A generalized finite mixture model

Jang SCHILTZ*
University of Luxembourg, LSF, Luxembourg, Luxembourg, jang.schiltz@uni.lu

Abstract

We present a generalization of Nagin's finite mixture model that allows non parallel trajectories for different values of covariates. We investigate some mathematical properties of this model and illustrate its use by giving typical salary curves for the employees in the private sector in Luxembourg between 1981 and 2006, as a function of their gender, as well as of Luxembourg's gross domestic product (GDP).
Keywords: Statistical Models; Developmental trajectories; Trajectory Modeling.

1. Introduction

Longitudinal data are the empirical basis of research on various subjects in sociology, psychology, economics, criminology and medicine and a host of statistical techniques are available for analyzing them (see Singer & Willet 2003). The common statistical aim of these various application fields is the modelization of the evolution of an age or time based phenomenon. In the 1990s, the generalized mixed model assuming a normal distribution of unobserved heterogeneity (Bryk & Raudenbush 1992), multilevel modeling (Goldstein 1995), latent growth curves modeling (Muthén 1989) and the nonparametric mixture model, based on a discrete distribution of heterogeneity (Nagin 1999) have emerged.

The nonparametric mixed model or semiparametric mixture model was originally discussed by Nagin and Land (1993)and is specifically designed to detect the presence of distinct subgroups among a set of trajectories. Compared to subjective classification methods, the nonparametric mixed model has the advantage of providing a formal framework for testing the existence of distinct groups of trajectories. This method does not assume a priori that there is necessarily more than one group in the population. Rather, an adjustment index is used to determine the number of sub-optimal groups.

While the conceptual aim of the analysis is to identify clusters of individuals with similar trajectories, the model's estimated parameters are not the result of a cluster analysis but of maximum likelihood estimation (Nagin 2005). Nagin and Odgers (2010) document numerous applications of group-based trajectory modeling in criminology and clinical research. They state that the appeal of group-based trajectory modeling for the future lies in the potential for the innovative application of trajectory models on their own, in conjunction with other statistical methods or embedded within creative study designs while carefully considering the perils and pitfalls inherent in the use of any methodology.

The remainder of this article is structured as follows. In section two, we present the basic version of Nagin's finite mixture model, as well as one of his generalizations, allowing to add covariates to the trajectories and we show two drawbacks of the model. In section three, we present a generalization of the model that overcomes these drawbacks and we discuss model selection and group member probabilities for the new model. Section four presents some basic statistical properties of the model. In section five, finally, we highlight typical features of the new model by means of a data example from economics.

2. Nagin's Finite Mixture Model

Starting from a collection of individual trajectories, the aim of Nagin's finite mixture model is to divide the population into a number of homogenous sub-populations and to estimate, at the same time, a typical trajectory for each sub-population (Nagin 2005).

More, precisely, consider a population of size N and a variable of interest Y. Let $Y_i = y_{i_1}, y_{i_2}, ..., y_{i_T}$ be T measures of the variable Y, taken at times $t_1, ..., t_T$ for subject number i. To estimate the parameters defining the shape of the trajectories, we need to fix the number r of desired subgroups. Denote the probability of a given subject to belong to group number j by π_j.

The objective is to estimate a set of parameters $\Omega = \{\pi_j, \beta_0^j, \beta_1^j, ...; j = 1, ..., r\}$ which allow to maximize the probability of the measured data. The particular form of Ω is distribution specific, but the β parameters

always perform the basic function of defining the shapes of the trajectories. In Nagin's finite mixture model, the shapes of the trajectories are described by a polynomial function of age or time. In this paper, we suppose that the data follow a normal distribution. Assume that for a subject in group j

$$y_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k + \varepsilon_{it}, \tag{1}$$

where s denotes the order of the polynomial describing the trajectories in group j and ε_{it} is a disturbance assumed to be normally distributed with a zero mean and a constant standard deviation σ. If we denote the density of the standard centered normal law by ϕ and $\beta^j t_{it} = \sum_{k=1}^{s} \beta_k^j t_{it}^k$, the likelihood of the data is given by

$$L = \frac{1}{\sigma} \prod_{i=1}^{N} \sum_{j=1}^{r} \pi_j \prod_{t=1}^{T} \phi\left(\frac{y_{it} - \beta^j t_{it}}{\sigma}\right). \tag{2}$$

The disadvantage of the basic model is that the trajectories are static and do not evolve in time. Thus, Nagin introduced several generalizations of his model in his book (Nagin 2005). Among others, he introduced a model allowing to add covariates to the trajectories. Let $z_1, ..., z_M$ be M covariates potentially influencing Y. We are then looking for trajectories

$$y_{it} = \sum_{k=0}^{s} \beta_k^j t_{it}^k + \alpha_1^j z_1 + ... + \alpha_M^j z_M + \varepsilon_{it}, \tag{3}$$

where ε_{it} is normally distributed with zero mean and a constant standard deviation σ. The covariates z_m may depend or not upon time t.

But even this generalized model still has two major drawbacks. First, the influence of the covariates in this model is unfortunately limited to the intercept of the trajectory. This implies that for different values of the covariates, the corresponding trajectories will always remain parallel by design, which does not necessarily correspond to reality.

Secondly, in Nagin's model, the standard deviation of the disturbance is the same for all the groups. That too is quite restrictive. One can easily imagine situations in which in some of the groups all individual are quite close to the mean trajectory of their group, whereas in other groups there is a much larger dispersion.

3. Our model

To address and overcome these two drawbacks, we propose the following generalization of Nagin's model.

Let $x_1 ... x_M$ and $z_{i_1}, ..., z_{i_T}$ be covariates potentially influencing Y. Here the x variables are covariates not depending on time like gender or cohort membership in a multicohort longitudinal study and the z variable is a covariate depending on time like being employed or unemployed. They can of course also designate time-dependent covariates not depending on the subjects of the data set which still influence the group trajectories, like GDP of a country in case of an analysis of salary trajectories.

The trajectories in group j will then be written as

$$y_{it} = \sum_{k=0}^{s} \left(\beta_k^j + \sum_{m=1}^{M} \alpha_{km}^j x_m + \gamma_k^j z_{i_t}\right) t_{it}^k + \varepsilon_{it}, \tag{4}$$

where the disturbance ε_{it} is normally distributed with mean zero and a standard deviation σ_j constant inside group j but different from one group to another. Since, for each group, this model is just a classical fixed effects model for panel data regression (see Woolridge 2002), it is well defined and we can get consistent estimates for the model parameters.

Our model allows obviously to overcome the drawbacks of Nagin's model. The standard deviation of the uncertainty can vary across groups and the trajectories depend in a nonlinear way on the covariates. In practice this dependance of all the power coefficients of the polynomials may considerably extend the computation time for the parameters, so it can be useful just to work with a first or second order dependance instead of using the full model.

Since our model is just a generalization of Nagin's finite mixture model, a lot of its main features and properties remain the same as in Nagin's model.

4. Statistical Properties

The model's estimated parameters are the result of maximum likelihood estimation. As such, they are consistent and asymptotically normally distributed (Greene 2012).

In our model, for a given group, the trajectories follow in fact a nonlinear regression model. As such, exact confidence interval procedures or exact hypothesis tests for the parameters are generally not available. There exist however approximative solutions. The standard error can be approximated for instance by a first-order Taylor series expansion (Greene 2012). This approximate standard error (ASE) is usually quite precise if the sample size is sufficiently large.

Consider model (4), for which $(2+M)s$ regression parameters have to be estimated. Then confidence intervals of level α for the parameters β_k^j are just

$$CI_\alpha(\beta_k^j) = \left[\hat{\beta}_k^j - t_{1-\alpha/2;N-(2+M)s} ASE(\hat{\beta}_k^j); \hat{\beta}_k^j + t_{1-\alpha/2;N-(2+M)s} ASE(\hat{\beta}_k^j)\right], \tag{5}$$

where $t_{1-\alpha;n}$ denotes as usually the $1-\alpha$ quantile of the Student distribution with n degrees of freedom.

The confidence intervals for the α_{kl}^j and γ_k^j are obtained in the same way.

The confidence intervals of level α for the disturbance factor σ_j is given by

$$CI_\alpha(\sigma_j) = \left[\sqrt{\frac{(N-(2+M)s-1)\hat{\sigma_j}^2}{\chi^2_{1-\alpha/2;N-(2+M)s-1}}}; \sqrt{\frac{(N-(2+M)s-1)\hat{\sigma_j}^2}{\chi^2_{\alpha/2;N-(2+M)s-1}}}\right], \tag{6}$$

where $\chi^2_{1-\alpha;n}$ denotes the $1-\alpha$ quantile of the Chi-Square distribution with n degrees of freedom.

5. A data example

For the following example, we use Luxembourg administrative data originating from the General Inspectorate of Social Security, IGSS (Inspection gnrale de la scurit sociale). The data have previously been described and exploited with Nagin's basic model by Guigou, Lovat and Schiltz (2010, 2012). The file contains the salaries of all employees of the Luxembourg private sector who started their work in Luxembourg between 1980 and 1990 at an age of less than 30 years. This choice was made to eliminate people with a long carrier in another country before moving to Luxembourg. The main variables are the net annual taxable salary, measured in constant (2006 equivalent) euros, gender, age at first employment, residentship and nationality, sector of activity, marital status and the years of birth of the children. The file consists of 1303010 salary lines corresponding to 85049 employees. In Luxembourg, the maximum contribution ceiling on pension insurance is 5 times the minimum wage, currently 7577 EUR (2006 equivalent euros) per month. Wages in our data are thus also capped at that number.

We will not present here an exhaustive analysis of the whole dataset, but an illustration of the possibilities of our generalized mixture model and its differences from Nagin's model. We concentrate on the first 20 years of the careers of the employees who started working in Luxembourg in 1987. That gives us a sample of 1716 employees. We will compute typical salary trajectories for them, taking into account the gender of the employees, as well as their dependancy from the GDP of the country.

Since we are in the somewhat special situation where we work with the complete population and not just a sample, it may seem a bit strange to speak about parameter significance and confidence intervals for this example. But first, this is just an illustration of the possibilities and main features of our model, so it makes sense to show what results we would get in a classical situation. And more importantly, in case of a use of the results to predict the future salary evolution, we are dealing in fact with just a subsample of the whole population. If we argue that for a reasonable time horizon, the typical salary trajectories just depend on the covariates that we included in our equations, then the complete set of people starting to work in 2006 is

just a part of the whole population of people starting to work in 2006 and the subsequent years. Confidence intervals for the salary trajectories then indicate prediction bounds.
Let us first highlight the differences with respect to Nagin's extended model.

Figure 1 shows a three group solution modeled by Nagin's generalized model representing the salary of employees in Luxembourg during the first 20 years of their professional career. We see that for the low salary group women and men are gaining exactly the same salary (with the consequence that there appears just one salary trajectory for the two lower salary groups on the graph instead of two) whereas in the middle and high salary groups, men earn more than women. Due to the limitations of the model, the evolution of the salaries seems to be exactly the same for men and women; their salary trajectories are strictly parallel.

Figure 1: Salary evolution by gender, modelized by Nagin's model.

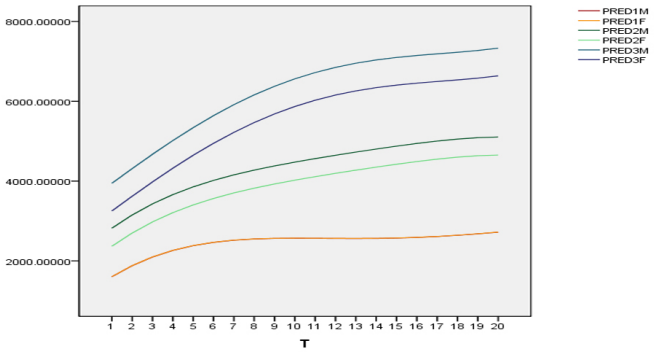
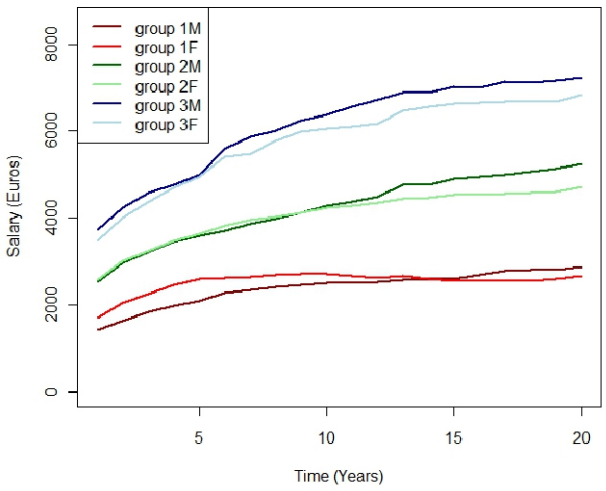


Figure 2 shows the three group solution for the 20 first year of Luxembourg employees calibrated with our model. We see a somewhat different and more realistic pattern emerging. For the high salary group the income of men and women remain more or less parallel, except for a short time interval around year five. This is however no longer the case for the middle and low salary groups. Here, we observe that the women in these groups have higher salaries then the men at the beginning of their career, but this is reversed somewhere in the middle and after 10 years for the middle salary group and 15 years for the low salary group the income of the men becomes higher then the one of the women.

Figure 2: Salary evolution by gender, modelized by our model



We obtained these results by calibrating the model

$$S_{it} = (\beta_0^j + \alpha_0^j x_i + \gamma_0 z_t) + (\beta_1^j + \alpha_1^j x_i + \gamma_1 z_t)t + (\beta_2^j + \alpha_2^j x_i + \gamma_2 z_t)t^2, \quad (7)$$

where S denotes the salary, x the gender and z_t is Luxembourg's GDP in year $t-1$ of the study. For figure 2, we replaced the variable z_t by the actual values of Luxembourg's GDP in the considered years.
Table 1 shows the values of the parameters for a 3-group solution of model 8.

Table 1: Parameter estimates for model

Results for group 1

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	476.183	132.857	191.413	760.856
α_0	220.302	1.387	202.568	227.896
γ_0	0.582	0.071	0.407	0.710
β_1	206.446	27.850	146.632	266.084
α_1	123.219	4.909	121.582	126.895
γ_1	-0.077	0.007	-0.092	-0.062
β_2	-3.828	1.760	-7.602	-0.053
α_2	-8.922	0.1838	-9.089	-8.753
γ_2	0.002	0.001	0.002	0.003

Results for group 2

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	2243.017	236.843	1734.795	2750.771
α_0	-380.402	116.972	-636.957	-122.585
γ_0	0.180	0.011	-0.074	0.433
β_1	370.016	49.685	263.469	475.590
α_1	12.846	8.153	-41.197	66.703
γ_1	-0.049	0.012	-0.074	-0.023
β_2	-11.018	3.140	-17.741	-4.272
α_2	-1.491	0.755	-4.902	1.947
γ_2	0.002	0.001	0.001	0.003

Results for group 3

Parameter	Estimate	Standard error	95% confidence interval	
			Lower	Upper
β_0	3293.714	335.402	2573.151	4011.944
α_0	-783.289	28.382	-892.997	-671.4
γ_0	0.189	0.025	-0.190	0.566
β_1	447.925	70.366	297.040	598.856
α_1	64.890	19.532	73.501	119.982
γ_1	0.036	0.017	-0.074	0.012
β_2	-13.873	4.447	-15.824	-9.174
α_2	-2.73	0.476	-4.196	-0.126
γ_2	0.001	0.001	0.000	0.002

The disturbance terms for the three groups are $\sigma_1 = 33.11$, $\sigma_2 = 54.18$ and $\sigma_3 = 78.85$ respectively. The dispersion is thus higher in the groups with higher salaries then in those with lower salaries. This makes

sense, since in the low salary group a lot of employees just earn the minimal wage. Hence, a lot of them have the same salary.

Moreover this example illustrates the dependence of the trajectories on Luxembourg's GDP. We see that in the three groups, this influence is non linear, since γ_2 is always significantly different from 0.
The trajectory equations from table 1 can now be used to predict the future evolution of the salaries for men and women as a function of GDP.

6. Conclusion

In this article, we presented Nagin's finite mixture model and some of its generalizations and showed some inherent shortcomings for possible applications. We addressed these by proposing a new generalized finite mixture model. A key characteristic is its ability to modelize nearly all kind of trajectories and to add covariates to the trajectories themselves in a nonlinear way.

We illustrated these possibilities through a data example about salary trajectories. We showed how to add a classical group membership predictor variable to the trajectories as well as a time serie that does not depend on the subjects of the analysis but influences the shape of the trajectories in some of the groups.

References

Bryk, A.S., & Raudenbush, S.W. (1992). Hierarchical linear models. Newbury Park, CA: Sage.

Goldstein, H. (1995). Multilevel Statistical Models. London: Arnold.

Greene, W.H. (2012). Econometric Analysis. 7th edition. New York: Macmillan.

Guigou, J.-D., Lovat, B., & Schiltz, J. (2010). The impact of ageing population on pay-as-you-go pension systems: The case of Luxembourg. Journal of International Finance and Economics, 10(1), 110–122.

Guigou, J.-D., Lovat, B., & Schiltz, J. (2012). Optimal mix of funded and unfunded pension systems: the case of Luxembourg. Pensions, 17(4), 208–222.

Jones, B.L., & Nagin, D.S. (2007). Advances in Group-Based Trajectory Modeling and an SAS Procedure for Estimating Them. Sociological Methods & Research, 35(4), 542–571.

Muthén, B.O. (1989). Latent Variable Modeling in Heterogeneous Populations. Psychometrika, 54(4), 557–585.

Nagin, D.S. (1999). Analyzing Developmental Trajectories: Semi-parametric. Groupe-based Approach. Psychological Method, 4, 139–157.

Nagin, D.S. (2005). Group-Based Modeling of Development. Cambridge, MA: Harvard University Press.

Nagin, D.S., & Land, K.C. (1993). Age, criminal careers and population heterogeneity: Specifiction and estimation of a nonparametric, mixed Poisson model. Criminology, 31, 327–362.

Nagin, D.S., & Odgers, C.L. (2010). Group-Based Trajectory Modeling (Nearly) Two Decades Later. Journal of Quantitative Criminology, 26, 445–453.

Singer, J.D., & Willet, J.B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York, NY: Oxford University Press.

Woolridge, J. (2002). Econometric Analysis of Cross-Section and Panel Data. Cambridge, MA: MIT press.