  | group   Parameter   Estimate  | group   Parameter   Estimate
                              |                               |
----------------------------- + ----------------------------- + ----------------------------
    1   Intercept    0.09116  |     1   Intercept    0.16163  |     1   Intercept    2.61661
           Linear    0.17219  |            Linear   -0.22154  |            Linear    0.00171
        Quadratic   -0.00996  |         Quadratic    0.21645  |         Quadratic    -0.0091
                              |                               |
    2   Intercept    3.45455  |     2   Intercept    4.59404  |     2   Intercept    0.07683
           Linear   -2.17899  |            Linear    0.16384  |            Linear    -0.0152
        Quadratic    0.45516  |         Quadratic   -0.10973  |         Quadratic   -0.00118
                              |                               |
    3   Intercept    -1.5896  |     3   Intercept    3.69664  |     3   Intercept    4.69277
           Linear    3.83929  |            Linear    3.14766  |            Linear    0.21049
        Quadratic   -0.63617  |         Quadratic   -0.82292  |         Quadratic   -0.03043
----------------------------- + ----------------------------- + ------------------------------
    1      sigma1    0.49507  |     1      sigma1      0.494  |     1      sigma1    0.49796
    2      sigma2    0.49507  |     2      sigma2      0.494  |     2      sigma2    0.49796
    3      sigma3    0.49507  |     3      sigma3      0.494  |     3      sigma3    0.49796
----------------------------- + ----------------------------- + ------------------------------
    1         pi1       0.34  |     1         pi1       0.34  |     1         pi1        0.1
    2         pi2    0.38999  |     2         pi2       0.36  |     2         pi2      0.815
    3         pi3       0.27  |     3         pi3        0.3  |     3         pi3      0.085
----------------------------- + ----------------------------- + ------------------------------
Likelihood : -933.3929          Likelihood : -932.8815          Likelihood : -843.0038
```

As starting values for the algorithms, we use the previously mentioned values but change the order of the groups to match the theoretical case. For example, for outcome 1, the fitted group 2 corresponds to the theoretical group 1.

Parameters	Outcome 1		Outcome 2		Outcome 3	
	EM	L	EM	L	EM	L
β_{11}	3.45456	3.45456	0.16163	0.16163	0.07689	0.07689
β_{12}	-2.17900	-2.179	-0.22154	-0.22154	-0.01524	-0.01524
β_{13}	0.45516	0.45516	0.21645	0.21645	-0.00117	-0.00117
β_{21}	-1.58961	-1.58961	3.69664	3.69663	2.61661	2.61661
β_{22}	3.83930	3.8393	3.14766	3.14766	0.00170	0.00170
β_{23}	-0.63617	-0.63617	-0.82292	-0.82292	-0.00910	-0.00910
β_{31}	0.09111	0.0911	4.59404	4.59404	4.69300	4.69301
β_{32}	0.17223	0.17224	0.16384	0.16384	0.21031	0.21031
β_{33}	-0.00996	-0.00996	-0.10973	-0.10973	-0.03040	-0.03040
σ	0.49507	0.49507	0.49400	0.494	0.49796	0.49786
θ_1	0.00000	0	0.00000	0	0.00000	0
θ_2	-0.22727	-0.2272	0.21016	0.21020	0.06437	0.06444
θ_3	-0.10618	-0.10608	0.17453	0.17463	-0.16475	-0.16472

and for the parameters ψ for the EM-algorithm,

ψ^{12}

	1	2	3
1	0.00	0.00	0.00
2	0.00	1.06318	-10.4754
3	0.00	-10.50618	1.29510

ψ^{13}

	1	2	3
1	0.00	0.00	0.00
2	0.00	-2.14664	-1.69506
3	0.00	-2.57744	-2.76700

ψ^{23}

	1	2	3
1	0.00	0.00	0.00
2	0.00	-2.03381	-3.01771
3	0.00	-10.92514	-2.12342

and for the likelihood algorithm,

ψ^{12}

	1	2	3
1	0.00	0.00	0.00
2	0.00	1.06309	-14.04538
3	0.00	-14.09561	1.29504

ψ^{13}

	1	2	3
1	0.00	0.00	0.00
2	0.00	-2.14658	-1.69503
3	0.00	-2.57705	-2.76658

ψ^{23}

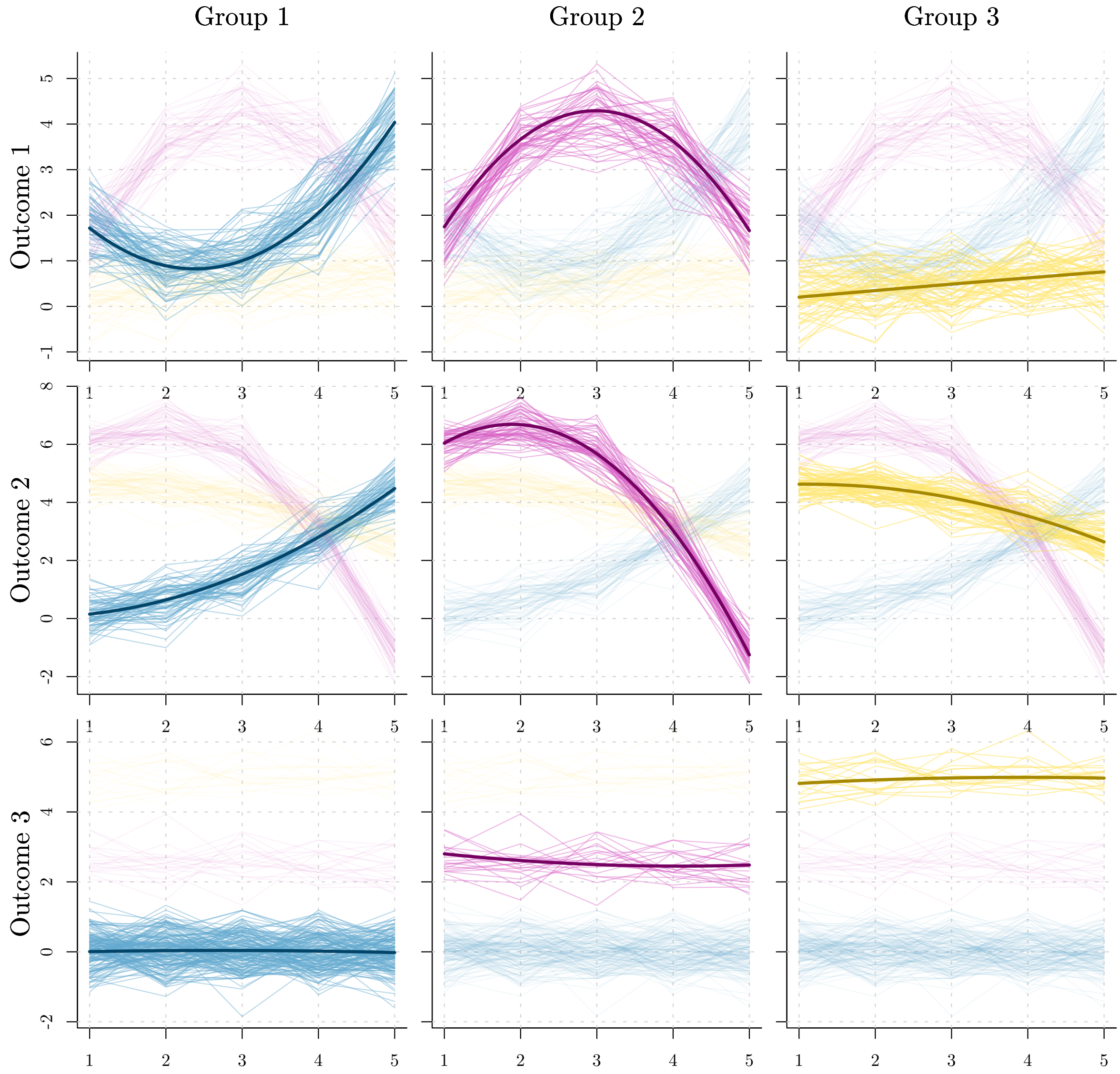
	1	2	3
1	0.00	0.00	0.00
2	0.00	-2.03375	-3.01733
3	0.00	-14.58479	-2.12344

The coefficient ψ^{12} represents the relationship between outcome 1 and outcome 2, with the first group being the base group. For example, $\psi^{12}_{22} = 1.06$, indicating that group 2 of outcome 1 is more likely to move with group 2 of outcome 2 than with group 1. Conversely, $\psi^{12}_{23} = -10.48$, suggesting that group 2 of outcome 1 rarely moves with group 3 of outcome 2. These parameters highlight the connection between groups 2 of outcome 1 and 2, as well as the groups 3 of outcome 1 and 2.

We can confirm these findings by examining the theoretical probabilities and constructing a table using the model to show the number of individuals in specific groups. For example, the model indicates that the first individual belongs to group 3 for outcome 1, group 3 for outcome 2, and group 1 for outcome 3. This supports the notion that certain groups are linked, as reflected in the parameters.

	Out. 2			Out. 3				Out. 3		
Out. 1	Gr 1	Gr 2	Gr 3	Gr 1	Gr 2	Gr 3	Out. 2	Gr 1	Gr 2	Gr 3
Gr 1	40	20	18	47	17	14	Gr 1	37	17	14
Gr 2	14	40	0	50	2	2	Gr 2	56	3	1
Gr 3	14	0	54	66	1	1	Gr 3	70	0	2

We conducted a similar analysis with a larger dataset, comprising 500 individuals, and different standard deviations within each group, where $\sigma^1 = \sigma^2 = 1$ and $\sigma^3 = 0.5$. For the parameters θ, β, and σ, we obtained the following results:


Group 1
Group 2
Group 3
Outcome 1
Outcome 2
Outcome 3
-1
0
1
2
3
4
5
-2
0
2
4
6
8
-2
0
2
4
6
1
2
3
4
5

Parameters	Outcome 1		Outcome 2		Outcome 3	
	EM	L	EM	L	EM	L
β_{11}	3.49454	3.49453	-0.14544	-0.14544	0.02716	0.02715
β_{12}	-2.23470	-2.23470	0.02445	0.02445	-0.03983	-0.03982
β_{13}	0.46777	0.46777	0.17627	0.17627	0.00902	0.00902
β_{21}	-1.67768	-1.67768	4.04774	4.04774	2.71487	2.71487
β_{22}	3.97145	3.97144	2.91493	2.91493	-0.18089	-0.18089
β_{23}	-0.66243	-0.66243	-0.78962	-0.78962	0.03207	0.03207
β_{31}	0.25214	0.25214	4.68652	4.68652	5.24299	5.24298
β_{32}	0.04869	0.04869	0.09741	0.09742	-0.13227	-0.13226
β_{33}	0.00909	0.00909	-0.09761	-0.09761	0.01754	0.01754
σ	0.49325	0.49324	0.49665	0.49665	0.49626	0.49626
θ_1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
θ_2	-0.25744	-0.25743	0.09665	0.09666	-0.22177	-0.22186
θ_3	-0.11759	-0.11752	-0.38528	-0.38527	-0.29925	-0.29919

and for the parameters ψ for the EM-algorithm,

ψ^{12}

	1	2	3
1	0.00	0.00	0.00
2	0.00	1.51337	-1.81561
3	0.00	-1.90342	1.83277

ψ^{13}

	1	2	3
1	0.00	0.00	0.00
2	0.00	-2.88368	-1.3438
3	0.00	-1.76906	-2.29668

ψ^{23}

	1	2	3
1	0.00	0.00	0.00
2	0.00	-1.97297	-2.57788
3	0.00	-1.27005	-1.68043

and for the likelihood algorithm,

ψ^{12}

	1	2	3
1	0.00	0.00	0.00
2	0.00	1.51337	-1.81635
3	0.00	-1.90350	1.83267

ψ^{13}

	1	2	3
1	0.00	0.00	0.00
2	0.00	-2.88410	-1.34378
3	0.00	-1.76904	-2.29673

ψ^{23}

	1	2	3
1	0.00	0.00	0.00
2	0.00	-1.97281	-2.57736
3	0.00	-1.27008	-1.68083

The model become more accurate.

7.1.8.2 Normal - normal - logit

In a similar experiment, we replaced the last outcome with one following a LOGIT model with 2 groups and polynomial shapes of degree 2 and 1. The parameters for this modified scenario are:

- $Y^1 : \beta_1^1 = (3.53, -2.25, 0.47)$, $\beta_2^1 = (-1.62, 3.9, -0.65)$, $\beta_3^1 = (0.263, 0.036, 0.01)$, $\sigma_1^1 = \sigma_2^1 = \sigma_3^1 = 1$;
- $Y^2 : \beta_1^2 = (0.015, -0.11, 0.2)$, $\beta_2^2 = (3.9, 3, -0.8)$, $\beta_3^2 = (4.7, 0.1, -0.1)$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$;
- $Y^3 : \beta_1^3 = (6.32, -5.8, 1)$, $\beta_2^3 = (-6.69, 1.92)$.

Additionally, we set all $\theta_k^j = 0$ and $\psi = (1.5, -2, -2, 1.5, 0, -2, -1, -2)$. We ran `trajeR` for each outcome separately and used the output as the initial values for the multi-outcome model. The results are as follows:

	Outcome 1			Outcome 2		
	Th.	EM	L	Th.	EM	L
β_{11}	3.53	3.56459	3.56460	0.015	-0.17405	-0.17406
β_{12}	-2.25	-2.31530	-2.31530	-0.110	0.03371	0.03372
β_{13}	0.47	0.48376	0.48376	0.200	0.18095	0.18095
β_{21}	-1.620	-1.68194	-1.68194	3.900	3.85913	3.85915
β_{22}	3.900	4.01305	4.01305	3.000	3.03961	3.03959
β_{23}	-0.650	-0.67322	-0.67322	-0.800	-0.80338	-0.80338
β_{31}	0.263	0.22700	0.22702	4.700	4.80305	4.80305
β_{32}	0.036	0.04902	0.04901	0.100	0.02368	0.02368
β_{33}	0.010	0.00630	0.00630	-0.100	-0.08565	-0.08565
σ	1	0.99597	0.99597	1	1.00209	1.00208
θ_1	0	0.00000	0.00000	0	0.00000	0.00000
θ_2	0	0.05413	0.05431	0	0.29607	0.29615
θ_3	0	0.16613	0.16633	0	0.09321	0.09341

	Outcome 3		
	Th.	EM	L
β_{11}	6.32	6.50352	6.50299
β_{12}	-5.8	-5.88881	-5.88849
β_{13}	1	1.01188	1.01184
β_{21}	-6.69	-6.48525	-6.48559
β_{22}	1.92	1.80831	1.80841
θ_1	0	0	0
θ_2	0	0.11835	0.1185

Parameters	ψ_{22}^{12}	ψ_{23}^{12}	ψ_{32}^{12}	ψ_{33}^{12}	ψ_{22}^{13}	ψ_{32}^{13}	ψ_{22}^{23}	ψ_{32}^{23}
Theoretical	1.5	-2	-2	1.5	0	-2	-1	-2
EM	1.27394	-2.25331	-3.41611	1.38477	-0.4814	-1.50997	-0.60894	-1.32099
Likelihood	1.27378	-2.25176	-3.41844	1.38448	-0.48173	-1.51061	-0.60897	-1.32194

7.1.8.3 Normal - ZIP - logit

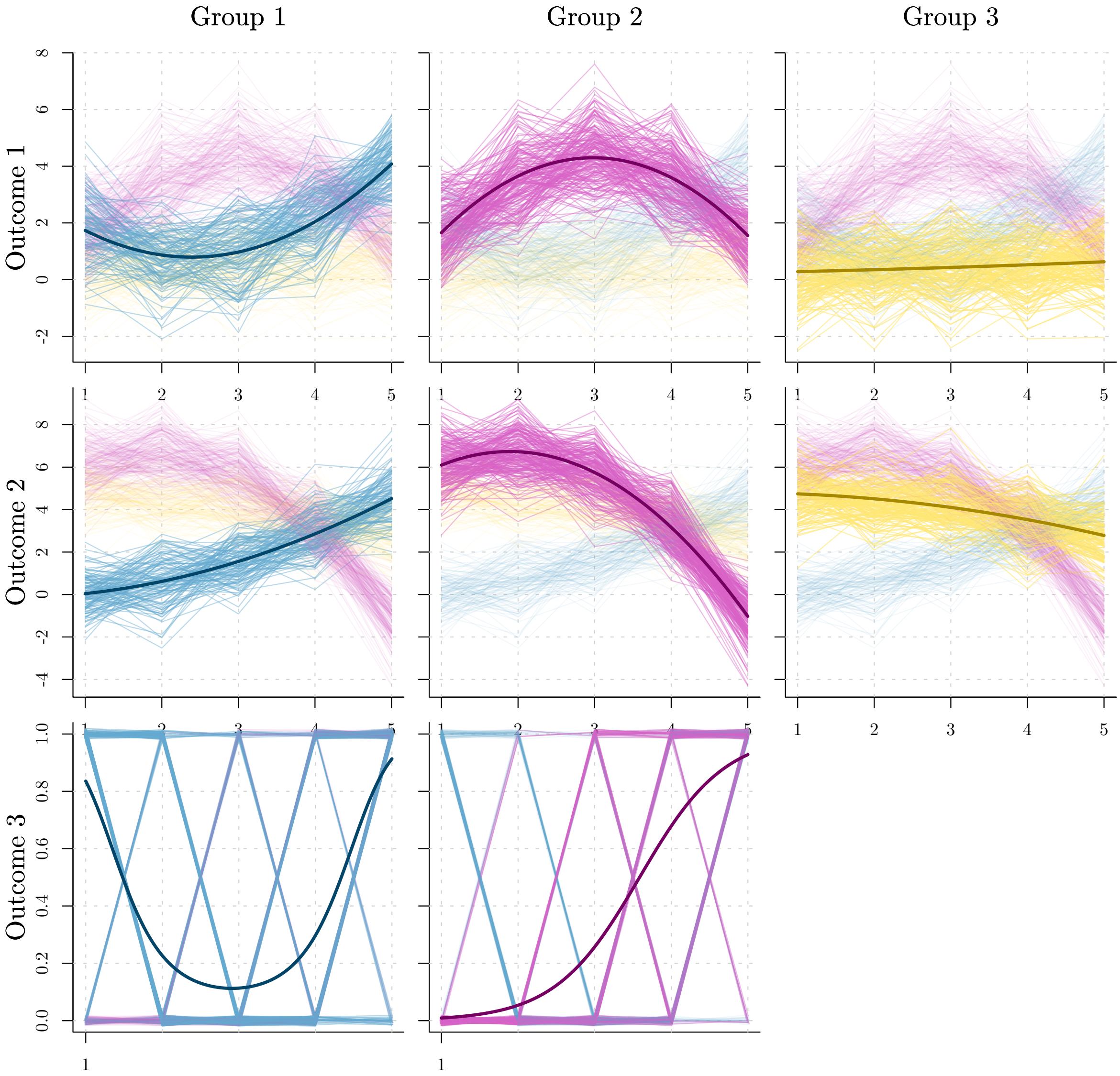
We repeat the experiment, replacing the second outcome with an outcome following a ZIP model with 3 groups. The degree of the polynomial shapes for the Poisson part is 4, 0, and 3 for the zero state in each group, and the degree of the polynomial shapes for the zero state is 2 in each group. The parameters are as follows:

- $Y^1 : \beta_1^1 = (3.53, -2.25, 0.47)$, $\beta_2^1 = (-1.62, 3.9, -0.65)$, $\beta_3^1 = (0.263, 0.036, 0.01)$, $\sigma_1^1 = \sigma_2^1 = \sigma_3^1 = 1$;

- $Y^2 : \beta_1^2 = (1.2, 2.3, -1.2, 0.5, -0.1)$, $\beta_2^2 = (2)$, $\beta_3^2 = (-7.5, 0, 2.2, -.4)$, $\nu_1 = (-2, 1)$, $\nu_2 = (-1, 0.1)$, $\nu_3 = (0, -1)$;

- $Y^3 : \beta_1^3 = (6.32, -5.8, 1)$, $\beta_2^3 = (-6.69, 1.92)$.

We also chose to set all θ_k^j to 0 and used the parameter values $\psi = (-3, 3, 4, 0, -2, 5, 1, 0)$. After running the `trajeR` algorithm for each outcome separately, we used this output as the initial value for the multi-outcome model. The results are as follows:

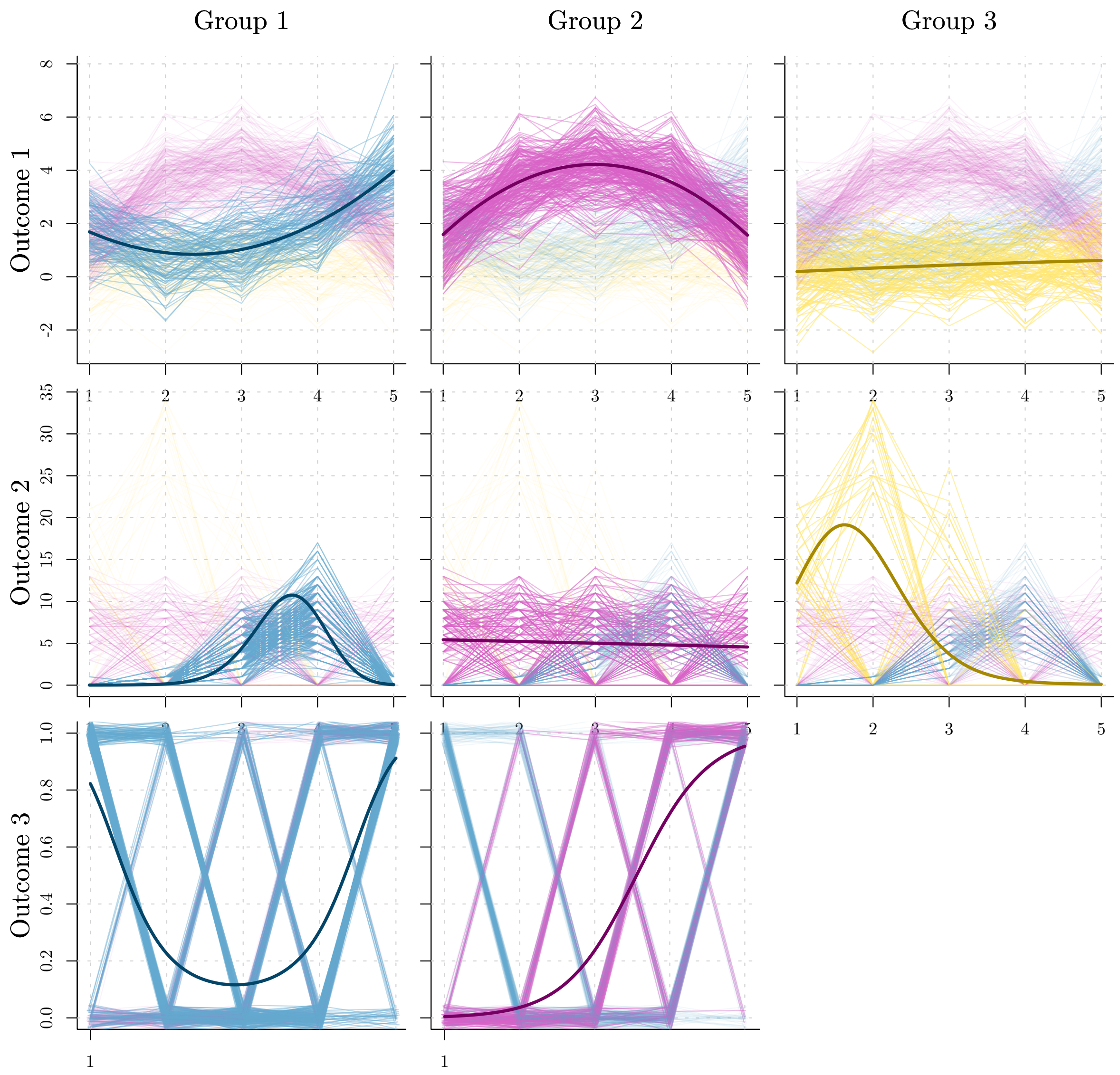

Group 1
Group 2
Group 3
Outcome 1
Outcome 2
Outcome 3
-2
0
2
4
6
8
-4
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5

	Outcome 1		
	Th.	EM	L
β_{11}	3.53	3.37162	3.37154
β_{12}	-2.25	-2.13530	-2.13523
β_{13}	0.47	0.45090	0.45089
β_{21}	-1.620	-1.71513	-1.71513
β_{22}	3.900	3.96422	3.96422
β_{23}	-0.650	-0.66183	-0.66183
β_{31}	0.263	0.04619	0.04624
β_{32}	0.036	0.15730	0.15725
β_{33}	0.010	-0.00882	-0.00881
σ	1	2.69943	2.69943
θ_1	0	0.00000	0.00000
θ_2	0	2.00645	2.00672
θ_3	0	-5.40659	-5.43048

	Outcome 2		
	Th.	EM	L
β_{11}	1.2	-3.92980	-3.92950
β_{12}	2.3	-4.11672	-4.11653
β_{13}	-1.2	3.88076	3.88051
β_{14}	0.5	-0.68312	-0.68305
β_{15}	-0.1	0.01616	0.01615
β_{21}	2	2.01483	2.01483
β_{31}	-7.5	-0.39232	-0.39204
β_{32}	0	4.55945	4.55895
β_{33}	2.2	-1.69650	-1.69623
β_{34}	-0.4	0.17702	0.17698
ν_{11}	-2	-0.98558	-0.98296
ν_{12}	1	-0.7915	-0.79232
ν_{21}	-1	-1.08749	-1.08734
ν_{22}	0.1	0.13039	0.13035
ν_{31}	0	-3.38530	-3.38515
ν_{32}	-1	1.53658	1.53652
θ_1	0	0.00000	0.00000
θ_2	0	-0.07507	-0.07461
θ_3	0	0.14593	0.14628

	Outcome 3		
	Th.	EM	L
β_{11}	6.32	6.28035	6.28024
β_{12}	-5.8	-5.73551	-5.73545
β_{13}	1	0.98963	0.98962
β_{21}	-6.69	-7.46613	-7.46609
β_{22}	1.92	2.09930	2.09930
θ_1	0	0.00000	0.00000
θ_2	0	-2.09538	-2.09295

Parameters	ψ_{22}^{12}	ψ_{23}^{12}	ψ_{32}^{12}	ψ_{33}^{12}	ψ_{22}^{13}	ψ_{32}^{13}	ψ_{22}^{23}	ψ_{32}^{23}
Theoretical	-3	3	4	0	-2	5	1	0
EM	-5.47292	4.33422	-4.57086	-4.46658	-2.12297	3.23218	1.18008	-7.96007
Likelihood	-5.46968	4.35722	-4.57081	-3.3298	-2.12013	3.23351	1.17703	-10.00915



7.1.8.4 Real data

We'll illustrate the use of this method with real data from a study by Jones and Nagin (available at `https://www.andrew.cmu.edu/user/bjones/strtxmpl2.htm`). The data comes from the Montreal Longitudinal Study, and we'll explore the potential link between hyperactivity and opposition scores. The hyperactivity scale ranges from 0 to 4, while the opposition behavior score ranges from 0 to 10.

In the figure 7.1, we've plotted the data for each outcome, and to aid in understanding, we've represented two trajectories in different colors: To obtain starting values for the joint model, we launched `trajeR` on the two behaviors separately, using single models. We won't go into detail about the choice of polynomial degree and the number of groups for the model, but we obtained the following starting values:

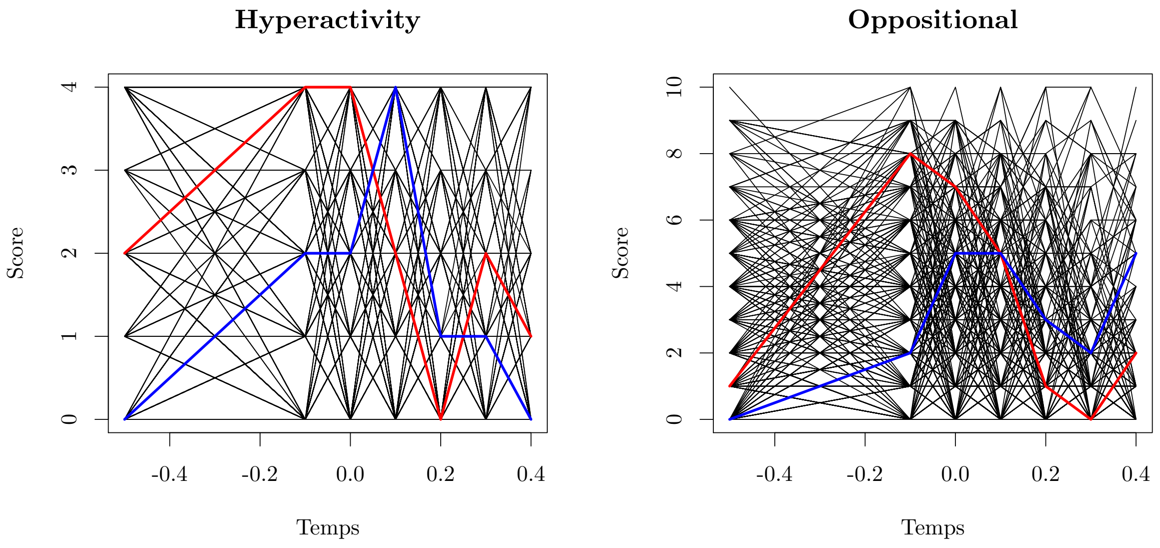


Figure 7.1: 3 outcomes. 3 groups for the two first and two for the third.

	Outcome 1	
	TrajeR	traj (SAS)
β_{11}	-2.69733	-2.69735
β_{21}	0.18631	0.18638
β_{22}	-6.36477	-6.36393
β_{23}	-1.33906	-1.34019
β_{24}	17.41459	17.4095
β_{31}	2.69890	2.69893
β_{32}	-2.21390	-2.21387
β_{33}	-5.57631	-5.57613
σ	2.32828	2.32828
π_1	0.24064	0.24064
π_2	0.41862	0.41863
π_3	0.34074	0.34073

	Outcome 2	
	TrajeR	traj (SAS)
β_{11}	-1.81222	-1.81222
β_{12}	-1.98170	-1.98171
β_{21}	1.99617	1.99617
β_{22}	-6.55146	-6.55134
β_{23}	-4.16659	-4.16671
β_{24}	20.70683	20.70610
β_{31}	4.84779	4.84780
β_{32}	-2.31239	-2.31240
β_{33}	-9.48859	-9.48858
β_{41}	6.56601	6.56603
σ	2.54834	2.54834
π_1	0.25657	0.25657
π_2	0.43094	0.43094
π_3	0.24189	0.2419
π_4	0.0706	0.0706

With these starting values in hand, we can proceed to launch the joint model. The results indicate:

	Outcome 1	
	TrajeR	traj (SAS)
β_{11}	-2.55343	-2.55348
β_{21}	0.48071	0.48074
β_{22}	-6.24514	-6.24425
β_{23}	-3.53665	-3.53676
β_{24}	14.90876	14.90347
β_{31}	2.66416	2.66411
β_{32}	-1.98842	-1.98840
β_{33}	-4.14192	-4.14164
σ	2.31949	2.3195

	Outcome 2	
	TrajeR	traj (SAS)
β_{11}	-1.67252	-1.67258
β_{12}	-1.48300	-1.48308
β_{21}	2.15443	2.15440
β_{22}	-6.79570	-6.79186
β_{23}	-4.29020	-4.29356
β_{24}	20.58775	20.56512
β_{31}	4.96622	4.96615
β_{32}	-2.12531	-2.12503
β_{33}	-9.72094	-9.71914
β_{41}	6.59242	6.59246
σ	2.54359	2.5436

With the `trajeR` package, you have 6 parameters of links between each group of each outcome. These parameters are essential for understanding the relationships and associations between different groups in your data.

Parameters	ψ_{22}^{12}	ψ_{23}^{12}	ψ_{24}^{12}	ψ_{32}^{12}	ψ_{33}^{12}	ψ_{34}^{12}
	19.97949	10.51569	8.94481	28.44746	31.69029	40.0845

The first three ψ-parameters indicate that the group 2 of outcome 1 most often varies with the group 2 of outcome 2. In contrast, ψ_{32}^{12} suggests that the group 2 of outcome 2 varies most often with the group 3 of outcome 1. Regarding the last three parameters, they imply that the group 3 of outcome 1 and the group 4 of outcome 2 vary more frequently together. However, it's worth noting that the proportion of group 4 in outcome 2 is relatively small.

Traj SAS computes only the membership probabilities, joint probabilities, and conditional probabilities. We are comparing these results with the two algorithms.

Membership's probabilities for the outcome 1

	Group 1	Group 2	Group 3
TrajeR	26.40014	42.02427	31.57559
Traj SAS	26.39963	42.02506	31.57531

Membership's probabilities for the outcome 2

	Group 1	Group 2	Group 3	Group 4
TrajeR	27.69402	43.01798	22.39573	6.89226
Traj SAS	27.7	43	22.4	6.9

Conditional probabilities, groups of outcome 1 given groups of outcome 2

Prob. group out. 1$\mid$group out. 2	$1\mid 1$	$2\mid 1$	$3\mid 1$	$1\mid 2$	$2\mid 2$	$3\mid 2$
TrajeR	95.32792	4.67203	6e-05	0	94.67946	5.32054
Traj SAS	95.3	4.7	0	0	94.7	5.3

Prob. group out. 1$\mid$group out. 2	$1\mid 3$	$2\mid 3$	$3\mid 3$	$4\mid 1$	$4\mid 2$	$4\mid 3$
TrajeR	0	0.00539	99.9946	0	0	100
Traj SAS	0	0	100	0	0	100

Conditional probabilities, groups of outcome 2 given groups of outcome 1

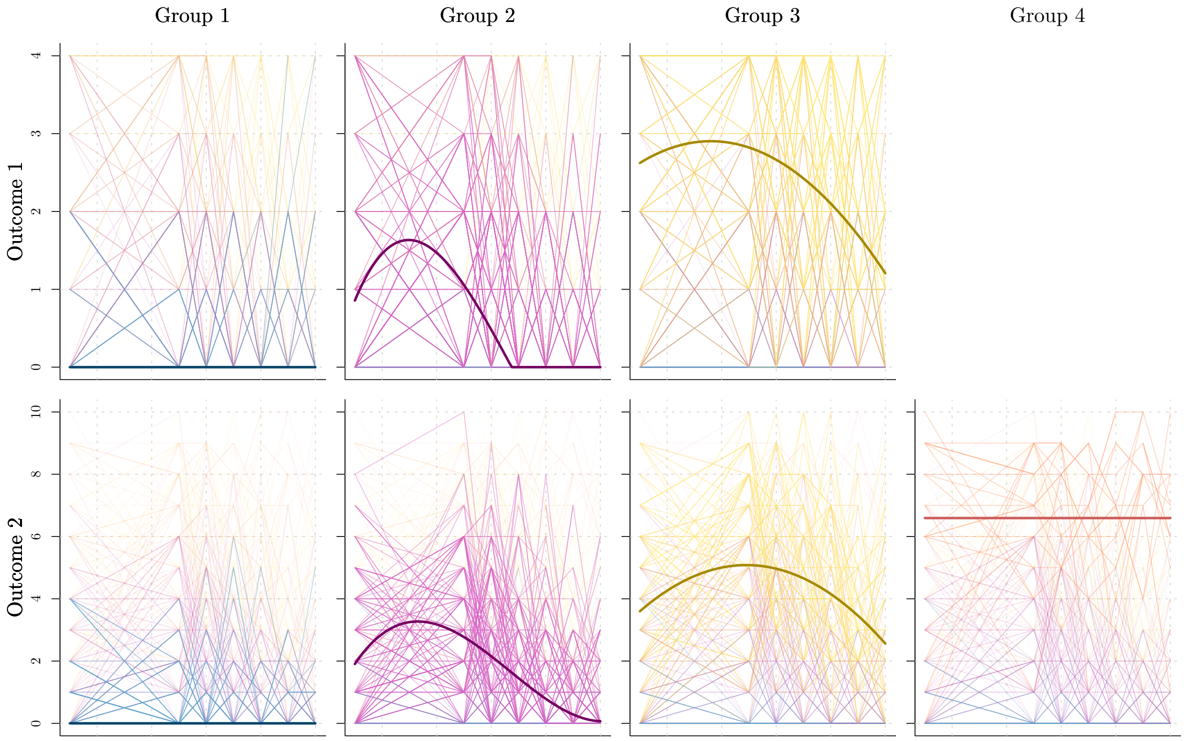
Prob. group out. 2$\mid$group out. 1	$1\mid 1$	$2\mid 1$	$3\mid 1$	$4\mid 1$	$1\mid 2$	$2\mid 2$
TrajeR	99.99999	1e-05	0	0	3.07887	96.91826
Traj SAS	99.99999	0	0.00001	0	3.07834	96.9216

Prob. group out. 2$\mid$group out. 1	$3\mid 2$	$4\mid 2$	$1\mid 3$	$2\mid 3$	$3\mid 3$	$4\mid 3$
TrajeR	0.00287	0	0	7.2486	70.92353	21.82783
Traj SAS	0.00005	0	0	7.24646	70.92763	21.82591

Joint probabilities

Prob. group out. 2 and group out. 1	$1;1$	$2;1$	$3;1$	$1;2$	$2;2$	$3;2$
TrajeR	26.40013	1.29387	2e-05	0	40.72919	2.28879
Traj SAS	26.4	1.3	0	0	40.7	2.3

Prob. group out. 1 and group out. 2	$1;3$	$2;3$	$3;3$	$4;1$	$4;2$	$4;3$
TrajeR	0	0.00121	22.39452	0	0	6.89226
Traj SAS	0	0	22.4	0	0	6.9


Group 1
Group 2
Group 3
Group 4
Outcome 1
Outcome 2
0
1
2
3
4
0
2
4
6
8
10

CHAPTER 8

trajeR: an R-package for finite mixture modeling

8.1 Introduction

Once the model was well-defined and the formulas correctly formulated, a significant aspect of this thesis involved the development of the **trajeR** R package. This package was constructed with the assistance of `Rcpp`, a package that enables the integration of `C++` code within R source code. You can learn more about `Rcpp` in references such as Eddelbuettel and François (2011) and Eddelbuettel (2013).

The **trajeR** package consists of nearly 9,000 lines of code, with approximately 76 % of it written in `C++` and 23.9 % in `R`. This package has been officially submitted and is available on CRAN (Comprehensive R Archive Network) at the following URL: `https://cran.r-project.org/web//packages/trajeR/index.html`. Additionally, the latest version of the package can be found on GitHub at: `https://github.com/gitedric/trajeR`.

8.2 Package design

The package **trajeR** is built around the core function **trajeR** which fits the model and computes its parameters for given degrees of the polynomial trajectories in the different groups. The function signature for **trajeR** is

```
trajeR(Y, A, Risk = NULL, TCOV = NULL, degre, degre.nu = 0,
       Model, Method = "L",
       ssigma = FALSE, ymax = max(Y) + 1, ymin = min(Y) - 1,
       hessian = TRUE, itermax = 100, paraminit = NULL,
       ProbIRLS = TRUE, refgr = 1,
       fct = NULL, diffct = NULL, nbvar = NULL, nls.lmiter = 50)
```

Some of these arguments are mandatory others optional.

The mandatory arguments are the main data matrices **Y** , **A** , as well as **degre** , **Model** and **Method** .

Here **Y** is the matrix containing the values of the variable of interest and **A** is the matrix containing the age or time variable. In most applications, this matrix just contains times of measurement that are the same for each individual in the sample, implying that all lines of the matrix

A are equal, but this is not necessarily the case. **A** can for instance contain the age of the different individuals at the times of measurement, which is generally different for each individual in the sample.

degre is a vector indicating the degree of the polynomials describing the typical trajectories in the different groups. Implicitly, the dimension of this vector also determines the number of groups into which we want to divide the population,

Model is a string defining the underlying distribution used in the model. The possible choices are LOGIT for the Logistic Regression Mixture Model, CNORM for the Censored Normal Mixture Model or ZIP for the Zero Inflated Poisson Mixture model.

Method , finally, is a string to decide which algorithm is used for estimating the model parameters. The possible choices are `L` for direct optimization, `EM` for the Expectation Maximization algorithm with quasi-Newton procedures (for LOGIT and ZIP models) and **EMIRLS** for the Expectation Maximization algorithm using Iterative Weighted Least Squares.

The optional arguments are **Risk** , **TCOV** , **degre.nu** , **ssigma** , **ymax** , **ymin** , **hessian** , **itermax** , **paraminit** , **ProbIRLS** , **refgr** , **fct** , **diffct** , **nls.lmiter** , **ng.nl** and **nbvar**

.

Risk is a data matrix that contains the values of the covariate X modifying the group membership probability. By default, there is no such variable and **Risk** is a one-column matrix with value 1.

ProbIRLS allows to decide which method is used to compute the predictor probabilities. If

its value is TRUE (default setting) we use the IRLS method and if it is FALSE we use the optimization method.
TCOV is an optional data matrix containing a time-dependent covariate W that influences the trajectories themselves. By default its value is NULL.
To ensure the identifiability of the parameters of the predictor, we have to fix a reference group. This can be done by the **refgr** command. It's default value is 1.
hessian indicates if we want to calculate the Hessian matrix, the default value being FALSE. If the method used is direct optimization, the Hessian matrix is computed by inverting the Fisher Information Matrix. If the method is EM or EMIRLS, the Hessian matrix is computed by using the Louis method Louis (1982).
itermax gives the maximal number of iterations for the **optim** function or for the EM algorithm.
The choice of the initial parameters is very important in optimization problems. We can specify these initial parameters by **paraminit** . By default **trajeR** calculates the initial value based on the range or the standard deviation of the data.

In case of a ZIP model, we have to specify the probability to belong to the excess zero state. This is done by using a polynomial logistic regression. **degre.nu** is the degree of the polynomial.

In case of a Censored Normal model, we have to define several arguments. **ssigma** indicates if we suppose to have the same standard deviation for the error terms in all groups. By default, its value is FALSE. **ymax** indicates the maximum of Y. It concerns only the model with censored data. By default its value is the maximum value of the data plus 1, i.e the model used is simply the normal model. Likewise, **ymin** indicates the minimum of Y.

There are also several arguments to define in order to use general nonlinear functions for the typical trajectories in the different groups. **fct** gives the definition of the function used to define the shape of the trajectories and its differential is defined in **diffct** . We also need to specify the number of groups to use **nl.ng** and the number of parameters to be estimated **nbvar** .

The output of **trajeR** is an object of class **Trajectory** that can be of four types depending on the model used: **Trajectory.LOGIT** , **Trajectory.CNORM** , **Trajectory.ZIP** or **Trajectory.NL** . These classes are described in chapter 3.

Some examples of the prompt for a given a distribution:

- Censored Normal:

```
# Likelihood different sigma
trajeR(Y = data[,2:11], A = data[,12:21],
       degre = c(0,3,4),
       Model = "CNORM", Method = "EM",
       hessian = TRUE, ssigma = FALSE)
```

- Logit:

```
# EM
trajeR(Y = data[,2:11], A = data[,12:21],
       degre = c(0,3,4),
       Model = "LOGIT", Method = "EM",
       hessian = TRUE)
```

- Zero Inflated Poisson:

```
# Likelihood
trajeR(Y = data[,2:6], A = data[,7:11],
       degre=c(2,2), degre.nu = c(1,1),
       Model="ZIP", Method = "L",
       hessian = TRUE)
```

- Beta:

```
# Likelihood
trajeR(Y = data[,2:6], A = data[,7:11],
       degre=c(2,3,1), degre.phi = c(1,1,1),
       Model="BETA", Method = "L",
       hessian = TRUE)
```

8.3 Plotting

trajeR provides a convenient way to visualize the output object, allowing you to create plots of the mean curves for each cluster, with or without the data points. The x-axis represents the time data, and you have the option to add the mean of the group for each time value to the plot.

The main function for creating these plots is:

```
plottrajeR(Object.trajeR, Y = NULL, A = NULL,
           mean = FALSE, ...)
```

8.4 Model selection criteria

The criteria usually used for model selection in finite mixture models are the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC). These 2 classical criteria are implemented in **`trajeR`** and can be accessed by the functions **`trajeRBIC`** and **`trajeRAIC`** respectively. For example, in a simulated example,

```
trajeRBIC(solLRisk)

## [1] 29255.68

trajeRAIC(solLRisk)

## [1] 29167.18
```

8.5 Model adequacy criteria

The posterior probabilities of group membership are among other a source of valuable information for judging the model's correspondence with the data Daniel S. Nagin (2005). We implemented the four diagnostics proposed in Daniel S. Nagin (2005) in **`trajeR`**.

8.5.1 Average Posterior Probability of Assignment

We call Average Posterior Probability of Assignment (AvePP) the average posterior probability of membership for each group for those individuals that were assigned to it. In a ideal situation the assignment probability for each individual would be 1 and the Average Posterior Probability (AvePP) would be 1 too. In our simulated example, we get

```
AvePP(solLRisk, Y = data[,2:11], A = data[,12:21], X = data[, 42:43])

## [1] 0.9957070 0.9920117 0.9999998
```

We can see that all values are very close to 1. It means that the model fits the data correctly, which is of course not astonishing since the data were simulated to fit the model.

8.5.2 Odds of Correct Classification

The Odds of Correct Classification for group k ($\mathrm{OCC_k}$) is the ratio between the odds of a correct classification into group k on the basis of the posterior probability rule and the odds of correct classification based on random assignment, with the probability of assignment to group k equal to its estimated population size $\hat{\pi}_k$. Hence,

$$OCC_k = \frac{AvePP_k/(1 - AvePP_k)}{\hat{\pi}_k/(1 - \hat{\pi}_k}).$$

Large values of OCC_k indicate a good assignment accuracy. Nagin suggests that in a real world application an OCC_k greater than 5 for all groups is indicative that the model has a high assigment accuracy Daniel S. Nagin (2005).

```
OCC(solL, Y = data[,2:11], A = data[,12:21])

## [1] 2.769705e+02 2.269346e+02 2.524712e+07
```

8.5.3 Estimated Group Probabilities versus the Proportion of the Sample Assigned to the Group

We compute the probability of group membership by two methods : using $\hat{\pi}_k$ or using the proportion P_k of the sample assigned to the group k. Ideally the these two values should be equal. The function **`propAssign`** computes P_k:

```
propAssign(solL, Y =data[,2:11], A = data[,12:21])

## gr
##     1     2     3
## 0.458 0.350 0.192
```

We can compare with $\hat{\pi}_k$ calculated by the model.

```
exp(solL$theta)/sum(exp(solL$theta))

## [1] 0.4589082 0.3489514 0.1921404
```

8.5.4 Confidence Intervals for Group Membership Probabilities

A narrow confidence interval of $\hat{\pi}_k$, for a given value of α implies that the probability is accurately estimated. These intervals are calculated by means of the bootstrap method.

```
ConfIntT(solL, Y = data[,2:11], A = data[,12:21],
         nb = 10000, alpha = 0.98)

##           [,1]      [,2]      [,3]
## 1%  0.4133988 0.3074413 0.1614733
## 99% 0.5045751 0.3928050 0.2269818
```

8.5.5 Adequacy Matrix

trajeR allows to summarize these diagnostics in one table.

```
adequacy(solL,  Y = data[,2:11], A = data[,12:21],
         nb = 10000, alpha = 0.98)

##                      1           2            3
## Prob. est.   0.4589082   0.3489514 1.921404e-01
## CI inf.      0.4137369   0.3074233 1.623457e-01
## CI sup.      0.5054918   0.3928927 2.251117e-01
## Prop.        0.4580000   0.3500000 1.920000e-01
## AvePP        0.9957610   0.9918456 9.999998e-01
## OCC        276.9705045 226.9346306 2.524712e+07
```

CHAPTER 9

Application examples

9.1 ZIP model

For our model's study, we utilized the Toronto Juvenile Offender Samples, which have been previously examined by Day, Jason D Nielsen, et al. (2011). The dataset can be accessed through the `crimCV` package or by contacting I. Bevc. This dataset tracked the criminal activity of 378 youth annually, spanning from ages 9 to 38, by recording their offenses. The data sources for this information included:

- The (Ontario) Ministry of Community and Social Services (MCSS).
- The (Ontario) Ministry of Community Safety and Correctional Services (MCSCS).
- The Canadian Police Information Centre (CPIC).
- The Predisposition Reports (PDRs) maintained by the children's mental health center.

The initial dataset provided the dates of court contacts but lacked the dates of offenses. To rectify this, we used the "divide and round" (DAR) approximation method proposed by A. K. Ward et al. (2010). This method involves dividing each count by the corresponding exposure time and rounding the result to the nearest integer. For example, an individual who had been free for 8 months and committed 10 offenses in a year would have their count adjusted to $10/0.66 = 6.6$, which is rounded to 7.

Given that a substantial portion of the dataset comprises zero counts, the ZIP model was considered an appropriate choice to account for this excess of zeros in the data.

```
> data(TO1adj, package = "crimCV")
> str(TO1adj)

##  num [1:378, 1:31] 0 0 0 0 0 0 0 0 0 0 ...
##   - attr(*, "dimnames")=List of 2
##    ..$ : NULL
##    ..$ : chr [1:31] "Offense8" "Offense9" "Offense10" "Offense11" ...
```

As the dataset has been the subject of prior studies, such as J. D. Nielsen et al. (2014), we decided to apply our model to this data. To account for the unique characteristics of the data, we chose a 4th-degree polynomial for both the Poisson state and the excess zero state. We experimented with various numbers of groups, ranging from 2 to 8. Unfortunately, models with 5 or more groups yielded cases where one or more groups had a size of 0, rendering them impractical.
Ultimately, we retained the models with 2, 3, or 4 groups and selected the most suitable one using the AIC and BIC criteria. The final choice was a model with 4 groups. Since the dataset does not include a time variable, we had to create one for our analysis. For our parameter estimation, we opted for the Likelihood method instead of the EM (Expectation-Maximization) algorithm, as it often yields faster convergence in the case of ZIP models.

```
> library(trajeR)
> A <- matrix(rep(8:38, nrow(TO1adj)), nrow = nrow(TO1adj),
+              byrow = TRUE)
> sol <- trajeR(
+          Y = TO1adj, A = A,
+          Model = "ZIP", Method = "L",
+          degre = c(4, 4, 4, 4), degre.nu = c(4, 4, 4, 4),
+          hessian = TRUE
+          )
> sol
```

```
> sol

## Call TrajeR with 4 groups and a 4,4,4,4 degrees of polynomial shape
## of trajectory.
## Model : Zero Inflated Poisson
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                param.=0
## ------------------------------------------------------------------
##       1   Intercept   -111.94768      10.6419   -10.51952          0
##              Linear   19.42967       2.0624      9.4209          0
##           Quadratic   -1.22354      0.14653    -8.35034          0
##               Cubic     0.0333       0.00451     7.38122          0
##             Quartic   -0.00033        5e-05    -6.54677          0
##                Nu11   -87.48518     17.79949    -4.91504          0
##                Nu12    13.4706      3.14214     4.28708      2e-05
##                Nu13   -0.77139      0.20784    -3.71148    0.00021
##                Nu14    0.01944      0.00606     3.20585    0.00135
##                Nu15   -0.00018        7e-05    -2.74479    0.00606
##
##       2   Intercept   -54.23659      5.48988    -9.87938          0
##              Linear    9.74532      0.99082     9.83558          0
##           Quadratic   -0.62546      0.06554    -9.54304          0
##               Cubic    0.01737      0.00188     9.22883          0
##             Quartic   -0.00018        2e-05    -8.91589          0
##                Nu21   24.23192     10.59853     2.28635    0.02225
##                Nu22   -4.12558      1.90011    -2.17123    0.02993
##                Nu23    0.24289       0.1238     1.96201    0.04978
##                Nu24    -0.0062      0.00348    -1.78343    0.07454
##                Nu25      6e-05        4e-05     1.69897    0.08935
##
##       3   Intercept   -19.94071      4.34736    -4.58685          0
##              Linear    3.06348      0.75081     4.08022      5e-05
##           Quadratic   -0.15403      0.04738    -3.25115    0.00115
##               Cubic    0.00333      0.00129     2.58024    0.00989
##             Quartic     -3e-05        1e-05    -2.04205    0.04117
##                Nu31   40.54459     10.63227     3.81335    0.00014
##                Nu32   -6.66735      1.93069    -3.45334    0.00056
##                Nu33    0.37821      0.12583      3.0057    0.00266
##                Nu34   -0.00919      0.00351    -2.61988    0.00881
##                Nu35      8e-05        4e-05     2.35308    0.01864
##
##       4   Intercept     4.2123     10.30998     0.40857    0.68287
##              Linear   -1.73859      2.04496    -0.85018    0.39524
##           Quadratic    0.17462       0.1502     1.16256    0.24503
##               Cubic   -0.00559      0.00484    -1.15481    0.24819
##             Quartic      5e-05        6e-05     0.85363    0.39333
##                Nu41   22.86393     21.34394     1.07121     0.2841
##                Nu42   -2.41689      4.55809    -0.53024    0.59595
##                Nu43   -0.00177      0.35593    -0.00497    0.99604
##                Nu44    0.00574      0.01204     0.47677    0.63353
##                Nu45   -0.00013      0.00015    -0.85799    0.39092
## --------------------------------------------------------------------
##       1         pi1     0.5773      0.02951           0          0
##       2         pi2    0.27582       0.0272   -27.15925          0
##       3         pi3    0.07333      0.01179   -175.0553          0
##       4         pi4    0.07355      0.01376   -149.71431          0
## --------------------------------------------------------------------
## Likelihood : -11041.57
```

The estimated parameter values are presented in the "Estimate" column, and their standard deviations are reported in the "Std. Error" column. In the following two columns, statistical tests were conducted to determine whether the parameters significantly differed from 0. These tests indicated that nearly all parameters were statistically different from 0, with the exception of the parameters associated with the first group in the logistic part. This suggests that polynomial degrees three and four may not be essential for the model.

To assess the model's goodness of fit, several methods were employed. Classic criteria such as BIC Schwarz (1978) and AIC Akaike (1974) were used to aid in model selection.

```
> trajeRAIC(sol)

## [1] 22169.14

> trajeRBIC(sol)

## [1] 22338.34
```

We can use the methods described in the chapter 6 to test the adequacy of the model.

Diagnostic 1: Average Posterior Probability of Assignment (AvePP)

Ideally, this indicator must be near to 1. All these values are greater than 0.76, and the groups 1, 3, and 4 achieve a good assignment of the individuals inside them. One open question is, does there exist some threshold for AvePP to consider good assignment? Nagin suggests that the minimum rule-of-thumb is that AvePP should be at least 0.7 for all groups.

```
> AvePP(sol, Y = TO1adj, A = A)

## [1] 0.9692146 0.9461970 0.9993172 0.9861020
```

Diagnostic 2: Odds of correct Classification (OCC)

OCC measures the odds of correct classification for a given group. If the maximum probability assignment rule has no predictive capacity beyond random chance, OCC equals 1. Here, all the values are greater than 0.8, indicating that the predictive capacity of the probability assignment rule is strong.

```
> OCC(sol, Y = TO1adj, A = A)

## [1]    23.05207    46.17459 18493.21392   893.73269
```

Diagnostic 3: Estimated Group Probabilities versus the Proportion of the Sample Assigned to the Group

We compare the estimated probability with the proportion assigned to a given group. Ideally, these two values should be equal.

```
> propAssign(sol, Y = TO1adj, A = A)

##              1         2          3          4
## [1,] 0.5820106 0.2751323 0.07142857 0.07142857
```

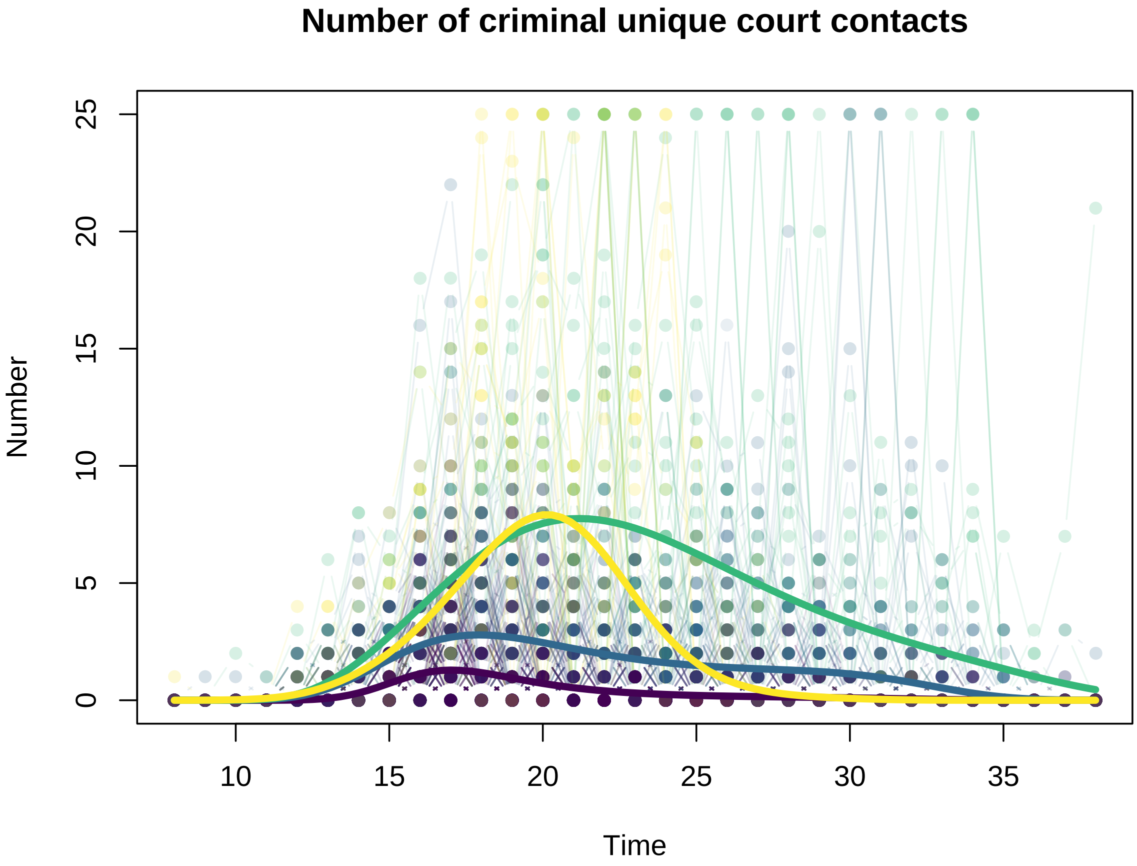
Diagnostic 4: Confidence intervals for group membership probabilities

The narrower the interval, the more accurately the probability of group membership is estimated. It is important to note that there is no specific criterion to determine whether an interval is sufficiently narrow. To assess the accuracy of the model, we present all the diagnostics in one table.

```
> round(adequacy(sol, Y = TO1adj, A = A, nb = 10000, alpha = 0.98), 5)

##                   1        2           3         4
## Prob. est.  0.57730  0.27582     0.07333   0.07355
## CI inf.     0.55752  0.25950     0.06968   0.06979
## CI sup.     0.59625  0.29271     0.07709   0.07746
## Prop.       0.58201  0.27513     0.07143   0.07143
## AvePP       0.96921  0.94620     0.99932   0.98610
## OCC        23.05207 46.17459 18493.21392 893.73269
```

Given that the model with 4 groups and 4-degree polynomials is accurate, we can proceed with plotting it.

```
> library(viridis)

> col_line <- viridis(4)
> col_bkg <- adjustcolor(col_line, alpha.f = 0.1)
> plotrajeR(sol, Y = TO1adj, A = A,
+           col = c(col_bkg, col_line),
+           main = "Number of criminal unique court contacts",
+           xlab = "Time", ylab = "Number")
```



9.2 CNORM model

We utilized a sample dataset from the Montreal Longitudinal and Experimental Study, as outlined in Richard E. Tremblay et al. (2003). This longitudinal study, initiated in 1984 in the Montreal area of Quebec, Canada, primarily aimed to examine the development of antisocial behavior from kindergarten to high school, with a specific focus on parent-child interactions. For our analysis, we focused on a subset of 138 subjects.

The dataset includes measurements of opposition behavior on a scale ranging from 0 to 10. These measurements were collected annually at ages 6 and 10-15 and encompassed five items: 'does not share,' 'irritable,' 'disobedient,' 'blames others,' and 'inconsiderate.' Concurrently, we collected two covariates related to the amount of schooling completed by the mother (SCOLMER) and father (SCOLPER).

The dataset can be accessed at `https://www.andrew.cmu.edu/user/bjones/sas/cnorm.sas`.

```
> str(data)

## 'data.frame': 138 obs. of  16 variables:
##  $ O1     : int  2 2 4 4 5 8 1 1 3 4 ...
##  $ O2     : int  0 1 2 0 9 10 2 0 1 2 ...
##  $ O3     : int  1 2 0 1 3 2 0 1 0 2 ...
##  $ O4     : int  0 0 0 0 4 0 0 0 0 3 ...
##  $ O5     : int  0 5 0 NA 4 1 0 0 0 6 ...
##  $ O6     : int  0 1 0 NA 3 2 0 0 0 NA ...
##  $ O7     : int  0 0 1 0 3 2 0 0 0 6 ...
##  $ T1     : num  -0.6 -0.6 -0.6 -0.6 -0.6 -0.6 -0.6 -0.6 -0.6 -0.6 ...
##  $ T2     : num  -0.2 -0.2 -0.2 -0.2 -0.2 -0.2 -0.2 -0.2 -0.2 -0.2 ...
##  $ T3     : num  -0.1 -0.1 -0.1 -0.1 -0.1 -0.1 -0.1 -0.1 -0.1 -0.1 ...
##  $ T4     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ T5     : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ T6     : num  0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 ...
##  $ T7     : num  0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 0.3 ...
##  $ SCOLMER: int  10 9 14 12 9 7 12 16 9 9 ...
##  $ SCOLPER: int  15 6 12 13 9 16 7 20 20 12 ...
```

Following the work of Jones, we have chosen to fit the model with three groups, each characterized by cubic trajectories and having the same standard deviation within each group. The dependent variable represents scores ranging from 0 to 10. Therefore, the censored normal distribution appears to be a sensible choice for modeling this type of data.

```
> sol <- trajeR(Y = data[, 1:7], A = data[, 8:14],
+     degre = c(3, 3, 3),
+     Model = "CNORM", Method = "L",
+     ssigma = TRUE, hessian = TRUE,
+     ymin = 0,
+     ymax = 10
+ )
```

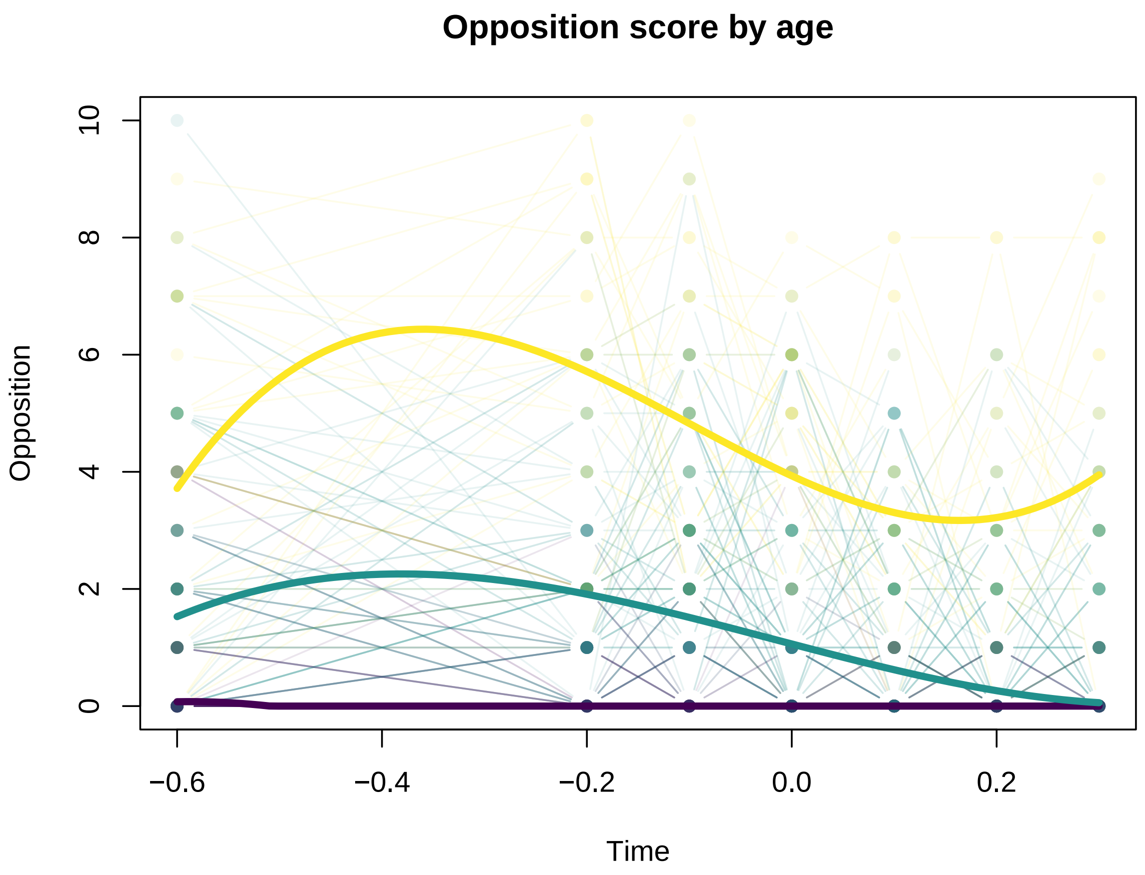
```
> sol

## Call TrajeR with 3 groups and a 3,3,3 degrees of polynomial shape
## of trajectory.
## Model : Censored Normal
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                param.=0
## -------------------------------------------------------------------
##       1   Intercept   -2.34029      0.59487    -3.93413      9e-05
##              Linear   -4.54303      2.97544    -1.52684    0.12713
##           Quadratic     5.5429     11.98453      0.4625    0.64382
##               Cubic   10.68041     24.10402      0.4431     0.6578
##
##       2   Intercept    1.06069       0.3167      3.3492    0.00084
##              Linear   -4.58363      1.41108    -3.24831     0.0012
##           Quadratic    0.64225      4.71449     0.13623    0.89167
##               Cubic   11.64248     10.20263     1.14113     0.2541
##
##       3   Intercept    3.93136      0.37738    10.41752          0
##              Linear   -8.06229      2.16777    -3.71916    0.00021
##           Quadratic   13.36513      6.82152     1.95926    0.05037
##               Cubic    45.6647     15.26152     2.99215    0.00284
## -------------------------------------------------------------------
##       1      sigma1    2.64271      0.14738    17.93163          0
##       2      sigma2    2.64271      0.14738    17.93163          0
##       3      sigma3    2.64271      0.14738    17.93163          0
## -------------------------------------------------------------------
##       1         pi1    0.26085      0.07987           0    0.00113
##       2         pi2    0.54254      0.06186    11.83801          0
##       3         pi3    0.19661      0.05052     -5.5971    0.00011
## -------------------------------------------------------------------
## Likelihood : -1593.537
```

To test the adequacy of the model, we use the function "adequacy" to directly display the table of all diagnostics.

```
> adequacy(sol,  Y = data[, 1:7], A = data[, 8:14])

##                     1         2          3
## Prob. est.  0.2608541 0.5425382  0.1966077
## CI inf.     0.2218094 0.4965786  0.1713491
## CI sup.     0.3041598 0.5875309  0.2233164
## Prop.       0.2608696 0.5579710  0.1811594
## AvePP        0.8702552 0.8790312  0.8994214
## OCC         19.0059288 6.1271043 36.5413764
```

The diagnostics suggest that the model has a very good capacity to accurately estimate group membership probabilities. The AvePP is greater than 0.87 for all groups, surpassing the 0.7 threshold (diagnostic 1). The OCC is well above 5 for all groups (diagnostic 2). There is a close correspondence between the estimated probabilities and the proportion assigned to the group based on the maximum posterior argument probability rule for each group (diagnostic 3). The 98 % confidence intervals are relatively narrow for each group, with widths less than 0.092 plus or minus the estimated probabilities (diagnostic 4). Then we can create the graphical representation.

```
> library(viridis)

> col_line <- viridis(3)
> col_bkg <- adjustcolor(col_line, alpha.f = 0.1)
> plotrajeR(sol, Y = data[, 1:7], A = data[, 8:14],
+      col = c(col_bkg, col_line),
+      xlab = "Time", ylab = "Opposition", main = "Opposition score by age")
```



For each parameter, we test its significance with a t-test. We observe that the p-values for the quadratic and cubic parameters of groups 1 and 2 are very high (greater than 0.1), indicating that they do not significantly differ from 0. Therefore, we can remove them in the next model.
We also investigate whether the three groups are influenced by certain covariates, particularly the amount of schooling completed by the mother and father. We introduce the covariates SCOLMER and SCOLPER in the model with the "Risk" option and check if the corresponding parameters are significant. To ensure a quick convergence of the algorithm, we start it from the results of the first model using **`paraminit`** .

```
> solRisk <- trajeR(Y = data[, 1:7], A = data[, 8:14],
+      Risk = data[, 15:16],
+      degre = c(1, 1, 3),
+      Model = "CNORM", Method = "L",
+      ssigma = TRUE, hessian = TRUE,
+      ymin = 0,
+      ymax = 10,
+      paraminit = c(
+          sol$theta[1], 0, 0,
+          sol$theta[2], 0, 0,
+          sol$theta[3], 0, 0,
+          sol$beta[1:2],
+          sol$beta[5:6],
+          sol$beta[9:12],
+          sol$sigma
+      )
+ )
```

```
> solRisk

## Call TrajeR with 3 groups and a 1,1,3 degrees of polynomial shape
## of trajectory.
## Model : Censored Normal
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                param.=0
## --------------------------------------------------------------------
##       1   Intercept   -2.19652      0.41702    -5.26717          0
##              Linear   -3.96666      0.92362    -4.29471      2e-05
```

```
##
##       2   Intercept    0.93213      0.25762     3.61827     0.00031
##              Linear   -1.73025      0.54896    -3.15184     0.00167
##
##       3   Intercept    3.88433      0.38775    10.01757          0
##              Linear   -8.79596      2.30474    -3.81646     0.00014
##           Quadratic   14.18118      7.40042     1.91627     0.05563
##               Cubic   49.61899     16.65507     2.97921     0.00296
## --------------------------------------------------------------------
##       1      sigma1    2.65235      0.14704    18.03802          0
##       2      sigma2    2.65235      0.14704    18.03802          0
##       3      sigma3    2.65235      0.14704    18.03802          0
## --------------------------------------------------------------------
##       1    Baseline          0           NA          NA         NA
##
##       2   Intercept    3.65064      1.20087        3.04    0.00243
##             SCOLMER   -0.05873      0.10072    -0.58305       0.56
##             SCOLPER   -0.20242      0.09283    -2.18058    0.02946
##
##       3   Intercept    3.82126      1.33147     2.86995     0.0042
##             SCOLMER    -0.1606      0.12165    -1.32016    0.18709
##             SCOLPER   -0.21765      0.10344    -2.10399    0.03564
## --------------------------------------------------------------------
## Likelihood : -1590.902
```

The covariate SCOLMER does not have any significant effect on the trajectory of the opposition score (p = 0.56 for group 1 and p = 0.187 for group 2). However, the covariate SCOLPER does have a predictive effect (p = 0.029 for group 1 and p = 0.036 for group 2). This indicates that the amount of schooling completed by the mother does not significantly influence the trajectory, while the amount of schooling completed by the father has a protective effect. The negative parameter estimates (-2.18 and -2.1) suggest that a higher level of schooling for the father is associated with a decrease or lower level of the opposition score.

9.3 LOGIT model

The Cambridge Study in Delinquent Development (Farrington and West (1990)) is a prospective longitudinal study of 411 London males. The study's primary objective is to investigate the development of offending and antisocial behavior from childhood to adulthood. The participants, all males, were interviewed at various time points from the age of 8 up to age 48, and their criminal records were also examined. In your sample, you are focusing on tracking convictions for each year from age 8 to 32.

```
> str(as.matrix(data))

##  int [1:411, 1:880] 8488 8488 8488 8488 8488 8488 8488 8488 8488 8488 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:880] "v1" "v2" "v3" "v4" ...

> library(kableExtra)
> library(dplyr)
```

For a comprehensive description of each variable, you can refer to the source cited, Farrington and West (1990).
In this example, we primarily use variables V13 to V26, which count the number of convictions for each year. We construct a longitudinal variable with a value of 1 if there is at least one conviction for a given year, and then fit the matrix of time variables to arbitrary values for your analysis.

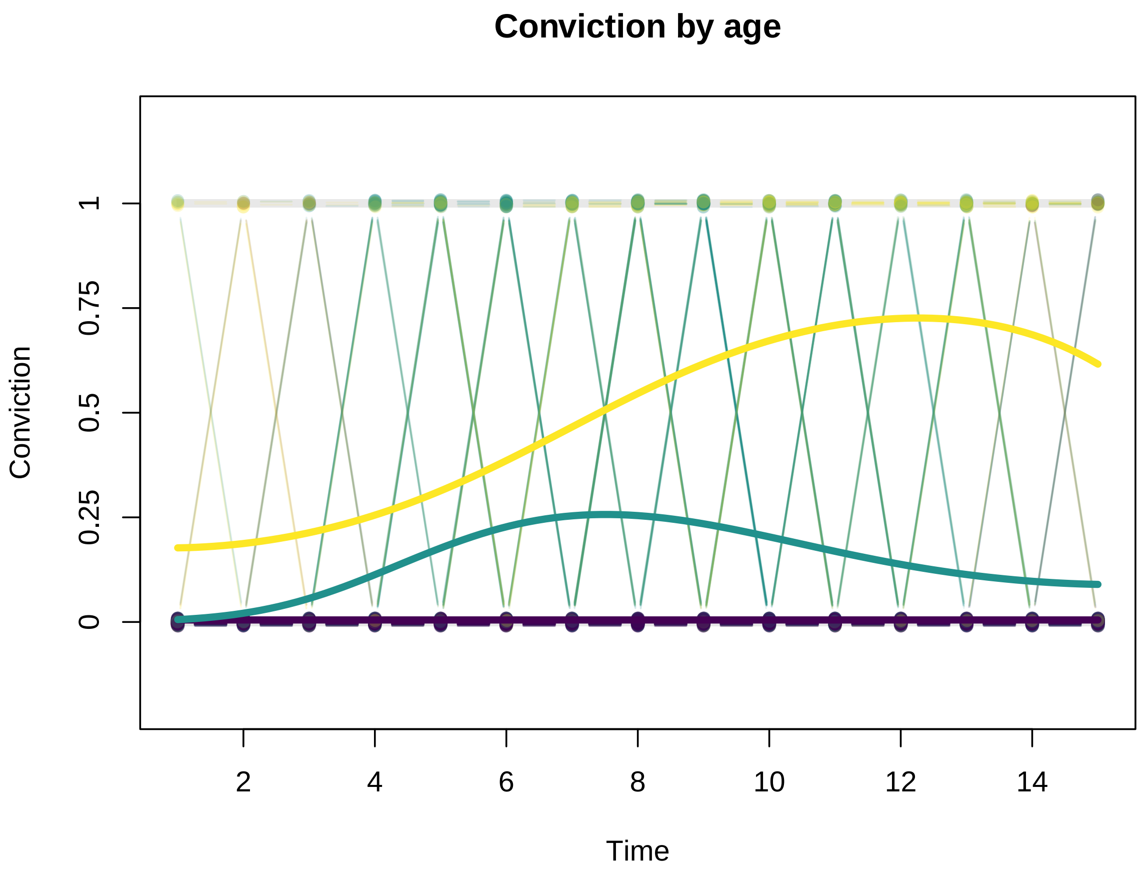
```
> Y <- data[, 13: 27]
> Y[Y >= 1] <- 1
> A <- matrix(rep(1:15, nrow(Y)), byrow = TRUE, ncol = ncol(Y))
```

We decided to fit the model with polynomials of degree 0, 3, and 3. For a more detailed discussion on how we arrived at these choices, please refer to chapter 3.

```
> sol <- trajeR(
+     Y = Y, A = A,
+     Model = "LOGIT", Method = "L",
+     degre = c(0, 3, 3),
+     hessian = TRUE
+ )
```

```
> sol

## Call TrajeR with 3 groups and a 0,3,3 degrees of polynomial shape
## of trajectory.
## Model : Logit
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                param.=0
## --------------------------------------------------------------------
##        1    Intercept   -5.31127      0.50068   -10.60815          0
##
##        2    Intercept   -6.83062        1.033    -6.61243          0
##                 Linear    1.83635      0.41019     4.47681      1e-05
##              Quadratic   -0.18223      0.04988    -3.65319    0.00026
##                  Cubic    0.00532      0.00188     2.83858    0.00455
##
##        3    Intercept   -1.49242      0.67662    -2.20569    0.02744
##                 Linear   -0.10687      0.33412    -0.31984     0.7491
##              Quadratic    0.06664      0.04746     1.40417    0.16032
##                  Cubic   -0.00339      0.00195     -1.7381    0.08224
## --------------------------------------------------------------------
##        1          pi1    0.66782      0.03873           0          0
##        2          pi2    0.26737      0.03727   -24.56397          0
##        3          pi3    0.06481      0.01377   -169.34427          0
## --------------------------------------------------------------------
## Likelihood : -1239.107
```

```
> library(viridis)

> col_line <- viridis(3)
> col_bkg <- adjustcolor(col_line, alpha.f = 0.1)
> plotrajeR(sol, Y = Y, A = A,
+         col = c(col_bkg, col_line),
+         xlab = "Time", ylab = "Conviction", main = "Conviction by age",
+         yaxt="n")
```



To fit the model, the ZIP, CNORM, and LOGIT methods can be computed using either the Likelihood or EM algorithm. In the case of ZIP and LOGIT, we also have an IRLS version of EM, which can be useful in cases where the Likelihood method fails to converge due to numerical issues, especially when dealing with sharp underlying functions.
Therefore, we fit the example above using the EM and EMIRLS methods for comparison with the Likelihood method.

```
> solEM <- trajeR(
+     Y = Y, A = A,
+     Model = "LOGIT", Method = "EM",
+     degre = c(0, 3, 3),
+     hessian = TRUE
+ )
>
> solEMIRLS <- trajeR(
+     Y = Y, A = A,
+     Model = "LOGIT", Method = "EMIRLS",
+     degre = c(0, 3, 3),
+     hessian = TRUE
+ )
```

SolL		SolEM		SolEMIRLS	
parameters	sd	parameters	sd	parameters	sd
Beta 1					
-5.31127	0.50068	-5.31776	0.46898	-5.31149	0.46612
Beta 2					
-6.83062	1.03300	-6.82778	1.05591	-6.83058	1.05581
1.83635	0.41019	1.83387	0.41861	1.83631	0.41860
-0.18223	0.04988	-0.18192	0.05108	-0.18223	0.05107
0.00532	0.00188	0.00531	0.00193	0.00532	0.00193
Beta 3					
-1.49242	0.67662	-1.52537	0.69858	-1.49245	0.69666
-0.10687	0.33412	-0.09055	0.34427	-0.10685	0.34398
0.06664	0.04746	0.06443	0.04891	0.06664	0.04891
-0.00339	0.00195	-0.00330	0.00200	-0.00339	0.00200
Pi					
0.66782	0.03873	0.66731	0.03044	0.66781	0.03043
0.26737	0.03727	0.26775	0.02874	0.26738	0.02872
0.06481	0.01377	0.06493	0.04186	0.06481	0.04184

9.3.0.1 Time longitudinal covariable

Some levels of reading skill are present in the dataset for specific ages: from 10 to 15, 18 to 19, 21 to 22, and 24. We recoded these levels into three categories: good aptitude (-1), normal aptitude (0), and bad aptitude (1). There were some missing values, which we chose to address through mean value imputation. In this method, missing values were replaced with the mean of the two adjacent values, creating a smooth reading index. For example, if a person had a reading level of 0 at age 19 and -1 at age 21, the missing 20-year reading level was imputed as 0.5, indicating a transition from a normal level of reading to a good level with a smooth transition.

The data are stored in the `TCOV` variable in this package.

The relationship between delinquency and learning disabilities is often studied by various authors. Brier (1989) attempted to explain this linkage through three hypotheses: susceptibility hypothesis, school failure hypothesis, and differential treatment hypothesis. In this context, we investigated whether our longitudinal reading ability variable could influence the trajectory of a given group.

We incorporated the TCOV variable into our model, using the parameters found in the simple case as initial parameters to ensure convergence.

```
> solTCOV <- trajeR(
+     Y = Y, A = A, TCOV = TCOV,
+     Model = "LOGIT", Method = "L",
+     degre = c(0, 3, 3),
+     paraminit = c(sol$theta[-1], sol$beta, 0, 0, 0),
+     hessian = TRUE
+ )
```

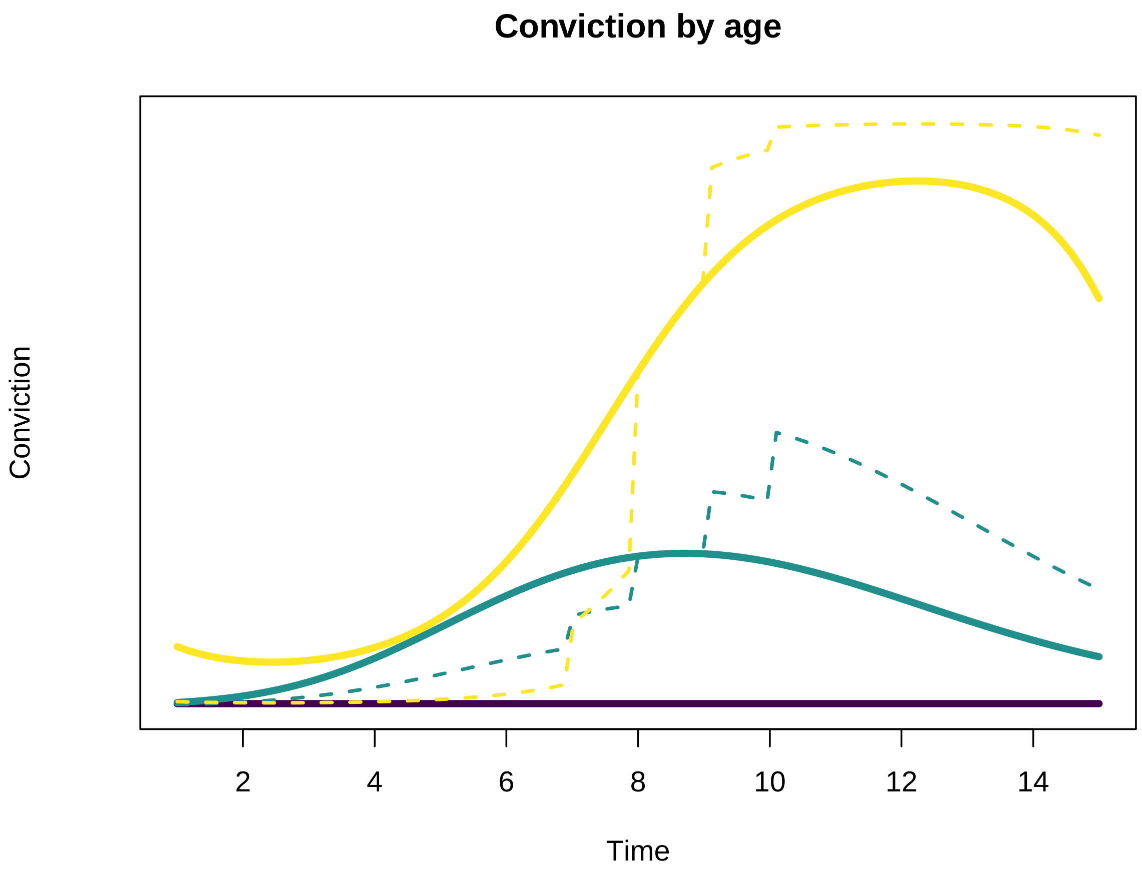
```
> solTCOV

## Call TrajeR with 3 groups and a 0,3,3 degrees of polynomial shape
## of trajectory.
## Model : Logit
## Method : Likelihood
##
##   group   Parameter   Estimate   Std. Error   T for H0:   Prob>|T|
##                                                 param.=0
## --------------------------------------------------------------------
##       1   Intercept   -5.64831      0.94758    -5.96075          0
##               TCOV1   -0.63241      0.96278    -0.65686     0.5113
##
##       2   Intercept   -6.56218      1.13247    -5.79456          0
##              Linear    1.46344      0.44117      3.3172    0.00091
##           Quadratic   -0.11782       0.0548    -2.14999    0.03159
##               Cubic    0.00259      0.00209     1.23648    0.21633
##               TCOV1    0.50063        0.122      4.1036      4e-05
##
##       3   Intercept   -1.50631      0.76965    -1.95714    0.05038
##              Linear   -0.89117      0.49181    -1.81203    0.07003
##           Quadratic    0.21889      0.08095     2.70412    0.00687
##               Cubic   -0.00994      0.00335    -2.96621    0.00303
##               TCOV1    1.38989      0.42495     3.27073    0.00108
## --------------------------------------------------------------------
##       1         pi1    0.65368      0.04456           0          0
##       2         pi2    0.28423      0.04115   -20.23849          0
##       3         pi3    0.06208      0.01469   -160.25724      2e-05
## --------------------------------------------------------------------
## Likelihood : -1215.196
```

For the first group, we observed that the time longitudinal covariate did not significantly differ from 0, indicating that it had no effect on the trajectories within this group. In contrast, for the two other groups, this covariate had a positive effect.

To further study the effect of a specific `TCOV` individual on the trajectory within their group, we examined an example where an individual had a good reading level from ages 8 to 12, which then progressively decreased to a bad level from ages 13 to 17. This level remained bad until age 32. The values of this time longitudinal covariate were as follows: -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.25, 0, 0.25, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5.

We plotted the trajectories and the impact of this particular longitudinal vector. The dashed lines represented the impact of this vector on the minimum trajectory.

```
> col_line <- viridis(3)
> col_bkg <- adjustcolor(col_line, alpha.f = 0.1)
> plotrajeR(solTCOV, col = c(col_bkg, col_line),
+     plotcov = c(rep(-0.5, 5), seq(from = -0.5, to = 0.5, length.out = 5),
+                 rep(0.5, 5)),
+     ylim = c(0, 1),
+     xlab = "Time", ylab = "Conviction", main = "Conviction by age",
+     yaxt="n")
```



9.4 BETA model

9.4.1 Data

We use daily data from 191 countries obtained from "Our World In data" Hasell et al. (2020). This data set is built from information supplied by the European Centre for Disease Prevention and Control, government reports, Oxford COVID-19 Government Response Tracker, World Bank, United Nations Statistics Division and Eurostat. Our main variable of interest is the rate of contamination with the COVID-19 virus. Moreover, we consider the covariates new cases, population size (in million inhabitants), total cases per million people, median age of the population, population density, number of inhabitants over 65 (in million inhabitants), government response stringency index, GDP per capita, extreme poverty index, cardiovascular death rate, diabetes prevalence rate, index of handwashing facilities, rate of hospital beds per thousand inhabitants, life expectancy and index of human development.

Finally, we use a stringency index, a composite measure of the nine government actions school closures, workplace closures, cancelation of public events, restrictions on public gatherings, closures of public transport, stay-at-home requirements, public information cam-

paigns, restrictions on internal movements and international travel controls.
First we need to do some data cleaning, replacing for instance some missing values for the median age, population density, population over 65, GDP per capita and life expectancy variables.

Alvarez et al. (2020), who use the same dataset, have created time series by considering a period of 100 days after the confirmation of the tenth COVID-19 case. We choose another approach. We transform all variables into monthly data and obtain thus a 16-period panel, starting in January 2020 and finishing in April 2021. This allows to smooth out periodicity problems that occur with daily data, for instance because on weekends some countries do not publish any COVID-19 related data. Moreover, for a lot of smaller countries, it is very difficult to determine the exact beginning of the epidemic. Since we interpret missing data as missing at random, and not as 0, this approach is more convenient anyway.

Using the notations of the chapter 3, we hence observe the rate of contamination Y_{it} for country i at time t, for $n = 190$ countires at $T = 16$ time points. $Y_i = (Y_{i1}, \cdots, Y_{i16})$ gives the time series for country i. Figure 9.1 shows the trajectories of the contamination rate for all 190 countries in our sample.

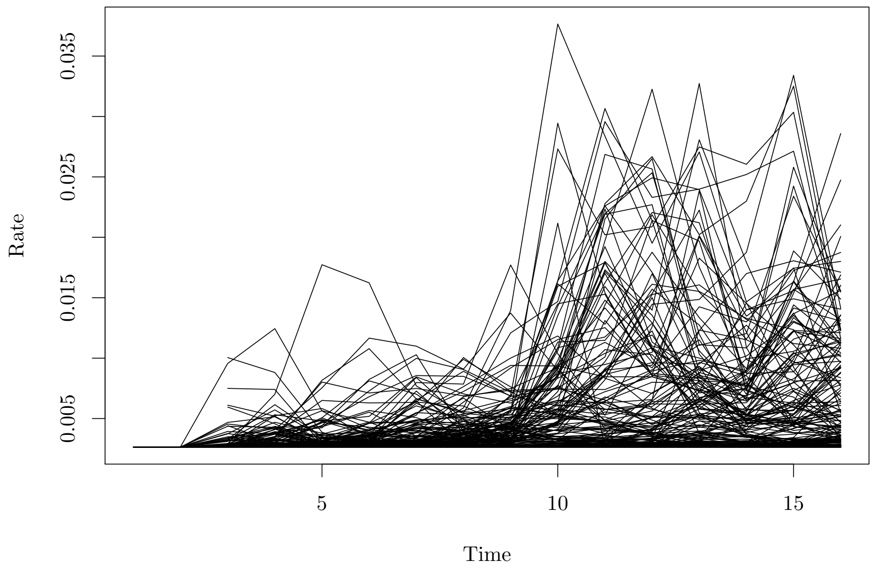


Figure 9.1: Contamination rates for all countries.

9.4.2 Model selection

To decide on the number of groups in our model, we use the BIC and AIC criteria together with the metric developed by Kass and Wasserman (1995). Indeed, let p_k be the probability that a model with k groups is the correct model. They show that p_k can be approximated by

$$p_k \approx \frac{e^{BIC_k - BIC_{max}}}{\sum_k e^{BIC_k - BIC_{max}}}.$$

To start with, we decide to model the typical trajectories with polynomials of degree 3 to be able to take into account two inflection points and to use polynomials of degree 2 to model the time dependency of the precision parameters zeta of the underlying Beta distribution.

Optimization methods can heavily depend upon the choice of the starting values of the different parameters. In case of a bad choice, the algorithm may not converge at all or converge to a local extremum. Unfortunately there is no established method to choose these starting points. We chose a three-step approach to solve this issue. In a first step, we chose k horizontal lines as starting values, i.e. take the parameters $\beta_k = (\beta_0, 0, 0, \dots)$ and $\zeta_k = (\zeta_0, 0, 0, \dots)$ for $k = 2, ..., 10$. Every time, we check if the model we found is well defined by computing its average posterior probability ie the average posterior probability of membership for each group for those individuals that were assigned to it. In a ideal situation the assignment probability for each individual would be 1 and the Average Posterior Probability (AvePP) would be 1 too.Daniel S. Nagin (2005) considers that it should be at least 0.7 for all groups, so we use that criteria too. We also remove models which contain empty groups.

In a second step time, we try some affine functions to take into account the growth of the data during the time period. The first two step exhibit mostly 2-group and some 3-group solutions. In a third step, we then use the parameters of the different groups found in the previous steps as starting values and also find 4- and 5-group solutions. The typical trajectories of all 7 possible models with less than 10 groups are presented in figure 9.2.

The BIC and AIC, as well as Kass and Wasserman's p_k, are presented in table 9.1.

The highest value of BIC and AIC are obtained for the model with 5 groups with BIC = 15558.41 and AIC = 31241.46. The Kass and Wasserman criterion also clearly indicates that we should retain the 5-group solution.

Looking at the parameter estimates for the selected 5-group solution, we see that actually in group 5, the first and second degree parameters of ζ are non significant, meaning that for this group the precision parameter Φ is constant over time. We take this into account

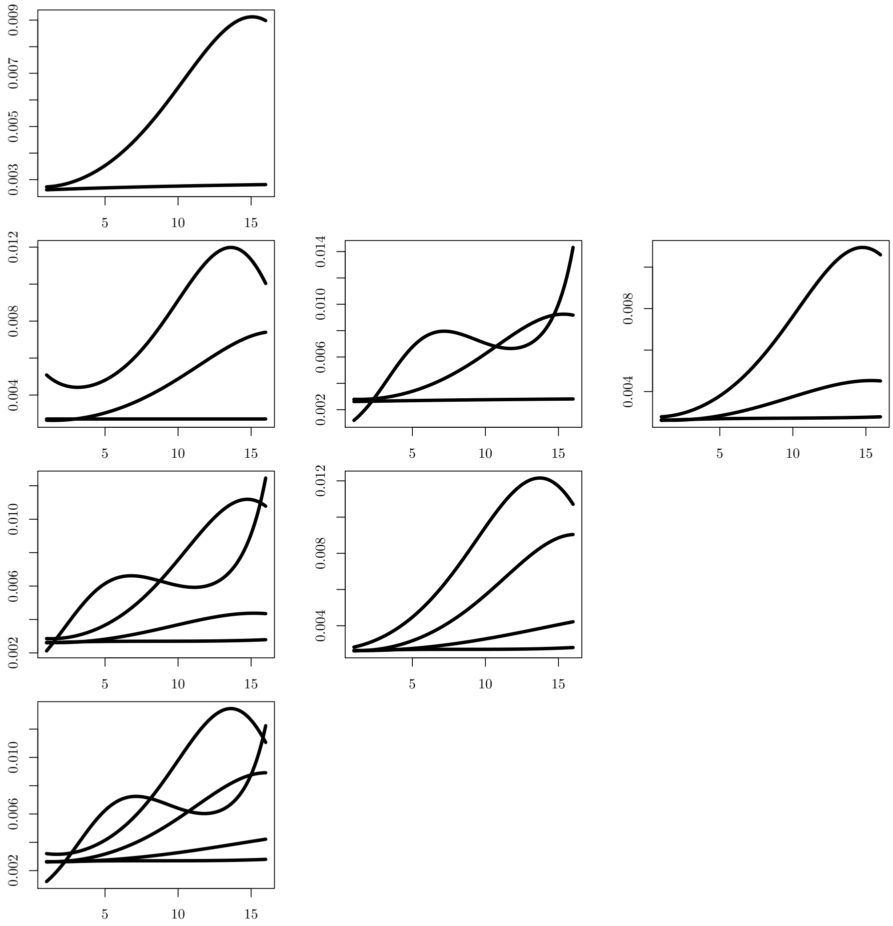


Figure 9.2: Typical trajectories of the different possible models

Number of groups	AIC	BIC	Prob
2	29851.99	14902.64	0.00000
3	30341.00	15142.28	0.00000
3	29945.96	14936.64	0.00000
3	30777.14	15352.23	0.00000
4	30839.69	15370.52	0.00000
4	31192.78	15547.06	0.00001
5	31241.46	15558.41	0.99999

Table 9.1: Model selection criteria

by rerunning the model with a constant ζ for group five and get the final model parameters presented in table 9.2.

Figure 9.3 shows the evolution of the precision parameters over time. We see that the trajectory of the parameters ϕ for group 5 is a horizontal line, and so, for this group, the precision is constant over time, while for most other groups, the trajectories become more homogeneous over time.

Param.	sd	Test
Beta 1		
-5.902	0.018	0.000
-0.052	0.013	0.000
0.020	0.002	0.000
-0.001	0.000	0.000
Beta 2		
-5.927	0.003	0.000
-0.015	0.003	0.000
0.005	0.001	0.000
0.000	0.000	0.005
Beta 3		
-5.659	0.133	0.000
-0.119	0.066	0.071
0.040	0.009	0.000
-0.002	0.000	0.000
Beta 4		
-5.962	0.011	0.000
0.018	0.005	0.000
-0.002	0.001	0.000
0.000	0.000	0.000
Beta 5		
-7.511	0.370	0.000
0.911	0.143	0.000
-0.102	0.016	0.000
0.004	0.001	0.000

Param.	sd	Test
Zeta 1		
14.648	0.309	0.000
-1.253	0.072	0.000
0.045	0.004	0.000
Zeta 2		
20.026	0.280	0.000
-1.818	0.088	0.000
0.066	0.005	0.000
Zeta 3		
9.454	0.374	0.000
-0.600	0.099	0.000
0.022	0.005	0.000
Zeta 4		
13.212	0.360	0.000
0.007	0.085	0.934
-0.008	0.004	0.061
Zeta 5		
7.422	0.146	0.000

Param.	sd	Test
Pi 1		
0.302	0.034	0.000
Pi 2		
0.183	0.028	0.000
Pi 3		
0.159	0.028	0.000
Pi 4		
0.319	0.034	0.000
Pi 5		
0.036	0.014	0.008

Table 9.2: Parameters of the final model

9.4.3 Description of the groups

The final model exhibits 4 main groups with 57, 35, 30 and 61 one countries respectively and one small group with 7 countries. Group membership of all countries in the sample can be found in table 9.4 in the appendix and can also be seen on the world map in figure 9.4.

Figure 9.5 shows the typical evolution of the contamination rate in the five groups, together with the initial trajectories of the 191 countries.

Group 1 contains countries with a quite fast and regularly increasing contamination rate, while the countries in group 2 exhibit a much smaller growth of the contamination rate. In the countries of group 3, the contamination rate grows even faster than in group 1, till the end of 2020 and starts declining at the beginning of 2021. Group 4 contains all countries that declare a contamination rate that stays very close to zero over the whole time interval in our study. The countries in group 5 have quite a particular behavior with two inflection points.

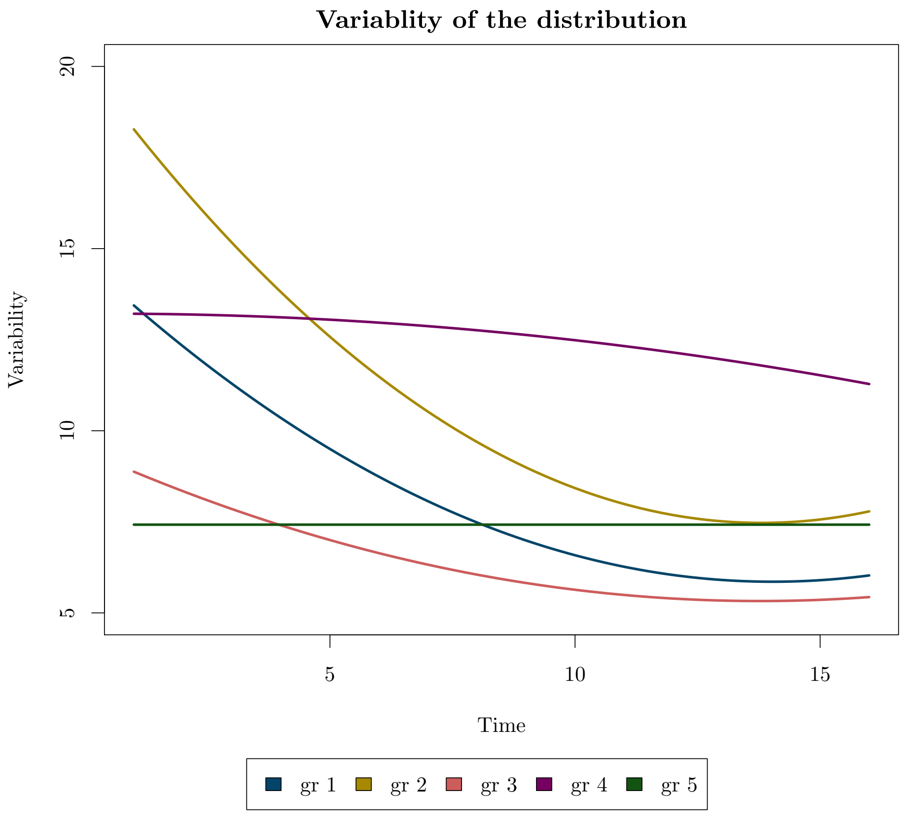


Figure 9.3: Variation of the precision parameters over time.

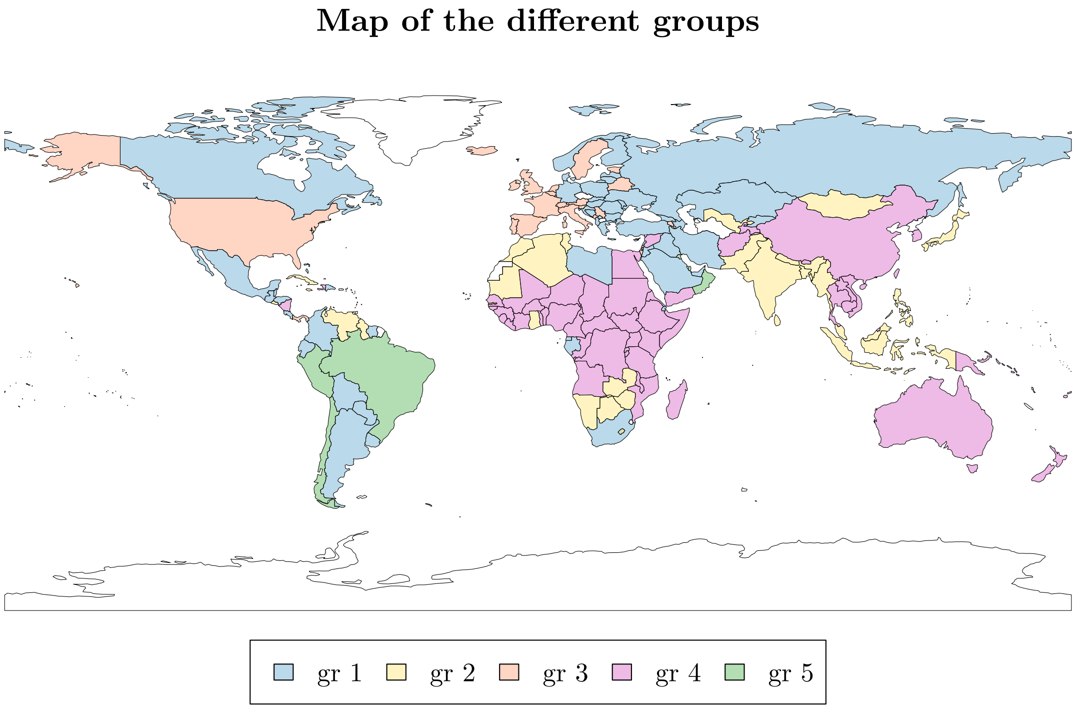


Figure 9.4: World map with the geographic distribution of the five groups

This group contains countries from South America like Chile, Peru or Brazil where the second increase of the contamination rate can be explained by the appearance of a new variant,

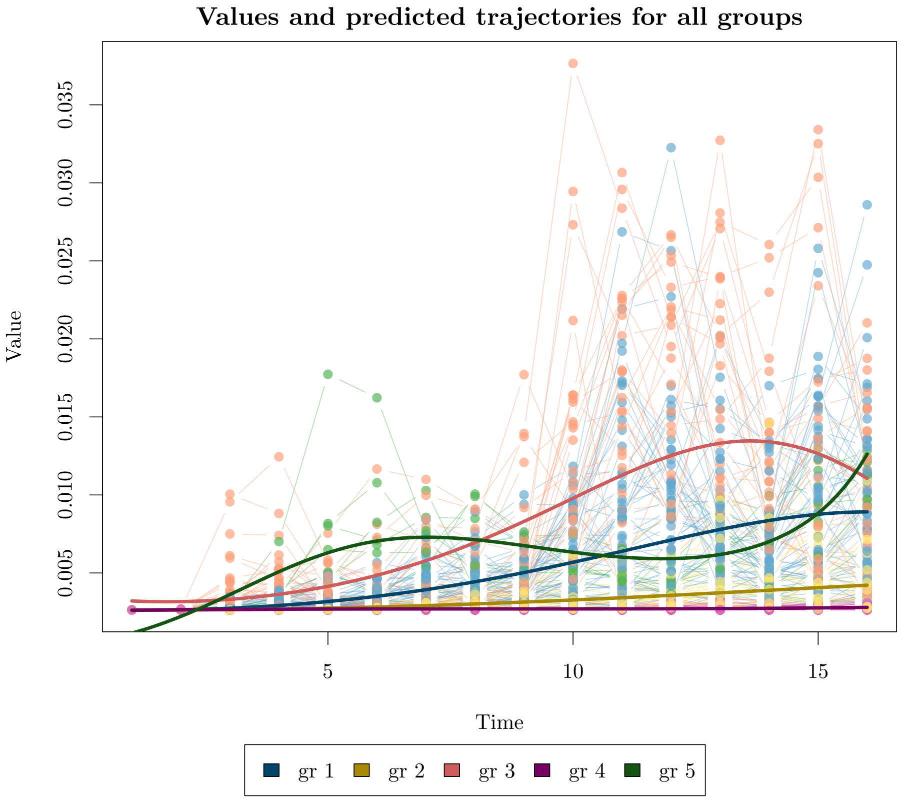


Figure 9.5: Evolution of the contamination rate for the five groups and all countries of the sample.

as described in Hojo de de Souza et al. (2021), but also countries from the Middle East like Kuwait, Qatar and Oman.

Table 9.4 shows the principal characteristics of the countries in the different groups.

Table 9.3: Means and standard deviations of the descriptive variables for each group

Variables	Group 1 mean (sd)	Group 2 mean (sd)	Group 3 mean (sd)	Group 4 mean (sd)	Group 5 mean (sd)
population size	19.97 (31.3)	76.41 (235.85)	23.38 (61.44)	48.49 (184.5)	39.64 (77.13)
median age	33.67 (7.76)	28.24 (6.88)	41.67 (5.29)	23.49 (7.51)	32.13 (2.19)
population density	122.67 (220.76)	207.31 (242.47)	1148.75 (3729.09)	126.64 (153.86)	286.19 (524.35)
aged 65 and older	10.54 (6.22)	6.75 (4.69)	17.08 (5.38)	4.83 (3.3)	5.27 (3.7)
aged 70 and older	6.87 (4.25)	4.16 (3.21)	10.76 (3.43)	2.81 (2.13)	3.23 (2.34)
gdp per capita	20583.62 (14511.5)	10267.22 (8325.13)	50138.17 (41050.34)	7655.52 (12468.09)	40673.95 (38570.57)
extreme poverty	3.63 (6.68)	12.96 (16.69)	0.81 (0.67)	31.28 (24.1)	2.73 (1.24)
cardiovascular death rate	259.68 (110.33)	285.53 (113.24)	172.99 (108)	306.11 (121.94)	161.7 (56.6)
diabetes prevalence	8.35 (3.3)	7.75 (3.63)	6.93 (2.84)	7.72 (5.59)	10.95 (4.08)
handwashing facilities	72.84 (22.95)	55.53 (28.91)	95.88 (2.6)	32.5 (27)	96.6 (1.13)
hospital beds	3.63 (2.18)	2.47 (2.5)	4.45 (2.67)	1.87 (2.1)	1.79 (0.38)
life expectancy	75.2 (4.56)	70.73 (6.42)	80.97 (3.16)	67.18 (7.41)	77.9 (1.95)
human development index	0.79 (0.09)	0.68 (0.1)	0.9 (0.05)	0.59 (0.14)	0.8 (0.04)
Stringency index	64.75 (16.98)	64.1 (15.74)	58.01 (18.17)	49.9 (16.22)	68.72 (22.02)

We observe that the contamination rate is higher in the groups with a higher median age or with a higher proportion of old people. The age variables are indeed highest in groups 3 and 1 and lowest in group 4. We also see that the countries in the groups with the largest increase of the contamination rate have the highest life expectancy and GDP per capita and the lowest poverty rate. Moreover, group 3 has by far the highest population density, whereas there is no visible trend linking the population density to the contamination rate in the other four groups.

On the other hand, we see that the average stringency index is almost the same for all five groups, so at a first glance, there does not seem to be a link between the contamination rate and the stringency index.

Carrillo-Larco and Castillo-Cara (2020) use a k-means clustering algorithm on cross-sectional data and find a division of the countries into clusters that are quite similar to our groups. They use some selected diseases, socio-economic status, air pollution and health system as descriptive variables, but do of course not get any typical evolution trajectories, since they perform a static analysis.

In the following paragraph, we will formally test which of the descriptive variables have a significant influence on group membership.

9.4.4 Predictors of Trajectory Group Membership

We test several covariates to understand if some of them influence the group membership probabilities. We use the median age, the population density, the proportion of people elder than 65 older, the life expectancy and the mean of the stringency index over the observed time period as risk measures X that potentially influence group membership in the generalized finite mixture model. Since that part of the model is essentially equivalent to a multinomial logit model, we have to use one of the groups as comparison group and test then for all covariates if and if yes, how much, they influence the membership probabilities for the other groups. We choose group 4 as comparison group. Table 9.5 reports the results.

The median age affects membership to groups 1 ($p = 0.02445$), 2 ($p = 0.088$), and especially 5 (p=0), in the sense that countries with an older population have a higher chance to belong to these groups. The high influence of the median age on the membership probability of group 5 is somewhat counterbalanced by the fact that a high proportion of people older than 65 has a highly significant $p = 0.002$ but negative influence on group membership in that group.

Group 1	Albania, Argentina, Azerbaijan, Bahamas, Belize, Bolivia, Bosnia and Herzegovina, Bulgaria, Canada, Cape Verde, Colombia, Costa Rica, Croatia, Cyprus, Denmark, Djibouti, Dominican Republic, Ecuador, Equatorial Guinea, Finland, Gabon, Georgia, Germany, Greece, Guatemala, Honduras, Hungary, Iran, Iraq, Jordan, Kazakhstan, Kosovo, Kyrgyzstan, Latvia, Lebanon, Libya, Lithuania, Malta, Mexico, Moldova, North Macedonia, Norway, Palestine, Paraguay, Poland, Romania, Russia, Sao Tome and Principe, Saudi Arabia, Seychelles, Slovakia, South Africa, Suriname, Turkey, Ukraine, United Arab Emirates, Uruguay
Group 2	Algeria, Antigua and Barbuda, Bangladesh, Barbados, Botswana, Comoros, Cuba, El Salvador, Eswatini, Gambia, Ghana, Guyana, India, Indonesia, Jamaica, Japan, Lesotho, Malaysia, Mauritania, Mongolia, Morocco, Myanmar, Namibia, Nepal, Pakistan, Philippines, Saint Lucia, Saint Vincent and the Grenadines, Sri Lanka, Trinidad and Tobago, Tunisia, Uzbekistan, Venezuela, Zambia, Zimbabwe
Group 3	Andorra, Armenia, Austria, Bahrain, Belarus, Belgium, Czechia, Estonia, France, Iceland, Ireland, Israel, Italy, Liechtenstein, Luxembourg, Monaco, Montenegro, Netherlands, Panama, Portugal, San Marino, Serbia, Singapore, Slovenia, Spain, Sweden, Switzerland, United Kingdom, United States, Vatican
Group 4	Afghanistan, Angola, Australia, Benin, Bhutan, Brunei, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, China, Congo, Cote d'Ivoire, Democratic Republic of Congo, Dominica, Egypt, Eritrea, Ethiopia, Fiji, Grenada, Guinea, Guinea-Bissau, Haiti, Kenya, Laos, Liberia, Madagascar, Malawi, Mali, Marshall Islands, Mauritius, Micronesia (country), Mozambique, New Zealand, Nicaragua, Niger, Nigeria, Papua New Guinea, Rwanda, Saint Kitts and Nevis, Samoa, Senegal, Sierra Leone, Solomon Islands, Somalia, South Korea, South Sudan, Sudan, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Timor, Togo, Uganda, Vanuatu, Vietnam, Yemen
Group 5	Brazil, Chile, Kuwait, Maldives, Oman, Peru, Qatar

Table 9.4: Countries belonging to each group.

Figure 9.6 shows the boxplots of the median age for all five groups of countries.

We see that the median of the median age covariant is by far highest in group five and actually nearly all countries in group five have a relatively high median age. More generally, we detect a clear link between a high median age of a population and membership in a group with a high contamination rate.

Finally, we see that a high value of the stringency index has a highly significant positive influence in group membership in all four groups. Accordingly, a high value of the stringency index also decreases the probability of membership in the baseline group. Now this results

	Group 1			Group 2		
	Estimate	Std. Error	Prob>|T|	Estimate	Std. Error	Prob>|T|
intercept	-16.812	4.681	0	-4.805	3.422	0.16
median age	0.193	0.086	**0.024**	0.172	0.101	**0.088**
population density	-0.003	0.002	0.093	0.000	0.001	0.869
aged 65 older	-0.021	0.132	0.871	-0.060	0.126	0.631
life expectancy	0.073	0.080	0.364	-0.073	0.071	0.304
stringency index	0.112	0.023	**0**	0.092	0.023	**0**

	Group 3			Group 5		
	Estimate	Std. Error	Prob>|T|	Estimate	Std. Error	Prob>|T|
intercept	-67.733	19.400	0	-73.689	23.469	0.002
median age	0.129	0.158	0.412	0.418	0.205	**0.041**
population density	0.000	0.001	0.784	0.000	0.001	0.926
aged 65 older	0.109	0.178	0.542	-0.640	0.206	**0.002**
life expectancy	0.646	0.223	**0.004**	0.646	0.283	**0.023**
stringency index	0.185	0.054	**0.001**	0.228	0.075	**0.002**

Table 9.5: Predictors of group membership.

seems of course surprising. But we actually did something here, that we are not supposed to do. In generalized finite mixture models, the covariates X that potentially influence group membership have to be measured before the beginning of the trajectories Y. This is not the case for the stringency index in our example. The government measures that are summarized in that index have not been taken before the pandemic, but during the first part of the outbreak. So, it is obvious that the stringency index cannot influence group memberships here. What our result shows is actually rather the contrary. The governments of countries in groups with a higher contamination rate have usually taken harsher measures, which explains that the coefficients in groups 1,2,3 and 5 are significant and strictly positive with respect to group

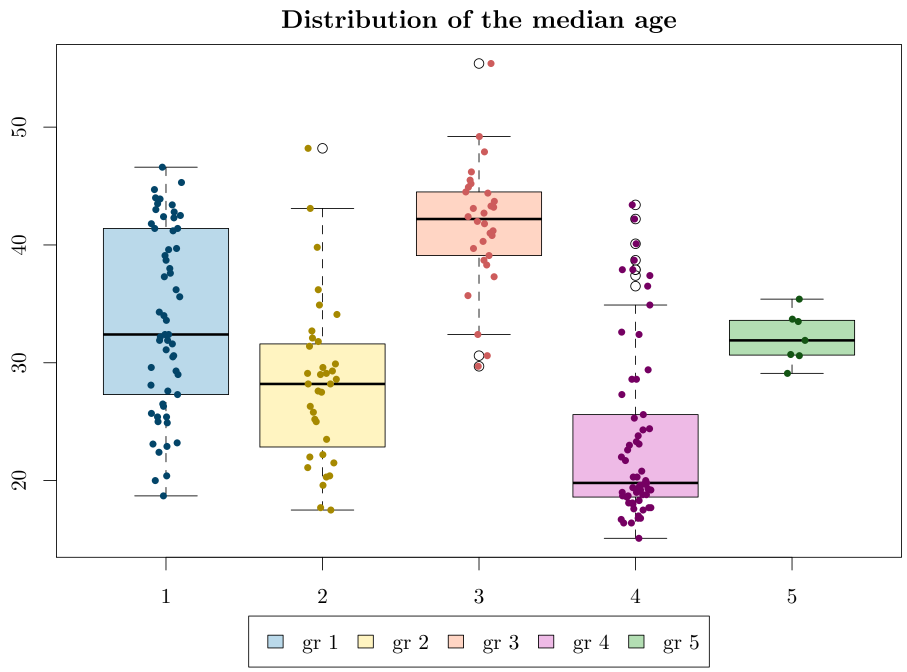


Figure 9.6: Boxplots of the median age for all 5 groups.

4, which contains the countries not affected a lot by the pandemic during the time of our study.

On the other hand these four significant theta coefficients are quite close one to each other. If we want to understand, if they differ significantly among themselves, we can use a χ^2-based test of multiple contrasts (Nagin, 2005). The degree of freedom of this test equals the number of equality constraints being tested or the number of different coefficients being tested minus 1. Here, the degree of freedom is 3, since we perform pairwise equality tests between the four theta coefficients. More precisely, we perform the test

$$H_0 : \theta_2^{st} = \theta_3^{st} \text{ and } \theta_2^{st} = \theta_4^{st} \text{ and } \theta_2^{st} = \theta_5^{st} \tag{9.1}$$

$$H_1 : \theta_2^{st} \neq \theta_3^{st} \text{ or } \theta_2^{st} \neq \theta_4^{st} \text{ or } \theta_2^{st} \neq \theta_5^{st} \tag{9.2}$$

If we use the notations $\theta = \left(\theta_2^{st}, \theta_3^{st}, \theta_4^{st}, \theta_5^{st}\right)'$, $q = (0,0,0)'$ and

$$H = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix},$$

our hypotheses can be written

$$H_0 : H\theta = q \tag{9.3}$$

$$H_1 : H\theta \neq 0. \tag{9.4}$$

The χ^2 distance is then defined by

$$\chi^2 = \left(H\hat{\theta} - q\right)' \left(HV_{\hat{\theta}}H'\right)^{-1} \left(H\hat{\theta} - q\right), \tag{9.5}$$

where $\hat{\theta}$ denotes the counterpart vector of the coefficients and $V_{\hat{\theta}}$ is the matrix variance/covariance of the parameter estimates in $\hat{\theta}$.
We obtain $\chi^2 = 5.62$ with 3 degrees of freedom, which is far short of significance. So, whereas we have shown that groups 1, 2, 3 and 5 have a significantly higher stringency index than group 1, we do not find any significant pairwise difference between these four groups.

9.4.5 Stringency as time dependent covariate

The generalized finite mixture model allows to test if time-dependent covariates have a significant relationship with the shape of the typical trajectories in the different groups. Here it makes sense to analyze if the different measures the governments took against the spread of the pandemic were successful. In other words, we are interested in testing if the typical contamination rate trajectories were flattened in case of a high stringency index.
Thus, here W_{it} denotes the stringency index at time t, for $1 \leq t \leq 16$, for country i. For this application, we had to remove countries without values for the stringency index[1] and in case of isolated missing data, we completed the variable by linear interpolation.

Table 9.6 shows the parameter estimations of the model. Their interpretation is the same as before except for the delta coefficients, which illustrate the relationship between the stringency index and the different typical group trajectories.

For the groups 1, 3 and 5, i.e. the groups with a high contamination rate, the delta parameters are significant and positive, although very small, while for the other two groups, they are not significant. Here again, the small but significant relationship between the delta parameters and the typical trajectories mostly indicates that the countries with a high contamination rate took slightly stricter measures against the spread of the virus than the countries which

[1] Antigua and Barbuda , Armenia , Comoros , Equatorial Guinea , Grenada , Guinea-Bissau , Liechtenstein , Maldives , Marshall Islands , Micronesia (country) , Montenegro , North Macedonia , Saint Kitts and Nevis , Saint Lucia , Saint Vincent and the Grenadines , Samoa , Sao Tome and Principe , Vatican.

Param.	sd	Test
Beta 1		
-5.843	0.026	0.000
-0.120	0.024	0.000
0.029	0.004	0.000
-0.001	0.000	0.000
Beta 2		
-5.927	0.003	0.000
-0.014	0.004	0.000
0.005	0.001	0.000
0.000	0.000	0.001
Beta 3		
-5.602	0.117	0.000
-0.421	0.070	0.000
0.076	0.009	0.000
-0.003	0.000	0.000
Beta 4		
-5.972	0.012	0.000
0.012	0.005	0.018
-0.001	0.001	0.043
0.000	0.000	0.027
Beta 5		
-7.304	0.366	0.000
0.701	0.147	0.000
-0.078	0.017	0.000
0.003	0.001	0.000
Phi 1		
14.337	0.317	0.000
-1.164	0.076	0.000
0.040	0.004	0.000
Phi 2		
19.866	0.570	0.000
-1.710	0.125	0.000
0.061	0.006	0.000
Phi 3		
9.624	0.369	0.000
-0.521	0.097	0.000
0.016	0.005	0.003
Phi 4		
12.887	0.372	0.000
0.148	0.085	0.082
-0.015	0.004	0.000
Phi 5		
7.384	0.137	0.000
Delta 1		
0.001	0.000	0.001
Delta 2		
0.000	0.000	0.955
Delta 3		
0.010	0.001	0.000
Delta 4		
0.000	0.000	0.955
Delta 5		
0.004	0.001	0.004
Prob. 1		
0.328	0.039	0.00
Prob. 2		
0.175	0.030	0.00
Prob. 3		
0.156	0.030	0.00
Prob. 4		
0.301	0.035	0.00
Prob. 5		
0.040	0.016	0.01

Table 9.6: parameters of the final model with time dependent covariates.

were less affected by the virus. But conversely, this also shows that we cannot see any sign of efficiency of the sanitary measures taken by the different countries against the propagation of the virus during the first part of the pandemic.

CHAPTER 10

Conclusion

This dissertation has made significant contributions to the field of trajectory analysis through the development of the TRAJER framework, an innovative R package designed for finite mixture modeling of longitudinal data. Tailored to address heterogeneity in time series data across domains such as finance, criminology, medicine, sports analytics, and others, TRAJER integrates advanced statistical methodologies with robust computational techniques. The following sections provide a detailed summary of the contributions from each chapter, followed by future research directions to extend the framework's capabilities.

Chapter 1 laid the foundation for the thesis by introducing the importance of trajectory analysis in longitudinal studies. It highlighted the prevalence of time series data in various fields and the need for models that capture both inter-individual and intra-individual variability without predefined group assumptions. The chapter introduced the Latent Growth Model (LGM) framework as a cornerstone for studying patterns of change over time and outlined the objectives of developing TRAJER to identify latent trajectory groups and their predictors, drawing on Daniel S. Nagin (2005).

Chapter 2 provided a comprehensive overview of finite mixture models (FMMs) as a statistical framework for clustering heterogeneous longitudinal data. The chapter detailed the theoretical underpinnings of FMMs, including their probabilistic structure and applications in trajectory modeling, referencing McLachlan (2004) and Lindsay (1995). It emphasized the flexibility of FMMs in accommodating various distributional assumptions, setting the stage for the specific models developed in subsequent chapters.

Chapter 3 was a cornerstone of the thesis, presenting the development of four distributional models within TRAJER: Censored Normal (CNORM), Zero-Inflated Poisson (ZIP), Logit, and Beta. Each model was tailored to specific data characteristics:

- **CNORM** addressed censored data, with sections on likelihood, EM algorithm, and numerical methods (Section 3.2.5) ensuring robust parameter estimation.

- **ZIP** modeled count data with excess zeros, incorporating link functions, Iteratively Reweighted Least Squares (IRLS, Section 3.3.5), and covariance estimation (Section 3.3.8) for precise inference, inspired by Lambert (1992).

- **Logit** handled binary outcomes, with detailed likelihood formulations and EM-based estimation (Section 3.3.4).

- **Beta** tackled bounded continuous responses, with numerical applications (Section 3.4.3) demonstrating its utility, following Smithson and Verkuilen (2006).

The chapter also explored non-linear mixture models (Section 3.5), enhancing TRAJER's flexibility for complex trajectories. Each model was accompanied by standard error estimation and covariance computations, ensuring rigorous statistical inference.

Chapter 4 addressed a critical theoretical challenge in FMMs: model identifiability. The chapter introduced the concept of identifiability (Section 4.1.1), characterized identifiable mixture families (Section 4.1.3), and explored identifiability in regression mixtures (Section 4.1.4), building on Teicher (1963a), Yakowitz and Spragins (1968) and C. Hennig (2000). By establishing conditions under which model parameters are uniquely determined, this chapter provided a theoretical foundation that enhances the reliability and interpretability of TRAJER's outputs.

Chapter 5 focused on practical implementation by developing strategies for initializing model parameters to ensure stable convergence. Specific approaches were tailored for CNORM (Section 5.1.1), Logit (Section 5.1.2), ZIP (Section 5.1.3), and Beta (Section 5.1.4) models. These strategies mitigated the sensitivity of EM or quasi-Newton algorithms to initial values, making TRAJER computationally robust and user-friendly for practitioners, with implementation details aligned with your expertise in R package development.

Chapter 6 addressed the critical task of selecting the optimal number of trajectory groups and evaluating model performance. Section 6.1 reviewed model selection criteria, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Calinski-Harabasz-DTW (CHDTW), tailored for time series clustering. Section 6.2 introduced performance metrics such as cross-validation error rates and Slope Heuristics (SH) for longitudinal data. These methods, implemented in TRAJER, provide users with rigorous tools to choose models that balance fit and parsimony, aligning with her statistical expertise.

Chapter 7 extended TRAJER's capabilities to model multiple outcomes simultaneously through group-based multi-trajectory modeling (GBMTM). Section 7.1 outlined the theoretical framework for GBMTM, which captures correlated trajectories across multiple variables. Section 7.2 detailed the extension of EM algorithms to handle multivariate longitudinal data.

Section 7.3 presented numerical applications, demonstrating GBMTM's ability to uncover joint trajectory patterns in datasets like criminology or medical studies. This chapter enhanced TRAJER's applicability to complex, multivariate time series, a key advancement for applied research.

Chapter 8 introduced `trajeR`, the R package implementing the TRAJER framework, emphasizing its design and functionality for statisticians and practitioners. Section 8.1 described the package's architecture, built on R's statistical ecosystem, ensuring accessibility and scalability. Section 8.2 detailed user-facing functions for fitting CNORM, ZIP, Logit, and Beta models, with diagnostics like Average Posterior Probability (AvePP) and Odds of Correct Classification (OCC). Section 8.5.3 (Estimated Group Probabilities) and 8.5.4 (Confidence Intervals) provided robust inference tools, while the Adequacy Matrix (Section 8.5.5) evaluated group coherence. Section 8.6 covered visualization tools, enhancing interpretability. The open-source nature of `trajeR` encourages community contributions.

Chapter 9 demonstrated TRAJER's versatility through real-world applications. The chapter applied the ZIP (Section 9.1), CNORM (Section 9.2), Logit (Section 9.3), and Beta (Section 9.4) models to diverse datasets, including COVID-19 contamination data for the Beta model. Subsections detailed data preparation (Section 9.4.1), model selection (Section 9.4.2), group descriptions (Section 9.4.3), predictors of group membership (Section 9.4.4), and time dependent covariates like stringency (Section 9.4.5). Tables (e.g., Table 9.1 for model selection criteria) and descriptive statistics (Table 9.3) provided actionable insights, showcasing TRAJER's ability to uncover latent trajectory groups and their dynamics across fields like medicine and criminology.

The TRAJER framework opens several avenues for future research to address emerging challenges in longitudinal data analysis:

- Dropout Modeling: Non-ignorable dropout is a common issue in longitudinal studies, potentially biasing trajectory estimates. Extending TRAJER to incorporate joint modeling of longitudinal and dropout processes, using shared random effects or latent class structures could improve robustness, particularly for sports analytics applications tracking athlete performance.

- Composite Likelihood Methods: For high-dimensional or complex datasets, full likelihood estimation can be computationally intensive. Composite likelihood approachescould approximate likelihoods to reduce computational burdens while maintaining statistical efficiency, especially for multivariate trajectory models.

- Supervised Group-Based Trajectory Modeling (GBTM): Integrating supervised GBTM could enhance TRAJER's predictive capabilities by incorporating external covariates or outcome variables to guide cluster formation, bridging exploratory and predictive modeling.

- Spline-Based Trajectory Modeling: Replacing polynomial bases with splinescould improve TRAJER's flexibility in capturing non-linear trajectories. B-splines or penalized splines could model complex patterns more effectively, particularly for irregular time series in fields like medicine or finance.

In conclusion, this thesis has established TRAJER as a robust and versatile framework for finite mixture modeling of longitudinal trajectories, implemented through the `trajeR` R package. By addressing theoretical, computational, and applied aspects across its chapters, the work provides a comprehensive toolkit for researchers in statistics, mathematics, and applied fields like sports analytics. The proposed extensions—dropout modeling, composite likelihood, supervised GBTM, and spline-based modeling—offer exciting opportunities to further refine TRAJER, ensuring its relevance in tackling complex longitudinal data challenges. As an R package, `trajeR` is well-positioned to empower researchers to uncover meaningful patterns in time series data, driving advancements in both methodological and applied research.

References

Abarda, Abdallah et al. (2020). “Latent Transition Analysis (LTA) : A Method for Identifying Differences in Longitudinal Change Among Unobserved Groups”. In: *Procedia Computer Science* 170, pp. 1116–1121.

Abdelkhalek, Touhami and Jean-Marie Dufour (Nov. 1998). “Statistical Inference for Computable General Equilibrium Models, with Application to A Model of the Moroccan Economy”. In: *Review of Economics and Statistics* 80.4, pp. 520–534.

Ahmad, Khalaf E. and Essam K. Al-Hussaini (Dec. 1982). “Remarks on the Non-Identifiability of Mixtures of Distributions”. In: *Annals of the Institute of Statistical Mathematics* 34.3, pp. 543–544.

Ahn, Hee-Kap et al. (Aug. 2020). “Middle Curves Based on Discrete Fréchet Distance”. In: *Computational Geometry* 89, p. 101621.

Akaike, Hirotugu (1974). “A New Look at the Statistical Model Identification”. In: *Selected Papers of Hirotugu Akaike*. Ed. by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. New York, NY: Springer New York, pp. 215–222.

Akhanli, Serhat Emre and Christian Hennig (June 23, 2020). *Comparing Clusterings and Numbers of Clusters by Aggregation of Calibrated Clustering Validity Indexes*. arXiv: `2002.01822 [stat]`. URL: `http://arxiv.org/abs/2002.01822` (visited on 09/19/2023). Prepublished.

Alerini, Julien, Madalina Olteanu, and James Ridgway (n.d.). “MARKOV AND THE DUCHY OF SAVOY: SEGMENTING A CENTURY WITH REGIME-SWITCHING MODELS”. In: ().

Allison, Paul David (2002). *Missing Data*. Sage University Papers. Quantitative Applications in the Social Sciences no. 07-136. Thousand Oaks, Calif: Sage Publications. 93 pp.

Allman, Elizabeth S., Catherine Matias, and John A. Rhodes (Dec. 1, 2009). “Identifiability of Parameters in Latent Structure Models with Many Observed Variables”. In: *The Annals of Statistics* 37 (6A). arXiv: `0809.5032 [math, stat]`.

Altman, Micah, Jeff Gill, and Michael P. McDonald (Dec. 12, 2003). *Numerical Issues in Statistical Computing for the Social Scientist*. 1st ed. Wiley Series in Probability and Statistics. Wiley.

Alvarez, Emiliano, Juan Gabriel Brida, and Erick Limas (2020). “Comparisons of COVID-19 Dynamics in the Different Countries of the World Using Time-Series Clustering”. In: *medRxiv : the preprint server for health sciences*, pp. 2020–08.

Anglada, Olivier and Jean François Maurras (Oct. 2003). “Enveloppe convexe des hyperplans d’un espace affine fini”. In: *RAIRO - Operations Research* 37.4, pp. 213–219.

Arbous, A. G. and J. E. Kerrich (Dec. 1951). “Accident Statistics and the Concept of Accident-Proneness”. In: *Biometrics* 7.4, p. 340. JSTOR: 3001656.

Ardila, Federico (Oct. 3, 2017). *Tutte Polynomials of Hyperplane Arrangements and the Finite Field Method*. arXiv: 1710.01424 [math]. URL: http://arxiv.org/abs/1710.01424 (visited on 09/19/2023). Pre-published.

Arenberg, Kasteelpark (n.d.). “REGULARIZATION TECHNIQUES IN MODEL FITTING AND PARAMETER ESTIMATION”. In: ().

Atienza, N., J. Garcia-Heras, and J. M. Muñoz-Pichardo (Apr. 2006). “A New Condition for Identifiability of Finite Mixture Distributions”. In: *Metrika* 63.2, pp. 215–221.

Auder, Benjamin, Elisabeth Gassiat, and Mor Absa Loum (Feb. 12, 2020). *Least Squares Moment Identification of Binary Regression Mixtures Models*. arXiv: 1811.01714 [math, stat]. URL: http://arxiv.org/abs/1811.01714 (visited on 09/19/2023). Pre-published.

— (n.d.). “Mixture of Generalized Linear Models: Identifiability and Applications”. In: ().

Audibert, Thierry (n.d.). “La me´thode de Newton et ses variantes pour l’optimisation”. In: ().

Augustin, Nicole H., Roger P. Cummins, and Donald D. French (Oct. 2001). “Exploring Spatial Vegetation Dynamics Using Logistic Regression and a Multinomial Logit Model: *Exploring Spatial Vegetation Dynamics*”. In: *Journal of Applied Ecology* 38.5, pp. 991–1006.

Aujol, Mr Jean-François et al. (n.d.). “Soutenue le 03 Décembre 2015”. In: ().

Aziz, Nur Ain Abd et al. (2016). “Modeling Multinomial Logistic Regression on Characteristics of Smokers after the Smoke-Free Campaign in the Area of Melaka”. In: ADVANCES IN INDUSTRIAL AND APPLIED MATHEMATICS: Proceedings of 23rd Malaysian National Symposium of Mathematical Sciences (SKSM23). Johor Bahru, Malaysia, p. 060020.

Ba, Demba et al. (Jan. 2014). “Convergence and Stability of Iteratively Re-weighted Least Squares Algorithms”. In: *IEEE Transactions on Signal Processing* 62.1, pp. 183–195.

Baey, Charlotte, Samis Trevezas, and Paul-Henry Cournède (Mar. 18, 2016). “A Non Linear Mixed Effects Model of Plant Growth and Estimation via Stochastic Variants of the EM Algorithm”. In: *Communications in Statistics - Theory and Methods* 45.6, pp. 1643–1669.

Bailet, Pauline (n.d.). “présentée et soutenue par”. In: ().

Balko, Martin, Josef Cibulka, and Pavel Valtr (Jan. 3, 2018). *Covering Lattice Points by Subspaces and Counting Point-Hyperplane Incidences.* arXiv: `1703.04767 [math]`. URL: `http://arxiv.org/abs/1703.04767` (visited on 09/19/2023). Pre-published.

Baraud, Yannick and Lucien Birgé (Nov. 29, 2017). *Rho-Estimators Revisited: General Theory and Applications.* arXiv: `1605.05051 [math, stat]`. URL: `http://arxiv.org/abs/1605.05051` (visited on 09/19/2023). Pre-published.

Baraud, Yannick, Lucien Birgé, and Mathieu Sart (Feb. 2017). “A New Method for Estimation and Model Selection: ρ-Estimation”. In: *Inventiones mathematicae* 207.2, pp. 425–517. arXiv: `1403.6057 [math, stat]`.

Barcenas, Diomedes (n.d.). “The Fundamental Theorem of Calculus for Lebesgue Integral”. In: ().

Barndorff-Nielsen, O (Aug. 1965a). “Identifiability of Mixtures of Exponential Families”. In: *Journal of Mathematical Analysis and Applications* 12.1, pp. 115–121.

— (Aug. 1965b). “Identifiability of Mixtures of Exponential Families”. In: *Journal of Mathematical Analysis and Applications* 12.1, pp. 115–121.

Bartolucci, Francesco, Francesca Chiaromonte, et al. (Apr. 3, 2017). “Composite Likelihood Inference in a Discrete Latent Variable Model for Two-Way “Clustering-by-Segmentation” Problems”. In: *Journal of Computational and Graphical Statistics* 26.2, pp. 388–402.

Bartolucci, Francesco and Monia Lupparelli (Jan. 2, 2016). “Pairwise Likelihood Inference for Nested Hidden Markov Chain Models for Multilevel Longitudinal Data”. In: *Journal of the American Statistical Association* 111.513, pp. 216–228.

Bartoszyński, Robert and Magdalena Niewiadomska-Bugaj (2008). *Probability and Statistical Inference.* 2nd ed. Hoboken, N.J: Wiley-Interscience. 647 pp.

Bashir, Shaheena and E. M. Carter (Sept. 15, 2012). “Robust Mixture of Linear Regression Models”. In: *Communications in Statistics - Theory and Methods* 41.18, pp. 3371–3388.

Bates, Douglas M. and Donald G. Watts (Aug. 26, 1988). *Nonlinear Regression Analysis and Its Applications.* 1st ed. Wiley Series in Probability and Statistics. Wiley.

Batista, Levy, Thierry Bastogne, and El-Hadi Djermoune (n.d.). “Parameters Identification of a Population of Systems Based on EM Algorithm”. In: ().

Baudry, Jean-Patrick, Cathy Maugis, and Bertrand Michel (Mar. 2012). “Slope Heuristics: Overview and Implementation”. In: *Statistics and Computing* 22.2, pp. 455–470.

Bauer, Daniel J. and Patrick J. Curran (2003). “Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes.” In: *Psychological Methods* 8.3, pp. 338–363.

Bazzoli, Caroline and Sophie Lambert-Lacroix (n.d.). “A Comparison of Methods for Analysing Logistic Regression Models with Both Clinical and Genomic Variables”. In: ().

Bel, Koen, Dennis Fok, and Richard Paap (May 28, 2018). “Parameter Estimation in Multivariate Logit Models with Many Binary Choices”. In: *Econometric Reviews* 37.5, pp. 534–550.

Bel, Koen and Richard Paap (2014). *A Multivariate Model for Multinomial Choices.*

— (n.d.). “A Multivariate Model for Multinomial Choices”. In: ().

Bel, L et al. (n.d.). “Le Modèle Linéaire et ses Extensions”. In: ().

Ben-Akiva, Moshe E and Steven R Lerman (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* Vol. 9. MIT press.

Benaglia, Tatiana et al. (2009). “**Mixtools** : An *R* Package for Analyzing Finite Mixture Models”. In: *Journal of Statistical Software* 32.6.

Berndt, E K, B H Hall, and R E Hall (n.d.). “Estimation and Inference in Nonlinear Structural Models”. In: ().

Bertrand, Frédéric et al. (2013). “Régression Bêta PLS”. In: 154.3.

Besag, Julian (1974a). “Spatial Interaction and the Statistical Analysis of Lattice Systems”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 192–225.

— (Jan. 1974b). “Spatial Interaction and the Statistical Analysis of Lattice Systems”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2, pp. 192–225.

Bhavani, Sivasubramanium V., Kyle A. Carey, et al. (Aug. 1, 2019). “Identifying Novel Sepsis Subphenotypes Using Temperature Trajectories”. In: *American Journal of Respiratory and Critical Care Medicine* 200.3, pp. 327–335.

Bhavani, Sivasubramanium V., Elbert S. Huang, et al. (Dec. 2020). “Novel Temperature Trajectory Subphenotypes in COVID-19”. In: *Chest* 158.6, pp. 2436–2439.

Biernacki, Christophe (2009). “Pourquoi les mod‘eles de m´elange pour la classification ?” In.

Biernat, Éric and Michel Lutz (2015). *Data science: fondamentaux et études de cas machine learning avec Python et R.* Paris: Eyrolles.

Bilmes, Jeff A (n.d.). “A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models”. In: ().

Bingham, N. H. and John M. Fry (2010). *Regression.* Springer Undergraduate Mathematics Series. London: Springer London.

Birgé, Lucien and Pascal Massart (May 2007). “Minimal Penalties for Gaussian Model Selection”. In: *Probability Theory and Related Fields* 138.1-2, pp. 33–73.

Blaze, Thomas James (n.d.). “ENUMERATING THE CORRECT NUMBER OF CLASSES IN A SEMIPARAMETRIC GROUP-BASED TRAJECTORY MODEL”. In: ().

Blischke, W R (1964). “Estimating the Parameters of Mixtures of Binomial Distributions”. In.

Bock, R. Darrell and Murray Aitkin (Dec. 1981). “Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm”. In: *Psychometrika* 46.4, pp. 443–459.

Bonat, Wagner Hugo, Paulo Justiniano RIBEIRO Jr, and Walmes Marques Zeviani (2012). “REGRESSION MODELS WITH RESPONSES ON THE UNITY INTERVAL: SPECIFICATION, ESTIMATION AND COMPARISON”. In.

Bousquet, Faustine, Christian Lavergne, and Sophie Lèbre (n.d.). “Classification de campagnes de publicité mobile: Modèle de mélange pour données longitudinales et non gaussiennes”. In: ().

Brakatsoulas, Sotiris et al. (n.d.). “On Map-Matching Vehicle Tracking Data”. In: ().

Brame, Robert, Daniel S. Nagin, and Larry Wasserman (Mar. 2006). “Exploring Some Analytical Characteristics of Finite Mixture Models”. In: *Journal of Quantitative Criminology* 22.1, pp. 31–59.

Brier, Norman (1989). “The Relationship between Learning Disability and Delinquency: A Review and Reappraisal”. In: *Journal of Learning Disabilities* 22.9, pp. 546–553.

Broyden, Charles George (1970). “The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations”. In: *IMA Journal of Applied Mathematics* 6.1, pp. 76–90.

Buchin, Kevin, Anne Driemel, Joachim Gudmundsson, et al. (July 20, 2018). *Approximating (k,ℓ)-Center Clustering for Curves.* arXiv: 1805.01547 [cs]. URL: http://arxiv.org/abs/1805.01547 (visited on 09/19/2023). Pre-published.

Buchin, Kevin, Anne Driemel, Natasja Van De L’Isle, et al. (Nov. 5, 2019). “Klcluster: Center-based Clustering of Trajectories”. In: *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* SIGSPATIAL ’19: 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Chicago IL USA: ACM, pp. 496–499.

Bull, John W., ed. (2009). *Linear and Non-Linear Numerical Analysis of Foundations.* London: Spon Press/Taylor & Francis. 452 pp.

Bunke, Olaf, Bernd Droge, and Jörg Polzehl (Jan. 1999). “Model Selection, Transformations and Variance Estimation in Nonlinear Regression”. In: *Statistics* 33.3, pp. 197–240.

Burckhardt, Philipp, Daniel S Nagin, and Rema Padman (n.d.). “Multi-Trajectory Models of Chronic Kidney Disease Progression”. In: ().

Burrus, C Sidney (n.d.). “Iterative Reweighted Least Squares”. In: ().

Calinski, T. and J. Harabasz (1974). “A Dendrite Method for Cluster Analysis”. In: *Communications in Statistics - Theory and Methods* 3.1, pp. 1–27.

Cantoni, Eva and Elvezio Ronchetti (Sept. 2001). “Robust Inference for Generalized Linear Models”. In: *Journal of the American Statistical Association* 96.455, pp. 1022–1030.

Carey, Vincent, Scott L. Zeger, and Peter Diggle (1993). “Modelling Multivariate Binary Data with Alternating Logistic Regressions”. In: *Biometrika* 80.3, pp. 517–526.

Caron-Diotte, Mathieu et al. (Feb. 1, 2023). “Handling Planned and Unplanned Missing Data in a Longitudinal Study”. In: *The Quantitative Methods for Psychology* 19.2, pp. 123–135.

Carpenter, James R. and Melanie Smuk (June 2021). “Missing Data: A Statistical Framework for Practice”. In: *Biometrical Journal* 63.5, pp. 915–947.

Carreira-Perpiñán, Miguel Á. and Steve Renals (Jan. 1, 2000). “Practical Identifiability of Finite Mixtures of Multivariate Bernoulli Distributions”. In: *Neural Computation* 12.1, pp. 141–152.

Carrillo-Larco, Rodrigo M. and Manuel Castillo-Cara (June 15, 2020). “Using Country-Level Variables to Classify Countries According to the Number of Confirmed COVID-19 Cases: An Unsupervised Machine Learning Approach”. In: *Wellcome Open Research* 5, p. 56.

Castilla, Elena et al. (Feb. 27, 2020). “Model Selection in a Composite Likelihood Framework Based on Density Power Divergence”. In: *Entropy* 22.3, p. 270.

Celeux, Gilles (Jan. 4, 2019). “EM Methods for Finite Mixtures”. In: *Handbook of Mixture Analysis*. Ed. by Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert. 1st ed. Boca Raton, Florida : CRC Press, [2019]: Chapman and Hall/CRC, pp. 21–39.

Celeux, Gilles, Sylvia Frühwirth-Schnatter, and Christian P. Robert (Jan. 4, 2019). “Model Selection for Mixture Models – Perspectives and Strategies”. In: *Handbook of Mixture Analysis*. Ed. by Sylvia Frühwirth-Schnatter, Gilles Celeux, and Christian P. Robert. 1st ed. Boca Raton, Florida : CRC Press, [2019]: Chapman and Hall/CRC, pp. 117–154.

Chamroukhi, Faicel and Hien D. Nguyen (July 2019). “Model-based Clustering and Classification of Functional Data”. In: *WIREs Data Mining and Knowledge Discovery* 9.4, e1298.

Charpentier, Arthur, Emmanuel Flachaire, and Antoine Ly (n.d.). “Économétrie & Machine Learning”. In: ().

Chauveau, Didier (July 1995). “A Stochastic EM Algorithm for Mixtures with Censored Data”. In: *Journal of Statistical Planning and Inference* 46.1, pp. 1–25.

Chauveau, Didier and Vy Thuy Lynh Hoang (Nov. 2016). “Nonparametric Mixture Models with Conditionally Independent Multivariate Component Densities”. In: *Computational Statistics & Data Analysis* 103, pp. 1–16.

Chavent, Guy (2010). *Nonlinear Least Squares for Inverse Problems: Theoretical Foundations and Step-by-Step Guide for Applications*. Scientific Computation. Dordrecht: Springer Netherlands.

Chen, K., L. Xu, and H. Chi (Nov. 1999). “Improved Learning Algorithms for Mixture of Experts in Multiclass Classification”. In: *Neural Networks* 12.9, pp. 1229–1252.

Chu, Man-Kee M. and John J. Koval (Mar. 16, 2014). “Trajectory Modeling of Longitudinal Binary Data: Application of the EM Algorithm for Mixture Models”. In: *Communications in Statistics - Simulation and Computation* 43.3, pp. 495–519.

Chu, Man-Kee Maggie (n.d.). “Statistical Methods for the Analysis of RNA Sequencing Data”. In: ().

Church, Kenneth W. and William A. Gale (June 1995). “Poisson Mixtures”. In: *Natural Language Engineering* 1.2, pp. 163–190.

Cole, Veronica T. and Daniel J. Bauer (July 3, 2016). “A Note on the Use of Mixture Models for Individual Prediction”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 23.4, pp. 615–631.

Cornillon, Pierre-André and Éric Matzner-Løber (2010). *Régression avec R*. Pratique R. Paris Berlin Heidelberg [etc.]: Springer.

Cox, David R (1972). “The Analysis of Multivariate Binary Data”. In: *Applied statistics*, pp. 113–120.

Cribari-Neto, Francisco and Achim Zeileis (2010). “Beta Regression in *R*”. In: *Journal of Statistical Software* 34.2.

Cuturi, Marco and Mathieu Blondel (n.d.). “Soft-DTW: A Differentiable Loss Function for Time-Series”. In: ().

Dali, Houcine Ben, Séverin Benzoni, and Joseph Lehec (n.d.). “La conjecture de corrélation gaussienne”. In: ().

Daniels, Michael J and Joseph W Hogan (n.d.). “Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis”. In: ().

Daskalakis, C. (Jan. 1, 2002). “Regression Analysis of Multiple-Source Longitudinal Outcomes: A "Stirling County" Depression Study”. In: *American Journal of Epidemiology* 155.1, pp. 88–94.

Davis, Charles Shaw (2003). *Statistical Methods for the Analysis of Repeated Measurements*. Corr. print. Springer Texts in Statistics. New York: Springer. 415 pp.

Day, David M, Irene Bevc, et al. (n.d.). “Comparison of Adult Offense Prediction Methods Based on Juvenile Offense Trajectories Using Cross-Validation”. In: ().

Day, David M, Jason D Nielsen, et al. (2011). “Trajectories of Criminal Activity in a Sample of 378 Adjudicated Ontario Youth”. In.

De Brito Trindade, Daniele, Raydonal Ospina, and Leila D. Amorim (Apr. 26, 2022). “Choosing the Right Strategy to Model Longitudinal Count Data in Epidemiology: An Application with CD4 Cell Counts”. In: *Epidemiology, Biostatistics, and Public Health* 12.4.

De Souza, Fernanda Sumika Hojo et al. (Apr. 2021). “Second Wave of COVID-19 in Brazil: Younger at Higher Risk”. In: *European Journal of Epidemiology* 36.4, pp. 441–443.

Dempster, A. P., N. M. Laird, and D. B. Rubin (Sept. 1977). “Maximum Likelihood from Incomplete Data Via the *EM* Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

“Description of the Algorithm Used by Mplus” (n.d.). In: ().

Desgraupes, Bernard and Maintainer Bernard Desgraupes (2018). *Package 'clusterCrit'.* R-Proj.

De Souza, Fernanda Sumika Hojo et al. (2021). “Second Wave of COVID-19 in Brazil: Younger at Higher Risk”. In: *European journal of epidemiology* 36, pp. 441–443.

Deville, Yannick (Apr. 2019). “From Separability/Identifiability Properties of Bilinear and Linear-Quadratic Mixture Matrix Factorization to Factorization Algorithms”. In: *Digital Signal Processing* 87, pp. 21–33.

Devogele, Thomas et al. (Nov. 7, 2017). “Optimized Discrete Fréchet Distance between Trajectories”. In: *Proceedings of the 6th ACM SIGSPATIAL Workshop on Analytics for Big Geospatial Data.* SIGSPATIAL'17: 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Redondo Beach CA USA: ACM, pp. 11–19.

Di, Chong-Zhi and Karen Bandeen-Roche (Mar. 2011). “Multilevel Latent Class Models with Dirichlet Mixing Distribution”. In: *Biometrics* 67.1, pp. 86–96.

Di Zio, Marco, Ugo Guarnera, and Roberto Rocci (Feb. 2007). “A Mixture of Mixture Models for a Classification Problem: The Unity Measure Error”. In: *Computational Statistics & Data Analysis* 51.5, pp. 2573–2585.

Dietz, Laura (n.d.). “Directed Factor Graph Notation for Generative Models”. In: ().

Diggle, Peter and Peter Diggle, eds. (2002). *Analysis of Longitudinal Data.* 2nd ed. Oxford Statistical Science Series 25. Oxford ; New York: Oxford University Press. 379 pp.

Dijk, Abram van (2009). *Essays on Finite Mixture Models: = Essays over Finite Mixture Modellen.* Tinbergen Institute Research Series 458. Amsterdam: Thela Thesis. 126 pp.

Ding, Ming, Jorge E. Chavarro, and Garrett M. Fitzmaurice (Dec. 18, 2019). *Development of a Mixture Model (SMM) Allowing for Smoothing Functions of Trajectories.* preprint. Epidemiology.

Dodge, Hiroko H, Changyu Shen, and Mary Ganguli (n.d.). “Application of the Pattern-Mixture Latent Trajectory Model in an Epidemiological Study with Non-Ignorable Missingness”. In: ().

Drton, Mathias, Rina Foygel, and Seth Sullivant (Apr. 1, 2011). “Global Identifiability of Linear Structural Equation Models”. In: *The Annals of Statistics* 39.2. arXiv: `1003.1146 [math, stat]`.

Duan, Jin-Chuan and Andras Fulop (n.d.). “A Stable Estimator for the Information Matrix under EM”. In: ().

Duncan, Terry E. and Susan C. Duncan (Dec. 2009). “The ABC’s of LGM: An Introductory Guide to Latent Variable Growth Curve Modeling: The ABC’s of LGM”. In: *Social and Personality Psychology Compass* 3.6, pp. 979–991.

“Econometric Analysis of Cross Section and Panel Data” (n.d.). In: ().

Eddelbuettel, Dirk (2013). *Seamless R and C++ Integration with Rcpp*. New York, NY: Springer New York.

Eddelbuettel, Dirk and James Joseph Balamuta (n.d.). “Extending R with C++: A Brief Introduction to Rcpp”. In: ().

Eddelbuettel, Dirk and Romain François (2011). “Rcpp: Seamless R and C++ Integration”. In: *Journal of statistical software* 40, pp. 1–18.

Edwards, D. and S. L. Lauritzen (Dec. 1, 2001). “The TM Algorithm for Maximising a Conditional Likelihood Function”. In: *Biometrika* 88.4, pp. 961–972.

Eiter, Thomas and Heikki Mannila (n.d.). “Computing Discrete Fr´echet Distance”. In: ().

Elmer, Jonathan, Bobby L. Jones, and Daniel S. Nagin (Dec. 2018). “Using the Beta Distribution in Group-Based Trajectory Models”. In: *BMC Medical Research Methodology* 18.1, p. 152.

Elmer, Jonathan, Bobby L. Jones, Vladimir I. Zadorozhny, et al. (Apr. 2019). “A Novel Methodological Framework for Multimodality, Trajectory Model-Based Prognostication”. In: *Resuscitation* 137, pp. 197–204.

Elshiewy, Ossama, Daniel Guhl, and Yasemin Boztug (2017). “Multinomial Logit Models in Marketing - From Fundamentals to State-of-the-Art”. In: *Marketing ZFP* 39.3, pp. 32–49.

Enders, Craig K. and Todd D. Little (2010). *Applied Missing Data Analysis*. Methodology in the Social Sciences. New York, N.Y.: Guilford Press. 377 pp.

Epstein, David and Jarmila Curtiss (n.d.). “Identifying Differences in Capital Growth Trajectories of Agricultural Enterprises in Russia”. In: ().

Erro, Roberto et al. (Aug. 1, 2013). “The Heterogeneity of Early Parkinson’s Disease: A Cluster Analysis on Newly Diagnosed Untreated Patients”. In: *PLoS ONE* 8.8. Ed. by Jeff A. Beeler, e70244.

Espinheira, Patrícia L., Evelyne G. Santos, and Francisco Cribari-Neto (May 2017). “On Nonlinear Beta Regression Residuals: On Nonlinear Beta Regression Residuals”. In: *Biometrical Journal* 59.3, pp. 445–461.

Etienne, Laurent and Thomas Devogele (n.d.). “Trajectoires médianes”. In: ().

Etienne, Laurent, Thomas Devogele, et al. (May 3, 2016). “Trajectory Box Plot: A New Pattern to Summarize Movements”. In: *International Journal of Geographical Information Science* 30.5, pp. 835–853.

Everitt, Brian (n.d.). “Cluster Analysis”. In: ().

Faraway, Julian J (n.d.[a]). “Extending the Linear Model with R”. In: ().

— (n.d.[b]). “Linear Models with R”. In: ().

Faria, Susana and Gilda Soromenho (Feb. 2010). “Fitting Mixtures of Linear Regressions”. In: *Journal of Statistical Computation and Simulation* 80.2, pp. 201–225.

— (Apr. 2012). “Comparison of EM and SEM Algorithms in Poisson Regression Models: A Simulation Study”. In: *Communications in Statistics - Simulation and Computation* 41.4, pp. 497–509.

Farrington, David P and Donald J West (1990). *The Cambridge Study in Delinquent Development: A Long-Term Follow-up of 411 London Males.* Springer.

Feller, W (2021). “On a General Class of "Contagious" Distributions”. In.

Feller, William (n.d.). “To .Probability Theory and Its Applications”. In: ().

Feng, Jiarui and Zhongyi Zhu (Jan. 2011). “Semiparametric Analysis of Longitudinal Zero-Inflated Count Data”. In: *Journal of Multivariate Analysis* 102.1, pp. 61–72.

Fereshtehnejad, Seyed-Mohammad et al. (Aug. 1, 2015). “New Clinical Subtypes of Parkinson Disease and Their Longitudinal Progression: A Prospective Cohort Comparison With Other Phenotypes”. In: *JAMA Neurology* 72.8, p. 863.

Ferrari, Silvia and Francisco Cribari-Neto (Aug. 2004). “Beta Regression for Modelling Rates and Proportions”. In: *Journal of Applied Statistics* 31.7, pp. 799–815.

Fieuws, Steffen and Geert Verbeke (June 2006). “Pairwise Fitting of Mixed Models for the Joint Modeling of Multivariate Longitudinal Profiles”. In: *Biometrics* 62.2, pp. 424–431.

Fischer, Aurélie (2014). “Deux méthodes d’apprentissage non supervisé : synthèse sur la méthode des centres mobiles et présentation des courbes principales”. In: 155.2.

Fitzmaurice, Garrett M., ed. (2009). *Longitudinal Data Analysis.* Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Boca Raton: CRC Press. 618 pp.

Fletcher, Roger (1970). “A New Approach to Variable Metric Algorithms”. In: *The computer journal* 13.3, pp. 317–322.

Follmann, Dean A. and Diane Lambert (Mar. 1991). “Identifiability of Finite Mixtures of Logistic Regression Models”. In: *Journal of Statistical Planning and Inference* 27.3, pp. 375–381.

Foulley, Jean-Louis (n.d.[a]). “Algorithme EM : théorie et application au modèle mixte”. In: ().

— (n.d.[b]). “LE MODELE LINEAIRE MIXTE”. In: ().

Fox, John (n.d.). “Nonlinear Regression and Nonlinear Least Squares”. In: ().

Frees, Edward W (2004). “Longitudinal and Panel Data: Analysis and Applications in the Social Sciences”. In.

Frühwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P. Robert, eds. (2019). *Handbook of Mixture Analysis*. Boca Raton: CRC Press, Taylor and Francis Group. 1 p.

Gaffney, Scott and Padhraic Smyth (Aug. 1999). “Trajectory Clustering with Mixtures of Regression Models”. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD99: The First Annual International Conference on Knowledge Discovery in Data. San Diego California USA: ACM, pp. 63–72.

Gajardo, Karla Andrea Munoz (n.d.). “AN EXTENSION OF THE NORMAL CENSORED REGRESSION MODEL. ESTIMATION AND APPLICATIONS”. In: ().

Gałecki, Andrzej and Tomasz Burzykowski (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer Texts in Statistics. New York, NY: Springer New York.

Gallant, A. Ronald (1987). *Nonlinear Statistical Models*. Wiley Series in Probability and Mathematical Statistics. New York: Wiley. 610 pp.

Galwey, Nick (2006). *Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance*. Chichester, England ; Hoboken, NJ: Wiley. 366 pp.

Gao, Xin (n.d.). “COMPOSITE LIKELIHOOD EM ALGORITHM WITH APPLICATIONS TO MULTIVARIATE HIDDEN MARKOV MODEL”. In: ().

Garet, Olivier (n.d.). “Probabilités et Processus Stochastiques”. In: ().

Gavin, Henri P (2022). “The Levenberg-Marquardt Algorithm for Nonlinear Least Squares Curve-Fitting Problems”. In.

Gebru, Israel D. et al. (Dec. 1, 2016). “EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.12, pp. 2402–2415. arXiv: `1509.01509 [cs, stat]`.

Genolini, Christophe, Xavier Alacoque, et al. (2015). “**Kml** and **Kml3d** : *R* Packages to Cluster Longitudinal Data”. In: *Journal of Statistical Software* 65.4.

Genolini, Christophe, René Ecochard, et al. (June 3, 2016). “kmlShape: An Efficient Method to Cluster Longitudinal Data (Time-Series) According to Their Shapes”. In: *PLOS ONE* 11.6. Ed. by Chun-Hsi Huang, e0150738.

Genolini, Christophe and Bruno Falissard (June 2010). “KmL: K-Means for Longitudinal Data”. In: *Computational Statistics* 25.2, pp. 317–328.

Gentle, James E. (2009). *Computational Statistics*. Statistics and Computing. New York, NY: Springer New York.

Geys, Helena, Geert Molenberghs, and Louise M. Ryan (Sept. 1999). “Pseudolikelihood Modeling of Multivariate Outcomes in Developmental Toxicology”. In: *Journal of the American Statistical Association* 94.447, pp. 734–745.

Ghosh, Subir and Hiya Banerjee (Dec. 2010). “Methods of Finding the Initial Values of Parameters in the Maximum Likelihood Estimating Equations for a Logistic Regression Model and Comparison of Their Final Solutions Using Different Criterion Functions”. In: *Journal of the Korean Statistical Society* 39.4, pp. 471–477.

Giguère, Charles-Édouard (n.d.). “Modèle de mélange de lois multinormales appliqué à l’analyse de comportements et d’habiletés cognitives d’enfants”. In: ().

Gill, Jeff and Steven Heeringa (n.d.). “Statistics in the Social and Behavioral Sciences Series”. In: ().

Gill, Jeff and Gary King (Aug. 2004). “What to Do When Your Hessian Is Not Invertible: Alternatives to Model Respecification in Nonlinear Estimation”. In: *Sociological Methods & Research* 33.1, pp. 54–87.

Goldfarb, Donald (1970). “A Family of Variable-Metric Methods Derived by Variational Means”. In: *Mathematics of computation* 24.109, pp. 23–26.

Gonzalez, Teofilo F. (1985). “Clustering to Minimize the Maximum Intercluster Distance”. In: *Theoretical Computer Science* 38, pp. 293–306.

Goodman, Leo A. (Aug. 2007). “1. On the Assignment of Individuals to Latent Classes”. In: *Sociological Methodology* 37.1, pp. 1–22.

Gormley, Isobel Claire and Sylvia Frühwirth-Schnatter (June 21, 2018). *Mixtures of Experts Models*. arXiv: `1806.08200 [stat]`. URL: `http://arxiv.org/abs/1806.08200` (visited on 09/19/2023). Pre-published.

Gorshenin, A. K., V. Yu. Korolev, and A. M. Tursunbaev (Nov. 2017). “Median Modifications of the EM-Algorithm for Separation of Mixtures of Probability Distributions and Their Applications to the Decomposition of Volatility of Financial Indexes*”. In: *Journal of Mathematical Sciences* 227.2, pp. 176–195.

Gray, Laura A and Monica Hernandez-Alava (n.d.). “BETAMIX: A Command for Fitting Mixture Regression Models for Bounded Dependent Variables Using the Beta Distribution”. In: ().

Greenberg, David F. (Mar. 2016). “Criminal Careers: Discrete or Continuous?” In: *Journal of Developmental and Life-Course Criminology* 2.1, pp. 5–44.

Greene, William (n.d.). “Preliminary. Comments Invited”. In: ().

Greene, William H. (2003). *Econometric Analysis*. 5th ed. Upper Saddle River, N.J: Prentice Hall. 1026 pp.

Gregoriou, Greg N. and Razvan Pascalau, eds. (2011). *Nonlinear Financial Econometrics: Markov Switching Models, Persistence and Nonlinear Cointegration*. London: Palgrave Macmillan UK.

Grosse, Roger and Nitish Srivastava (n.d.). “Lecture 16: Mixture Models”. In: ().

Grün, Bettina and Kurt Hornik (Mar. 1, 2012). “Modelling Human Immunodeficiency Virus Ribonucleic Acid Levels with Finite Mixtures for Censored Longitudinal Data”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 61.2, pp. 201–218.

Grün, Bettina, Ioannis Kosmidis, and Achim Zeileis (2012). “Extended Beta Regression in *R* : Shaken, Stirred, Mixed, and Partitioned”. In: *Journal of Statistical Software* 48.11.

Grün, Bettina and Friedrich Leisch (2008a). “Finite Mixtures of Generalized Linear Regression Models”. In: Shalabh and Christian Heumann. *Recent Advances in Linear Models and Related Areas*. Heidelberg: Physica-Verlag HD, pp. 205–230.

— (Nov. 2008b). “Identifiability of Finite Mixtures of Multinomial Logit Models with Varying and Fixed Effects”. In: *Journal of Classification* 25.2, pp. 225–247.

— (n.d.). “Finite Mixture Model Diagnostics Using Resampling Methods”. In: ().

Guigou, Jean-Daniel, Bruno Lovat, and Jang Schiltz (Nov. 2012). “Optimal Mix of Funded and Unfunded Pension Systems: The Case of Luxembourg”. In: *Pensions: An International Journal* 17.4, pp. 208–222.

Guillaume, Saint Pierre (n.d.). “Mélanges Gaussiens”. In: ().

Gyllenberg, Mats et al. (1994). “Non-Uniqueness in Probabilistic Numerical Identification of Bacteria”. In: *Journal of Applied Probability* 31.2, pp. 542–548.

El-Habil, Abdalla M (Mar. 28, 2012). “An Application on Multinomial Logistic Regression Model”. In: *Pakistan Journal of Statistics and Operation Research* 8.2, p. 271.

Hadfield, Jarrod D. (2010). “MCMC Methods for Multi-Response Generalized Linear Mixed Models: The **MCMCglmm** *R* Package”. In: *Journal of Statistical Software* 33.2.

Hall, Daniel B. (Dec. 2000). “Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study”. In: *Biometrics* 56.4, pp. 1030–1039.

Hall, Daniel B. and Jing Shen (Sept. 27, 2009). “Robust Estimation for Zero-Inflated Poisson Regression: Robust ZIP Regression”. In: *Scandinavian Journal of Statistics* 37.2, pp. 237–252.

Hamedi-Shahraki, Soudabeh et al. (Mar. 2, 2021). “Kumaraswamy Regression Modeling for Bounded Outcome Scores”. In: *Pakistan Journal of Statistics and Operation Research*, pp. 79–88.

Hansen, Niels Richard (n.d.). “Aspects of Algebraic Statistics”. In: ().

Harden, Simon (n.d.). “Weighted Composite Likelihoods”. In: ().

Hasan, Asad, Zhiyu Wang, and Alireza S. Mahani (2016). “Fast Estimation of Multinomial Logit Models: *R* Package **Mnlogit**”. In: *Journal of Statistical Software* 75.3.

Hasell, Joe et al. (Oct. 8, 2020). “A Cross-Country Database of COVID-19 Testing”. In: *Scientific Data* 7.1, p. 345.

Haviland, Amelia M., Bobby L. Jones, and Daniel S. Nagin (May 2011). “Group-Based Trajectory Modeling Extended to Account for Nonrandom Participant Attrition”. In: *Sociological Methods & Research* 40.2, pp. 367–390.

Henderson, R. (June 1, 2003). “A Serially Correlated Gamma Frailty Model for Longitudinal Count Data”. In: *Biometrika* 90.2, pp. 355–366.

Henna, Jogi (1994). *Examples of Identifiable Mixture.* Japan Statistical Society: 2. URL: https://doi.org/10.11329/jjss1970.24.193 (visited on 09/19/2023). Pre-published.

Hennig, C. (July 1, 2000). “Identifiablity of Models for Clusterwise Linear Regression”. In: *Journal of Classification* 17.2, pp. 273–296.

Henningsen, Arne (n.d.). “Estimating Censored Regression Models in R Using the censReg Package”. In: ().

Herby, Jonas, Lars Jonung, and Steve H Hanke (n.d.). “A Literature Review and Meta-Analysis of the Effects of Lockdowns on COVID-19 Mortality”. In: ().

Hickson, Ryan P. et al. (Mar. 2020). “Opening the Black Box of the Group-based Trajectory Modeling Process to Analyze Medication Adherence Patterns: An Example Using Real-world Statin Adherence Data”. In: *Pharmacoepidemiology and Drug Safety* 29.3, pp. 357–362.

Hipp, John R. and Daniel J. Bauer (2006). “Local Solutions in the Estimation of Growth Mixture Models.” In: *Psychological Methods* 11.1, pp. 36–53.

Holzmann, Hajo, Axel Munk, and Bernd Stratmann (2004). “Identifiability of Finite Mixtures - with Applications to Circular Distributions”. In: *Sankhya* 66, pp. 440–449.

Houssou, Noudéhouénou, Jean-Loup Guillaume, and Armelle Prigent (n.d.). “Review and Comparison of Similarity Measures and Community Detection Algorithms for Clustering of Network Constrained Trajectories”. In: ().

Huet, S. (2004). *Statistical Tools for Nonlinear Regression: A Practical Guide with S-PLUS and R Examples*. 2nd ed. Springer Series in Statistics. New York: Springer. 232 pp.

Hui, Francis K. C. (Feb. 2016). “MIXING IT UP: NEW METHODS FOR FINITE MIXTURE MODELLING OF MULTI-SPECIES DATA IN ECOLOGY”. In: *Bulletin of the Australian Mathematical Society* 93.1, pp. 167–168.

Iliopoulos, Costas S. and William F. Smyth, eds. (2011). *Combinatorial Algorithms*. Vol. 7056. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg.

Imkamp, Maike et al. (Nov. 2018). “Uncovering the Heterogeneity of Disease Impact in Axial Spondyloarthritis: Bivariate Trajectories of Disease Activity and Quality of Life”. In: *RMD Open* 4.2, e000755.

Inouye, David I. et al. (May 2017). “A Review of Multivariate Distributions for Count Data Derived from the Poisson Distribution”. In: *WIREs Computational Statistics* 9.3.

“Institut National de Recherche En Informatique et En Automatique” (Dec. 1992). In: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 37.1, pp. 55–57.

“Institut National de Recherche En Informatique et En Automatique” (Dec. 1992). In: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 37.1, pp. 55–57.

Irizarry, Rafael A and Michael I Love (n.d.). “Data Analysis for the Life Sciences”. In: ().

Jank, Wolfgang (Dec. 2006). “Implementing and Diagnosing the Stochastic Approximation EM Algorithm”. In: *Journal of Computational and Graphical Statistics* 15.4, pp. 803–829.

Jannes, Gil and Jesús Barreal (June 29, 2020). *Beta Regression with Spatio-Temporal Effects as a Tool for Hospital Impact Analysis of Initial Phase Epidemics: The Case of COVID-19 in Spain*. preprint. Health Economics.

Jiang, Jianfei (n.d.). “GROUP-BASED TRAJECTORY MODELING FOR LONGITUDINAL DATA OF HEALTHCARE FINANCIAL CHARGES IN PATIENTS WITH INFLAMMATORY BOWEL DISEASE”. In: ().

Jiang, W. and M.A. Tanner (Nov. 1999). “On the Identifiability of Mixtures-of-Experts”. In: *Neural Networks* 12.9, pp. 1253–1258.

Jin, Zi (n.d.). “Aspects of Composite Likelihood Inference”. In: ().

Jóhannesson, Benedikt and Narayan Giri (Jan. 1995). “On Approximations Involving the Beta Distribution”. In: *Communications in Statistics - Simulation and Computation* 24.2, pp. 489–503.

Jones, Bobby L and Daniel S Nagin (n.d.). “A Stata Plugin for Estimating Group-Based Trajectory Models”. In: ().

— (May 2007). “Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them”. In: *Sociological Methods & Research* 35.4, pp. 542–571.

— (Nov. 2013). “A Note on a Stata Plugin for Estimating Group-based Trajectory Models”. In: *Sociological Methods & Research* 42.4, pp. 608–613.

Jones, Bobby L., Daniel S. Nagin, and Kathryn Roeder (Feb. 2001). “A SAS Procedure Based on Mixture Models for Estimating Developmental Trajectories”. In: *Sociological Methods & Research* 29.3, pp. 374–393.

Jones-White, Daniel R. et al. (Mar. 2010). “Redefining Student Success: Applying Different Multinomial Regression Techniques for the Study of Student Graduation Across Institutions of Higher Education”. In: *Research in Higher Education* 51.2, pp. 154–174.

Jørgensen, Bent et al. (Sept. 1996). “State-Space Models for Multivariate Longitudinal Data of Mixed Types”. In: *Canadian Journal of Statistics* 24.3, pp. 385–402.

Jr, Frank E Harrell (n.d.). “Regression Modeling Strategies”. In: ().

Juhel, Jacques (2014). “La recherche d’invariants différentiels dans les variations développementales : de la population à l’individu et réciproquement !” In: *Développement et variabilités*. Ed. by Sandrine Le Sourn-Bissaoui et al. Presses universitaires de Rennes, pp. 13–41.

Jung, Tony and K. A. S. Wickrama (Jan. 2008). “An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling: Latent Trajectory Classes”. In: *Social and Personality Psychology Compass* 2.1, pp. 302–317.

Kaciroti, Niko A. et al. (Dec. 1, 2008). “A Bayesian Model for Longitudinal Count Data with Non-Ignorable Dropout”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 57.5, pp. 521–534.

Kaptein, Maurits and Paul Ketelaar (2018a). “Maximum Likelihood Estimation of a Finite Mixture of Logistic Regression Models in a Continuous Data Stream”. Version 1. In.

— (Feb. 2018b). *Maximum Likelihood Estimation of a Finite Mixture of Logistic Regression Models in a Continuous Data Stream*. arXiv.

Karagrigoriou, Alex, Teresa Oliveira, and Christos H Skiadas (n.d.). “Statistical, Stochastic and Data Analysis Methods and Applications”. In: ().

Karlis, Dimitris and Loukia Meligkotsidou (June 2007). “Finite Mixtures of Multivariate Poisson Distributions with Application”. In: *Journal of Statistical Planning and Inference* 137.6, pp. 1942–1960.

Karlis, Dimitris and Evdokia Xekalaki (n.d.). “Choosing Initial Values for the EM Algorithm for Ÿnite Mixtures”. In: ().

Kass, Robert E and Larry Wasserman (1995). “A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion”. In: *Journal of the american statistical association* 90.431, pp. 928–934.

Kelava, Augustin, Benjamin Nagengast, and Holger Brandt (July 3, 2014). “A Nonlinear Structural Equation Mixture Modeling Approach for Nonnormally Distributed Latent Predictor Variables”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 21.3, pp. 468–481.

Khalili, Abbas and Jiahua Chen (Sept. 2007). “Variable Selection in Finite Mixture of Regression Models”. In: *Journal of the American Statistical Association* 102.479, pp. 1025–1038.

Khoshaein, Vafa (n.d.). “Trajectory Clustering Using a Variation of Fre´chet Distance”. In: ().

Kim, Daeyoung and Bruce G. Lindsay (Aug. 2015). “Empirical Identifiability in Finite Mixture Models”. In: *Annals of the Institute of Statistical Mathematics* 67.4, pp. 745–772.

Kinnunen, Jani et al. (2021). “Dynamic Indexing and Clustering of Government Strategies to Mitigate Covid-19”. In: *Entrepreneurial Business and Economics Review* 9.2, pp. 7–20.

Kleiber, Christian and Achim Zeileis (2008). *Applied Econometrics with R*. New York, NY: Springer New York.

Klein, John P. and Melvin L. Moeschberger (2005). *Survival Analysis: Techniques for Censored and Truncated Data*. Second edition, corrected third printing. Statistics for Biology and Health. New York Berlin Heidelberg: Springer. 536 pp.

Kleinbaum, David G. and Mitchel Klein (2010). *Logistic Regression*. Statistics for Biology and Health. New York, NY: Springer New York.

Klijn, Sven L et al. (Oct. 2017). “Introducing the Fit-Criteria Assessment Plot – A Visualisation Tool to Assist Class Enumeration in Group-Based Trajectory Modelling”. In: *Statistical Methods in Medical Research* 26.5, pp. 2424–2436.

Klonecki, Witold (Apr. 1970). “Mixtures and Characteristic Functions”. In: *Proceedings of the National Academy of Sciences* 65.4, pp. 831–836.

Kopciuszewski, Pawel (Jan. 2004). “An Extension of the Factorization Theorem to the Non-Positive Case”. In: *Journal of Multivariate Analysis* 88.1, pp. 118–130.

Kruskal, Joseph B. (1977). “Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics”. In: *Linear Algebra and its Applications* 18.2, pp. 95–138.

Kuo, Chien-Wen Jean (n.d.). “ANALYZING TRAJECTORIES OF CAREGIVER PSYCHOLOGICAL DISTRESS OVER TIME USING GROUP-BASED MODELING METHODS”. In: ().

Labouriau, Rodrigo (May 4, 2014). *A Note on the Identifiability of Generalized Linear Mixed Models*. arXiv: `1405.0673 [stat]`. URL: `http://arxiv.org/abs/1405.0673` (visited on 09/19/2023). Pre-published.

Labunets, V. and E. Osthaimer (2017). “Systematic Approach to Nonlinear Filtering Associated with Aggregation Operators. Part 2. Fréchet MIMO-filters”. In: *Procedia Engineering* 201, pp. 385–397.

Lachos, Víctor Hugo, Luis Benites Sanchez, and Celso Rômulo Barbosa Cabral (n.d.). “Robust Regression Modeling for Censored Data Based on Mixtures of Student-t Distributions”. In: ().

Lai, Dongbing et al. (Oct. 25, 2016). “A Multivariate Finite Mixture Latent Trajectory Model with Application to Dementia Studies”. In: *Journal of Applied Statistics* 43.14, pp. 2503–2523.

Lambert, Diane (Feb. 1992). “Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing”. In: *Technometrics* 34.1, p. 1. JSTOR: `1269547`.

Lange, Kenneth (n.d.). “A QUASI-NEWTON ACCELERATION OF THE EM ALGORITHM”. In: ().

Lavielle, Marc and Leon Aarons (Feb. 2016). “What Do We Mean by Identifiability in Mixed Effects Models?” In: *Journal of Pharmacokinetics and Pharmacodynamics* 43.1, pp. 111–122.

Lee, Andy H. et al. (Feb. 2006). “Multi-Level Zero-Inflated Poisson Regression Modelling of Correlated Count Data with Excess Zeros”. In: *Statistical Methods in Medical Research* 15.1, pp. 47–61.

Lee, Do Q (n.d.). “NUMERICALLY EFFICIENT METHODS FOR SOLVING LEAST SQUARES PROBLEMS”. In: ().

Lee, Gyemin and Clayton Scott (Sept. 2012). “EM Algorithms for Multivariate Gaussian Mixture Models with Truncated and Censored Data”. In: *Computational Statistics & Data Analysis* 56.9, pp. 2816–2829.

Lee, Jae-Gil, Jiawei Han, and Kyu-Young Whang (n.d.). “Trajectory Clustering: A Partition-and-Group Framework”. In: ().

Leisch, Friedrich (2004). “FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in *R*”. In: *Journal of Statistical Software* 11.8.

— (n.d.). “Creating R Packages: A Tutorial”. In: ().

Leveque, Emilie (n.d.). “Modélisation statistique de l’intensité des expositions prolongées en étiologie du cancer: application au tabac, à l’amiante, au cancer du poumon, et au mésothéliome pleural”. In: ().

Lewis, Robert H., Bela Paláncz, and Joseph Awange (Aug. 14, 2015). "Application of Dixon Resultant to Maximization of the Likelihood Function of Gaussian Mixture Distribution". In: *ACM Communications in Computer Algebra* 49.2, pp. 57–57.

Li, Chin-Shang (Aug. 1, 2012). "Identifiability of Zero-Inflated Poisson Models". In: *Brazilian Journal of Probability and Statistics* 26.3.

Li, Chin-Shang et al. (Feb. 1999). "Multivariate Zero-Inflated Poisson Models and Their Applications". In: *Technometrics* 41.1, pp. 29–38.

Li, Gen (Nov. 21, 2018). "Application of Finite Mixture of Logistic Regression for Heterogeneous Merging Behavior Analysis". In: *Journal of Advanced Transportation* 2018, pp. 1–9.

Li, Haifeng, Keshu Zhang, and Tao Jiang (n.d.). "The Regularized EM Algorithm". In: ().

Li, Ming, Jeffrey R Harring, and George B Macready (May 1, 2014). "Investigating the Feasibility of Using Mplus in the Estimation of Growth Mixture Models". In: *Journal of Modern Applied Statistical Methods* 13.1, pp. 484–513.

Lim, Hwa Kyung, Wai Keung Li, and Philip L H Yu (n.d.). "Zero-Inflated Poisson Regression Mixture Model". In: ().

Lima Passos, Valéria et al. (Apr. 2017). "At the Heart of the Problem - A Person-Centred, Developmental Perspective on the Link between Alcohol Consumption and Cardio-Vascular Events". In: *International Journal of Cardiology* 232, pp. 304–314.

Lin, Haiqun, Charles E. McCulloch, and Robert A. Rosenheck (June 2004). "Latent Pattern Mixture Models for Informative Intermittent Missing Data in Longitudinal Studies". In: *Biometrics* 60.2, pp. 295–305.

Lindsay, Bruce G. (1988). "Composite Likelihood Methods". In: *Contemporary Mathematics.* Ed. by N. U. Prabhu. Vol. 80. Providence, Rhode Island: American Mathematical Society, pp. 221–239.

— (1995). *Mixture Models: Theory, Geometry, and Applications.* NSF-CBMS Regional Conference Series in Probability and Statistics v. 5. Hayward, Calif. : Alexandria, Va: Institute of Mathematical Statistics ; American Statistical Association. 163 pp.

Lipovetsky, Stan (2009). "Numerical Recipes: The Art of Scientific Computing". In: *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences* 51.4, p. 481.

Liu, Fangfang (2015). "Statistical Methods in Detecting Differential Expressed Genes, Analyzing Insertion Tolerance for Genes and Group Selection for Survival Data". Doctor of Philosophy. Ames: Iowa State University, Digital Repository, p. 7986485.

López-Oriona, Ángel and José A. Vilar (Dec. 2021). “Quantile Cross-Spectral Density: A Novel and Effective Tool for Clustering Multivariate Time Series”. In: *Expert Systems with Applications* 185, p. 115677.

Louis, Thomas A. (Jan. 1982). “Finding the Observed Information Matrix When Using the *EM* Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 226–233.

Loum, Mor Absa (n.d.). “Mod‘ele de M´elange et Mod‘eles Lin´eaires G´en´eralis´es, Application Aux Donn´ees de Co-Infection (Arbovirus & Paludisme)”. In: ().

Lourakis, Manolis IA et al. (2005). “A Brief Description of the Levenberg-Marquardt Algorithm Implemented by Levmar”. In: *Foundation of Research and Technology* 4.1, pp. 1–6.

Lu, Chenguang (n.d.). “From the EM Algorithm to the CM-EM Algorithm for Global Convergence of Mixture Models”. In: ().

M. Gad, Ahmed and Rasha B. El Kholy (Aug. 31, 2012). “Generalized Linear Mixed Models for Longitudinal Data”. In: *International Journal of Probability and Statistics* 1.3, pp. 41–47.

Madsen, Kaj, Hans Bruun Nielsen, and Ole Tingleff (2004). “Methods for Non-Linear Least Squares Problems”. In.

Mahlknecht, Philipp, Klaus Seppi, and Werner Poewe (Oct. 17, 2015). “The Concept of Prodromal Parkinson’s Disease”. In: *Journal of Parkinson’s Disease* 5.4, pp. 681–697.

Marquardt, Donald W (1963). “An Algorithm for Least-Squares Estimation of Nonlinear Parameters”. In: *Journal of the society for Industrial and Applied Mathematics* 11.2, pp. 431–441.

Marron, Megan M (n.d.). “CROSS-VALIDATION IN GROUP-BASED LATENT TRAJECTORY MODELING WHEN ASSUMING A CENSORED NORMAL MODEL”. In: ().

Martin, Daniel P. and Timo Von Oertzen (Apr. 3, 2015). “Growth Mixture Models Outperform Simpler Clustering Algorithms When Detecting Longitudinal Heterogeneity, Even With Small Sample Sizes”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 22.2, pp. 264–275.

Mathieu, Edouard et al. (May 10, 2021). “A Global Database of COVID-19 Vaccinations”. In: *Nature Human Behaviour* 5.7, pp. 947–953.

Mbogning, Cyprien (n.d.). “Inférence dans les modèles conjoints et de mélange non-linéaires à effets mixtes”. In: ().

McCullagh, P. and John A. Nelder (1998). *Generalized Linear Models*. 2nd ed. Monographs on Statistics and Applied Probability 37. Boca Raton: Chapman & Hall/CRC. 511 pp.

McCulloch, Charles (Feb. 2008). “Joint Modelling of Mixed Outcome Types Using Latent Variables”. In: *Statistical Methods in Medical Research* 17.1, pp. 53–73.

McKnight, Patrick E., ed. (2007). *Missing Data: A Gentle Introduction*. Methodology in the Social Sciences. New York: Guilford Press. 251 pp.

McLachlan, Geoffrey (2004). *Finite Mixture Models*. Hoboken: John Wiley & Sons.

McLaughlin, Michael P (n.d.). "Compendium of Common Probability Distributions". In: ().

Meeker, William Q. and Luis A. Escobar (Feb. 1995). "Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation". In: *The American Statistician* 49.1, p. 48. JSTOR: `2684811`.

Melnykov, Volodymyr and Ranjan Maitra (Jan. 1, 2010). "Finite Mixture Models and Model-Based Clustering". In: *Statistics Surveys* 4 (none).

Menard, Scott (2008). *Handbook of Longitudinal Research: Design, Measurement, and Analysis*. Burlington, MA: Academic Press - Elsevier.

Meng, Lingyao (Aug. 4, 2016). *Method for Computation of the Fisher Information Matrix in the Expectation-Maximization Algorithm*. arXiv: `1608.01734 [stat]`. URL: `http://arxiv.org/abs/1608.01734` (visited on 09/29/2023). Pre-published.

Milligan, Glenn W and Martha C Cooper (n.d.). "An Examination of Procedures for Determining the Number of Clusters in a Data Set". In: ().

Mitnik, Pablo A. and Sunyoung Baek (Feb. 2013). "The Kumaraswamy Distribution: Median-Dispersion Re-Parameterizations for Regression Modeling and Simulation-Based Estimation". In: *Statistical Papers* 54.1, pp. 177–192.

Mitra, Debanjan (n.d.). "LIKELIHOOD INFERENCE FOR LEFT TRUNCATED AND RIGHT CENSORED LIFETIME DATA". In: ().

"Mixture Model, Longitudinal Trajectory, PROC TRAJ, Quasi-Newton, EM Algorithm" (2015). In: *American Journal of Mathematics and Statistics*.

"Mixture Models: Theory, Geometry and Applications" (n.d.). In: ().

Molenaar, Peter C. M. (June 1985). "A Dynamic Factor Model for the Analysis of Multivariate Time Series". In: *Psychometrika* 50.2, pp. 181–202.

Molenberghs, Geert and Geert Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics. New York ; London: Springer. 683 pp.

Montopoli, George and Donald A. Anderson (Mar. 31, 2001). "THE ANALYSIS OF DISCRETE CHOICE EXPERIMENTS WITH CORRELATED ERROR STRUCTURE". In: *Communications in Statistics - Theory and Methods* 30.4, pp. 615–626.

Morais, Joanna, Christine Thomas-Agnan, and Michel Simioni (n.d.). "A Tour of Regression Models for Explaining Shares". In: ().

Moreno, Carlos Julio (n.d.). "The Zeros of Exponential Polynomials (I)". In: ().

Morgan, Charity J. et al. (June 15, 2014). “A Hierarchical Finite Mixture Model That Accommodates Zero-Inflated Counts, Non-Independence, and Heterogeneity”. In: *Statistics in Medicine* 33.13, pp. 2238–2250.

Mouatassim, Younès and El Hadj Ezzahid (2012). “Poisson Regression and Zero-inflated Poisson Regression: Application to Private Health Insurance Data”. In: *European actuarial journal* 2.2, pp. 187–204.

Mufudza, Chipo and Hamza Erol (2016). “Poisson Mixture Regression Models for Heart Disease Prediction”. In: *Computational and Mathematical Methods in Medicine* 2016, pp. 1–10.

Müller, Hans-Georg and Fang Yao (July 2006). “Regressing Longitudinal Response Trajectories on a Covariate”. In: Fan, Jianqing and Hira L Koul. *Frontiers in Statistics*. PUBLISHED BY IMPERIAL COLLEGE PRESS and DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO., pp. 305–324.

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. Cambridge, MA: MIT Press. 1067 pp.

Musliner, Katherine L (n.d.). “HETEROGENEITY IN LONG-TERM TRAJECTORIES OF DEPRESSION: A REVIEW AND APPLICATION OF GROUP-BASED TRAJECTORY MODELING”. In: ().

Muthén, Bengt (2004). “Latent Variable Analysis: Growth Mixture Modeling and Related Techniques for Longitudinal Data”. In: Kaplan, David. *The SAGE Handbook of Quantitative Methodology for the Social Sciences*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc., pp. 346–369.

— (Nov. 2006). “The Potential of Growth Mixture Modelling”. In: *Infant and Child Development* 15.6, pp. 623–625.

Nagin, Daniel and Richard E. Tremblay (Sept. 1999). “Trajectories of Boys’ Physical Aggression, Opposition, and Hyperactivity on the Path to Physically Violent and Nonviolent Juvenile Delinquency”. In: *Child Development* 70.5, pp. 1181–1196.

Nagin, Daniel S, Bobby L Jones, et al. (July 2018). “Group-Based Multi-Trajectory Modeling”. In: *Statistical Methods in Medical Research* 27.7, pp. 2015–2023.

Nagin, Daniel S and Richard E Tremblay (2001). “Analyzing Developmental Trajectories of Distinct but Related Behaviors: A Group-Based Method”. In: *Psychological methods* 6.1, p. 18.

Nagin, Daniel S. (2005). *Group-Based Modeling of Development*. Cambridge, Mass: Harvard University Press. 201 pp.

Nagin, Daniel S. (2014). “Group-Based Trajectory Modeling: An Overview”. In: *Annals of Nutrition and Metabolism* 65.2-3, pp. 205–210.

Nagin, Daniel S. and Candice L. Odgers (Dec. 2010). “Group-Based Trajectory Modeling (Nearly) Two Decades Later”. In: *Journal of Quantitative Criminology* 26.4, pp. 445–453.

Nagin, Daniel S. and Richard E. Tremblay (Nov. 2005). “DEVELOPMENTAL TRAJECTORY GROUPS: FACT OR A USEFUL STATISTICAL FICTION?*”. In: *Criminology* 43.4, pp. 873–904.

Naik, Prasad A, Peide Shi, and Chih-Ling Tsai (Mar. 2007). “Extending the Akaike Information Criterion to Mixture Regression Models”. In: *Journal of the American Statistical Association* 102.477, pp. 244–254.

Nash, J. C. (Dec. 13, 2018). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*. 2nd ed. Routledge.

Nash, John C (Author) (n.d.). “Nonlinear Parameter Optimization Using R Tools”. In: ().

Nawa, Victor Mooto (2014). “A Mixture Model for Longitudinal Trajectories”. In: *International journal of statistics and applications* 4, pp. 181–191.

— (2015). “A Mixture Model for Longitudinal Trajectories with Covariates”. In: *American Journal of Mathematics and Statistics* 5, pp. 293–305.

Nering, Evar D. (1970). *Linear Algebra and Matrix Theory*. 2. ed., 1. print. New York: Wiley. 352 pp.

Neyman, J. and E. S. Pearson (July 1928). “On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I”. In: *Biometrika* 20A.1/2, p. 175. JSTOR: 2331945.

Ng, H.K.T., P.S. Chan, and N. Balakrishnan (June 2002). “Estimation of Parameters from Progressively Censored Data Using EM Algorithm”. In: *Computational Statistics & Data Analysis* 39.4, pp. 371–386.

Ng, Shu Kay, Thriyambakam Krishnan, and Geoffrey J. McLachlan (2012). “The EM Algorithm”. In: *Handbook of Computational Statistics*. Ed. by James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 139–172.

Ng, Shu-Kay (Nov. 2013). “Recent Developments in Expectation-Maximization Methods for Analyzing Complex Data: Recent Developments in EM Methods”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6, pp. 415–431.

Nguena Nguefack, Hermine Lore et al. (Oct. 2020). “Trajectory Modelling Techniques Useful to Epidemiological Research: A Comparative Narrative Review of Approaches”. In: *Clinical Epidemiology* Volume 12, pp. 1205–1222.

Nguyen, Hien Duy (May 8, 2015). “Finite Mixture Models for Regression Problems”. PhD thesis. The University of Queensland.

Nielsen, Hans Bruun (2000). *UCMINF - an Algorithm for Unconstrained, Nonlinear Optimization.* Informatics and Mathematical Modelling, Technical University of Denmark, DTU.

Nielsen, J. D. et al. (Oct. 18, 2014). “Group-Based Criminal Trajectory Analysis Using Cross-validation Criteria”. In: *Communications in Statistics - Theory and Methods* 43.20, pp. 4337–4356.

“Non-Uniqueness in Probabilistic Numerical Identification of Bacteria” (n.d.). In: ().

Nunez, O G and D Concordet (n.d.). “When Is a Nonlinear Mixed-Effects Model Identifiable ?” In: ().

Nylund, Karen L., Tihomir Asparouhov, and Bengt O. Muthén (Oct. 23, 2007). “Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 14.4, pp. 535–569.

O’Brien, Liam M. and Garrett M. Fitzmaurice (Jan. 1, 2004). “Analysis of Longitudinal Multiple-Source Binary Data Using Generalized Estimating Equations”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 53.1, pp. 177–193.

Oehlert, Gary W. (Feb. 1992). “A Note on the Delta Method”. In: *The American Statistician* 46.1, pp. 27–29.

Ohlsson, Esbjörn and Björn Johansson (2010). *Non-Life Insurance Pricing with Generalized Linear Models.* EAA Lecture Notes. Berlin, Heidelberg: Springer Berlin Heidelberg.

Olino, Thomas M. et al. (May 4, 2014). “Trajectories of Depression and Anxiety Symptoms in Adolescent Girls: A Comparison of Parallel Trajectory Approaches”. In: *Journal of Personality Assessment* 96.3, pp. 316–326.

Oliveira, Amílcar and Teresa Oliveira (n.d.). “Method for Detection of Mixtures of Normal Distributions with Application to Vine Varieties”. In: ().

Olteanu, Madalina and James Ridgway (2012). “Hidden Markov Models for Time Series of Counts with Excess Zeros”. In: *Computational Intelligence.*

Pan, Yi and Jianxin Pan (n.d.). “Roptim: General Purpose Optimization with C++”. In: ().

Panhard, X. and A. Samson (May 23, 2008). “Extension of the SAEM Algorithm for Nonlinear Mixed Models with 2 Levels of Random Effects”. In: *Biostatistics* 10.1, pp. 121–135.

Park, Chanseok and Seong Beom Lee (Mar. 17, 2012). *Parameter Estimation from Censored Samples Using the Expectation-Maximization Algorithm.* arXiv: `1203.3880 [stat]`. URL: `http://arxiv.org/abs/1203.3880` (visited on 09/19/2023). Pre-published.

Paul, Thomas J. (May 11, 2012). *Enumerative Geometry of Hyperplane Arrangements:* Fort Belvoir, VA: Defense Technical Information Center.

Paulino, Carlos Daniel Mimoso and Carlos Alberto Bragança Pereira (Feb. 1994). “On Identifiability of Parametric Statistical Models”. In: *Journal of the Italian Statistical Society* 3.1, pp. 125–151.

Penrose, R. (July 1955). “A Generalized Inverse for Matrices”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 51.3, pp. 406–413.

Petersen, Kaare Brandt and Michael Syskind Pedersen (n.d.). “[http://matrixcookbook.com]”. In: ().

Peterson, Leif E. (1997). “**PIRLS** : Poisson Iteratively Reweighted Least Squares Computer Program for Additive, Multiplicative, Power, and Non-Linear Models”. In: *Journal of Statistical Software* 2.5.

Petitjean, François (n.d.). “Description des alignements formés par DTW”. In: ().

Petitjean, François, Alain Ketterlin, and Pierre Gançarski (Mar. 2011). “A Global Averaging Method for Dynamic Time Warping, with Applications to Clustering”. In: *Pattern Recognition* 44.3, pp. 678–693.

Pham, D T, S S Dimov, and C D Nguyen (Jan. 1, 2005). “Selection of K in K -Means Clustering”. In: *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 219.1, pp. 103–119.

Picard, Franck (n.d.). “An Introduction to Mixture Models”. In: 7 ().

Preisser, John S. and Bahjat F. Qaqish (June 1999). “Robust Regression for Clustered Data with Application to Binary Responses”. In: *Biometrics* 55.2, pp. 574–579.

“Proceedings Of 1993 International Joint Conference On Neural Networks” (1993). In: *Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan)*. Proceedings of 1993 International Conference on Neural Networks (IJCNN-93-Nagoya, Japan). Nagoya, Japan: IEEE, pp. i–xxxxiii.

Proust, Cécile et al. (Dec. 2006). “A Nonlinear Model with Latent Process for Cognitive Evolution Using Multivariate Longitudinal Data”. In: *Biometrics* 62.4, pp. 1014–1024.

Proust-Lima, Cécile, Luc Letenneur, and Hélène Jacqmin-Gadda (May 10, 2007). “A Nonlinear Latent Class Model for Joint Analysis of Multivariate Longitudinal Data and a Binary Outcome”. In: *Statistics in Medicine* 26.10, pp. 2229–2245.

Proust-Lima, Cécile, Viviane Philipps, and Benoit Liquet (2017). “Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The *R* Package **Lcmm**”. In: *Journal of Statistical Software* 78.2.

Puma, Michael et al. (n.d.). “What to Do When Data Are Missing in Group Randomized Controlled Trials”. In: ().

Puri, Madan Lal and Pranab Kumar Sen (1971). “Nonparametric Methods in Multivariate Analysis”. In: *(No Title)*.

Ram, Nilam and Kevin J. Grimm (Nov. 2009). “Methods and Measures: Growth Mixture Modeling: A Method for Identifying Differences in Longitudinal Change among Unobserved Groups”. In: *International Journal of Behavioral Development* 33.6, pp. 565–576.

Ramsay, James, Giles Hooker, and Spencer Graves (2009). *Functional Data Analysis with R and MATLAB*. New York, NY: Springer New York.

Ramsay, James O. and B. W. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics. New York Berlin Heidelberg: Springer. 190 pp.

Ramsay, James O. and Bernard W. Silverman (2006). *Functional Data Analysis*. 2. ed., [Nachdr.] Springer Series in Statistics. New York, NY: Springer. 426 pp.

Rasmussen, Carl Edward (n.d.). “The Infinite Gaussian Mixture Model”. In: ().

Reinoso, G. et al. (Mar. 2015). “Clinical Evolution of Parkinson’s Disease and Prognostic Factors Affecting Motor Progression: 9-Year Follow-up Study”. In: *European Journal of Neurology* 22.3, pp. 457–463.

Rencher, Alvin C. and G. Bruce Schaalje (2008). *Linear Models in Statistics*. 2. ed. Hoboken, NJ: Wiley-Interscience. 672 pp.

Resear, G.B.I.F., 1880- Greenwood, and H.M. Woods (2016). *The Incidence of Industrial Accidents upon Individuals, with Special Reference to Multiple Accidents*. Creative Media Partners, LLC.

Riazoshams, Hossein, Habshah Midi, and Gebrenegus Ghilagaber (2018). *Robust Nonlinear Regression: With Application Using R*. Hoboken: John Wiley & Sons.

Ridolfi, Andrea and Jérôme Idier (n.d.). “Penalized Maximum Likelihood Estimation for Normal Mixture Distributions”. In: ().

Rizopoulos, Dimitris (2017). “An Introduction to the Joint Modeling of Longitudinal and Survival Data, with Applications in R”. In.

Roche, Alexis (n.d.). “EM Algorithm and Variants: An Informal Tutorial”. In: ().

Roeder, Kathryn, Kevin G Lynch, and Daniel S Nagin (n.d.). “Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology”. In: ().

Rossi, Fabrice et al. (n.d.). “Hidden Process Regression for Curve Modeling, Classification and Tracking”. In: ().

Rubin, Donald B. (June 1975a). “INFERENCE AND MISSING DATA”. In: *ETS Research Bulletin Series* 1975.1, pp. i–19.

— (June 1975b). “INFERENCE AND MISSING DATA”. In: *ETS Research Bulletin Series* 1975.1, pp. i–19.

Ruckstuhl, Andreas (n.d.). “Introduction to Nonlinear Regression”. In: ().

Rudin, Walter (2013). *Real and Complex Analysis*. 3. ed., internat. ed., [Nachdr.] McGraw-Hill International Editions Mathematics Series. New York, NY: McGraw-Hill. 416 pp.

Russell, Gary J and Ann Petersen (2000). “Analysis of Cross Category Dependence in Market Basket Selection”. In: *Journal of Retailing* 76.3, pp. 367–392.

Ryoo, Ji Hoon et al. (May 8, 2018). “Longitudinal Model Building Using Latent Transition Analysis: An Example Using School Bullying Data”. In: *Frontiers in Psychology* 9, p. 675.

Santos, Frédéric (n.d.). “L’algorithme EM : une courte présentation”. In: ().

Sardá-Espinosa, Alexis (n.d.). “Comparing Time-Series Clustering Algorithms in R Using the Dtwclust Package”. In: ().

Sarul, Latife Sinem (Dec. 23, 2015). “AN APPLICATION OF CLAIM FREQUENCY DATA USING ZERO INFLATED AND HURDLE MODELS IN GENERAL INSURANCE”. In: *Pressacademia* 4.4, pp. 732–732.

Sarul, Latife Sinem and Serap Sahin (2015). “An Application of Claim Frequency Data Using Zero Inflated and Hurdle Models in General Insurance”. In: *Journal of Business Economics and Finance* 4.4.

Schiltz, Jang (n.d.). “Robustness of Groups and Trajectories in Nagin’s Finite Mixture Model”. In: ().

Schwarz, Gideon (1978). “Estimating the Dimension of a Model”. In: *The Annals of Statistics* 6.2, pp. 461–464. JSTOR: 2958889.

Seber, G. A. F. and C. J. Wild (Feb. 15, 1989). *Nonlinear Regression*. 1st ed. Wiley Series in Probability and Statistics. Wiley.

Severson, Kristen A et al. (Sept. 2021). “Discovery of Parkinson’s Disease States and Disease Progression Modelling: A Longitudinal Data Study Using Machine Learning”. In: *The Lancet Digital Health* 3.9, e555–e564.

Shanno, David F (1970). “Conditioning of Quasi-Newton Methods for Function Minimization”. In: *Mathematics of computation* 24.111, pp. 647–656.

Shedden, Kerby and Robert A. Zucker (Dec. 2008). “Regularized Finite Mixture Models for Probability Trajectories”. In: *Psychometrika* 73.4, pp. 625–646.

Shi, Qiuling et al. (Nov. 2013). “Using Group-Based Trajectory Modeling to Examine Heterogeneity of Symptom Burden in Patients with Head and Neck Cancer Undergoing Aggressive Non-Surgical Therapy”. In: *Quality of Life Research* 22.9, pp. 2331–2339.

Shim, Yosung, Jiwon Chung, and In-Chan Choi (2005). “A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm”. In: *International Conference on Computational Intelligence for Modelling, Control and Automation and Interna-*

tional Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06). Vol. 1. IEEE, pp. 199–204.

Sieber, Beth-Anne et al. (Oct. 2014). "Prioritized Research Recommendations from the National Institute of Neurological Disorders and Stroke *P Arkinson's D Isease 2014 Conference*: NINDS PD 2014 Recommendations". In: *Annals of Neurology* 76.4, pp. 469–472.

Simas, Alexandre B., Wagner Barreto-Souza, and Andréa V. Rocha (Feb. 2010). "Improved Estimators for a General Class of Beta Regression Models". In: *Computational Statistics & Data Analysis* 54.2, pp. 348–366. arXiv: `0809.1878 [stat]`.

Simone, Rosaria (Sept. 2022). "On Finite Mixtures of Discretized Beta Model for Ordered Responses". In: *TEST* 31.3, pp. 828–855.

Simuni, Tanya et al. (May 2018). "Longitudinal Change of Clinical and Biological Measures in Early Parkinson's Disease: Parkinson's Progression Markers Initiative Cohort: Early PD and MDS-UPDRS and Dat Binding Change". In: *Movement Disorders* 33.5, pp. 771–782.

Smithson, Michael and Jay Verkuilen (Mar. 2006). "A Better Lemon Squeezer? Maximum-likelihood Regression with Beta-Distributed Dependent Variables." In: *Psychological Methods* 11.1, pp. 54–71.

Sohil, Fariha, Muhammad Umair Sohali, and Javid Shabbir (Jan. 2, 2022). "An Introduction to Statistical Learning with Applications in R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7". In: *Statistical Theory and Related Fields* 6.1, pp. 87–87.

Solomon, Justin (June 24, 2015). *Numerical Algorithms: Methods for Computer Vision, Machine Learning, and Graphics*. 0th ed. A K Peters/CRC Press.

Sridharan, Ramesh (n.d.). "Gaussian Mixture Models and the EM Algorithm". In: ().

Stanley, R P (Mar. 19, 1996). "Hyperplane Arrangements, Interval Orders, and Trees." In: *Proceedings of the National Academy of Sciences* 93.6, pp. 2620–2625.

Stanley, Richard P (n.d.). "An Introduction to Hyperplane Arrangements". In: ().

Statistical Models of Shape (2008). *Statistical Models of Shape: Optimisation and Evaluation*. London: Springer London.

Stemmler, Mark, Alexander Von Eye, and Wolfgang Wiedermann, eds. (2015). *Dependent Data in Social Sciences Research: Forms, Issues, and Methods of Analysis*. Vol. 145. Springer Proceedings in Mathematics & Statistics. Cham: Springer International Publishing.

Tahmasebi, Behrooz, Seyed Abolfazl Motahari, and Mohammad Ali Maddah-Ali (July 14, 2018). *On the Identifiability of Finite Mixtures of Finite Product Measures*. arXiv: `1807.05444 [math, stat]`. URL: `http://arxiv.org/abs/1807.05444` (visited on 09/19/2023). Pre-published.

Teicher, Henry (Mar. 1960). "On the Mixture of Distributions". In: *The Annals of Mathematical Statistics* 31.1, pp. 55–73.

— (Dec. 1963a). "Identifiability of Finite Mixtures". In: *The Annals of Mathematical Statistics* 34.4, pp. 1265–1269.

— (Dec. 1963b). "Identifiability of Finite Mixtures". In: *The Annals of Mathematical Statistics* 34.4, pp. 1265–1269.

— (Aug. 1967). "Identifiability of Mixtures of Product Measures". In: *The Annals of Mathematical Statistics* 38.4, pp. 1300–1302.

Teuling, Niek Den, Steffen Pauws, and Edwin van den Heuvel (Nov. 9, 2021). *Clustering of Longitudinal Data: A Tutorial on a Variety of Approaches.* arXiv: 2111.05469 [cs, stat]. URL: http://arxiv.org/abs/2111.05469 (visited on 09/19/2023). Pre-published.

"The Incidence of Industrial Accidents upon Individuals, with Special Reference to Multiple Accidents" (n.d.). In: ().

Tremblay, Richard E. et al. (2003). "The Montreal Longitudinal and Experimental Study: Rediscovering the Power of Descriptions". In: Thornberry, Terence P. and Marvin D. Krohn. *Taking Stock of Delinquency.* Boston, MA: Springer US, pp. 205–254.

Tsuda, Masaki E (n.d.). "Rcpp for Everyone". In: ().

Tuwei, Kipkorir E (n.d.). "Power Series Distribution Sand Zero-Inflated Models". In: ().

Van de L'Isle, Natasja (n.d.). "Algorithms for Center-Based Trajectory Clustering". In: ().

Van Der Nest, Gavin et al. (Mar. 2020). "An Overview of Mixture Modelling for Latent Evolutions in Longitudinal Data: Modelling Approaches, Fit Statistics and Software". In: *Advances in Life Course Research* 43, p. 100323.

Van Dorp, RenÉ J. and Thomas A. Mazzuchi (Sept. 2000). "Solving For the Parameters of a Beta a Distribution under Two Quantile Constraints". In: *Journal of Statistical Computation and Simulation* 67.2, pp. 189–201.

Van Montfort, Kees, Johan H.L. Oud, and Albert Satorra, eds. (2010). *Longitudinal Research with Latent Variables.* Berlin, Heidelberg: Springer Berlin Heidelberg.

Van Leeuwen, Marijtje (n.d.). "Estimating Standard Errors of Parameters Obtained by the EM-Algorithm". In: ().

Vannoorenberghe, Patrick (n.d.). "Estimation de modèles de mélanges finis par un algorithme EM crédibiliste". In: ().

Van Wieringen, Wessel N. (June 27, 2023). *Lecture Notes on Ridge Regression.* arXiv: 1509.09169 [stat]. URL: http://arxiv.org/abs/1509.09169 (visited on 09/19/2023). Pre-published.

Varin, Cristiano, Nancy Reid, and David Firth (n.d.). “AN OVERVIEW OF COMPOSITE LIKELIHOOD METHODS”. In: ().

Verbeke, Geert et al. (Feb. 2014). “The Analysis of Multivariate Longitudinal Data: A Review”. In: *Statistical Methods in Medical Research* 23.1, pp. 42–59.

Wakefield, J C (2004). “NON-LINEAR REGRESSION MODELLING”. In.

Wald, Abraham (Mar. 1941). “Asymptotically Most Powerful Tests of Statistical Hypotheses”. In: *The Annals of Mathematical Statistics* 12.1, pp. 1–19.

Walsh, Michael J (n.d.). “Computing the Observed Information Matrix for Dynamic Mixture Models”. In: ().

Wang, Peiming et al. (June 1996). “Mixed Poisson Regression Models with Covariate Dependent Rates”. In: *Biometrics* 52.2, p. 381. JSTOR: 2532881.

Wang, Shaoli, Weixin Yao, and Mian Huang (Oct. 2014). “A Note on the Identifiability of Nonparametric and Semiparametric Mixtures of GLMs”. In: *Statistics & Probability Letters* 93, pp. 41–45.

Wang, Wei (Jan. 1, 2013). “Identifiability of Linear Mixed Effects Models”. In: *Electronic Journal of Statistics* 7 (none).

Wang, Zhichao et al. (June 2020). “Linear Mixed-Effects Model for Longitudinal Complex Data with Diversified Characteristics”. In: *Journal of Management Science and Engineering* 5.2, pp. 105–124.

Ward, Ashley K. et al. (Nov. 2010). “Criminal Trajectories and Risk Factors in a Canadian Sample of Offenders”. In: *Criminal Justice and Behavior* 37.11, pp. 1278–1300.

Ward, Gill et al. (June 2009). “Presence-Only Data and the EM Algorithm”. In: *Biometrics* 65.2, pp. 554–563.

Wegman, Edward J. (Sept. 1990). “Hyperdimensional Data Analysis Using Parallel Coordinates”. In: *Journal of the American Statistical Association* 85.411, pp. 664–675.

Weinberg, Clarice A., David M. Umbach, and Sander Greenland (Sept. 1994). “When Will Nondifferential Misclassification of an Exposure Preserve the Direction of a Trend?” In: *American Journal of Epidemiology* 140.6, pp. 565–571.

Wolfe, Philip (1969). “Convergence Conditions for Ascent Methods”. In: *SIAM review* 11.2, pp. 226–235.

— (1971). “Convergence Conditions for Ascent Methods. II: Some Corrections”. In: *SIAM review* 13.2, pp. 185–188.

Wu, Hulin and Jin-Ting Zhang (n.d.). “Nonparametric Regression Methods for Longitudinal Data Analysis”. In: ().

Wylie, Timothy Randall (n.d.). “The Discrete Fréchet Distance with Applications”. In: ().

Xia, Michelle and P Richard Hahn (n.d.). “A Finite Mixture Model Approach to Regression under Covariate Misclassification”. In: ().

Xiang, Sijia and Weixin Yao (n.d.). “Mixture of Regression Models with Single-index”. In: ().

Yakowitz, Sidney J. and John D. Spragins (Feb. 1968). “On the Identifiability of Finite Mixtures”. In: *The Annals of Mathematical Statistics* 39.1, pp. 209–214.

Yi, Grace Y. (Feb. 15, 2017). “Composite Likelihood/Pseudolikelihood”. In: *Wiley StatsRef: Statistics Reference Online*. Ed. by N. Balakrishnan et al. 1st ed. Wiley, pp. 1–14.

Yosung Shim, Jiwon Chung, and In-Chan Choi (2005). “A Comparison Study of Cluster Validity Indices Using a Nonhierarchical Clustering Algorithm”. In: *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06)*. International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC’06). Vol. 1. Vienna, Austria: IEEE, pp. 199–204.

Yu, Qingzhao and Bin Li (2014). “Regularization and Estimation in Regression with Cluster Variables”. In: *Open Journal of Statistics* 04.10, pp. 814–825.

Zhang, Hanze and Yangxin Huang (n.d.). “Finite Mixture Models and Their Applications: A Review”. In: ().

Zhang, Yue and Kiros Berhane (Mar. 11, 2016). “Dynamic Latent Trait Models with Mixed Hidden Markov Structure for Mixed Longitudinal Outcomes”. In: *Journal of Applied Statistics* 43.4, pp. 704–720.

Zhu, Hao (n.d.). “Create Awesome LaTeX Table with Knitr::Kable and kableExtra”. In: ().