PhD-FDEF-2021-013
The Faculty of Law, Economics and Finance

DISSERTATION

Defence held on 22/07/2021 in Luxembourg
to obtain the degree of

Docteur de l'université du Luxembourg en sciences économiques

by

Andreï Victorovitch Kostyrka
Born on 12 July 1992 in Moscow (Russia)

Efficient estimation with non-standard sampling or missing endogenous variables, and conditional density modelling with unobserved copula-connected shocks

Dissertation defence committee

Prof. Dr Antonio Cosma, supervisor
Université du Luxembourg

Prof. Dr Gautam Tripathi, chairman
Université du Luxembourg

Prof. Dr Benjamin Holcblat, vice-chairman
Université du Luxembourg

Prof. Dr Shin Kanaya, member
University of Essex

Prof. Dr Valentin Patilea, member
École Nationale de Statistique et Analyse de l'Information

Acknowledgements

First and foremost, I am extremely grateful to my supervisors, professor Antonio Cosma and professor Gautam Tripathi, for their invaluable advice, continuous support, patience, kindness, and humanly attitude during my Ph. D. studies. Their wealth of experience and expertise has encouraged me at all times of my research and daily life, and I learned a lot from scientific discussions with them. They have always been thoughtful and have always put much time into my research-related and personal issues. I could not have completed this thesis save for their help. I also thank professor Michel Beine for the help and advice that he provided as a member of my CET committee.

My gratitude extends to the National Research Fund of Luxembourg (FNR) for the funding opportunity to undertake my studies at the Faculty of Law, Economics and Finance (FDEF) through a PRIDE grant for the Migration and Labour (MINLAB) doctoral training unit.

Finally, I would like to express my gratitude to my parents, for they were the ones who fostered my curiosity and encouraged my interest in science, without which it would be impossible for me to complete my studies.

Contents

List of Figures

List of Tables

Abstract

In Chapter 1, it is shown how to use a smoothed empirical likelihood approach to conduct efficient semi-parametric inference in models characterised as conditional moment equalities when data are collected by variable probability sampling. Results from a simulation experiment suggest that the smoothed-empirical-likelihood-based estimator can estimate the model parameters very well in small to moderately sized stratified samples.

In Chapter 2, a novel univariate conditional density model is proposed to decompose asset returns into a sum of copula-connected unobserved 'good' and 'bad' shocks. The novelty of this approach comes from two factors: correlation between unobserved shocks is modelled explicitly, and the presence of copula-connected discrete jumps is allowed for. The proposed framework is very flexible and subsumes other models, such as 'bad environments, good environments'. The proposed model shows certain hidden characteristics of returns, explains investors' behaviour in greater detail, and yields better forecasts of risk measures. The in-sample and out-of-sample performance of the proposed model is better than that of 40 popular GARCH variants. A Monte Carlo simulation shows that the proposed model recovers the structural parameters of the unobserved dynamics. This model is estimated on S&P 500 data, and time-dependent non-negative covariance between 'good' and 'bad' shocks with a leverage-like effect is found to be an essential component of the total variance. Asymmetric reaction to shocks is present almost in all characteristics of returns. The conditional distribution of returns seems to be very time-dependent with skewness both in the centre and tails. Continuous shocks are more important than discrete jumps for return modelling, at least at the daily frequency.

In Chapter 3, the semi-parametric efficiency bound is derived for estimating finite-dimensional parameters identified via a system of conditional moment equalities when at least one of the endogenous variables (which can either be endogenous outcomes, or endogenous explanatory variables, or both) is missing for some individuals in the sample. An interesting result is obtained that if there are no endogenous variables that are not missing, i. e. all the endogenous variables in the model are missing, then estimation using only the validation subsample (the sub-sample of observations for which the endogenous variables are non-missing) is asymptotically efficient. An estimator based on the full sample is proposed, and it is shown that it achieves the semi-parametric efficiency bound. A simulation study reveals that the proposed estimator can work well in medium-sized samples and that the resulting efficiency gains (measured as the ratio of the variance of an efficient estimator based on the validation sample and the variance of our estimator) are comparable with the maximum gain the simulation design can deliver.

Chapter 1

Inference in conditional moment restriction models when there is selection due to stratification

This chapter is based on joint work with Antonio Cosma and Gautam Tripathi.
Citation: Cosma, A., Kostyrka, A. V. & Tripathi, G. (2019). Inference in conditional moment restriction models when there is selection due to stratification. *The Econometrics of Complex Survey Data: Theory and Applications, 39*, 137–171. https://doi.org/10.1108/S0731-905320190000039010

1.1 Introduction

The gold standard for collecting data, at least for the ease of doing subsequent statistical analysis, is simple random sampling, whereby each observation in the 'target' population, namely, the population of interest, has an equal chance of being chosen. Consequently, the probability distribution of the chosen observation, regarded as belonging to a 'realised' population, is the same as the probability distribution of an observation in the target population, which facilitates statistical analysis.

However, when estimating or testing economic relationships, economists often discover that the data they plan to use is not drawn from the target population they wish to study. Instead, the observations are found to be sampled from a related but different population. Sometimes this is done deliberately to make the sample more informative. E. g. when studying the impact of welfare legislation, it is desirable to oversample minorities and low-income families. Similarly, if we want to examine the effect of disability laws on demand for public transportation, it makes sense to oversample households with disabled members. At other times, a distinction between the target and realised populations can be created unintentionally. E. g. in sampling the duration of unemployment at a randomly chosen time, economists are more likely to observe longer unemployment spells than shorter ones. Using a dataset to answer questions for which it was not originally designed, a typical situation in economics where data is often costly to collect, may also lead to such a situation (Newey, 1993, p. 419). For instance, if the reason for collecting data is to estimate mean income for an underlying population, oversampling low-income and undersampling high-income families can improve the precision of estimators. However, at some later stage, this income data can be used by another researcher as the dependent variable in a regression model without realising that the original sample was drawn from a distribution other than the target population.

Whatever its cause, if the distinction between the target and realised populations is not taken into account when analysing the data, statistical inference can be seriously off the mark. This phenomenon is commonly called selection bias. Cf. Heckman (1976, 1979) and Manski (1989, 1995) for a classic exposition of the selection problem.

In this paper, we describe an efficient semi-parametric approach for conducting inference in conditional moment restriction models when data is collected by a variable probability sampling scheme such that the observations from the target population have unequal chances of being chosen. In other words, we show how to efficiently deal with the selection bias caused by the sampling scheme used to collect the data because the sampling scheme induces a probability distribution on the realised population that differs from the target distribution for which inference is to be made.

The remainder of the paper is organised as follows. In Section 1.2, we describe the conditional moment restriction model and the variable probability sampling scheme. Section 1.3 discusses how to do inference using the smoothed empirical likelihood approach, and the finite-sample properties of the proposed estimator are examined in Section 1.4. Section 1.5 concludes the paper. Related technical details are in the appendices.

1.2 The model

1.2.1 Conditional moment equalities

Let $Z^* \stackrel{\text{def}}{=} (Y^*, X^*)_{(\dim Y^* + \dim X^*)\times 1}$ be a random (column) vector that denotes an observation from the target population, where Y^* is the vector of endogenous variables and X^* the vector of exogenous variables. Assume that

$$\mathrm{H}_0: \quad \exists \theta^* \in \mathbb{R}^{\dim \theta^*} \quad \text{s.t.} \quad \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*, \theta^*) \mid X^*] = 0 \qquad P^*_{X^*}\text{-a.s.}, \tag{1.2.1}$$

where g is a vector of functions known up to θ^*, the notation $\mathbb{E}_{P^*_{Y^*|X^*}}$ indicates that the conditional expectation is with respect to the conditional distribution $P^*_{Y^*|X^*} \stackrel{\text{def}}{=} \mathrm{Law}(Y^* \mid X^*)$, and $P^*_{X^*}$ denotes the marginal distribution of X^*. The conditional distribution of $Y^* \mid X^*$ and the marginal distribution of X^* are unknown.[1] Throughout the paper, random variables and probability measures associated with the target population appear with the superscript '$*$'. The parameter of interest θ^* has an asterisk attached to it because it is a functional of $P^*_{Y^*|X^*}$.[2]

A large class of models in applied economics can be characterised in terms of conditional moment equalities of the form (1.2.1). E. g. in linear regression models where some or all of the regressors are endogenous, we have $g(Z^*, \theta^*) \stackrel{\text{def}}{=} Y_1^* - \alpha^* - X_1^{*\prime}\beta^* - Y_2^{*\prime}\delta^*$, where $Y^* \stackrel{\text{def}}{=} (Y_1^*, Y_2^*)$ with Y_1^* the outcome variable and Y_2^* the vector of endogenous regressors; $X^* \stackrel{\text{def}}{=} (X_1^*, X_2^*)$ with X_1^* the exogenous regressors, i. e. the 'included instruments', and X_2^* the 'excluded instruments' for Y_2^*; and $\theta^* \stackrel{\text{def}}{=} (\alpha^*, \beta^*, \delta^*)$. If all the regressors are endogenous, then X_1^* is the empty vector and the definition of θ^* has to be adjusted accordingly by dropping β^*. Similarly, for non-linear regression models, $g(Z^*, \theta^*) \stackrel{\text{def}}{=} Y_1^* - \psi(Y_2^*, X_1^*, \theta^*)$, where the non-linear function $\psi(Y_2^*, X_1^*, \cdot)$ is known up to θ^*. Multivariate extensions include systems of equations or transformation models, linear or non-linear, of the form $g(Z^*, \theta^*) = \varepsilon^*$, where g is a vector of known functions and the identifying assumption is that $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^* \mid X^*] = 0$ $P^*_{X^*}$-a.s.. Several examples of econometric models defined via conditional moment restrictions may be found in Newey (1993, Section 3), Pagan and Ullah (1999, Chapter 3), and Wooldridge (2010).

1.2.2 Variable probability sampling

Instead of observing Z^* directly from the target population, we possess a random vector $Z \stackrel{\text{def}}{=} (Y, X)$ that is collected by variable probability (VP) sampling, also known as Bernoulli sampling. For more on VP and other stratified sampling schemes, cf., e. g. DeMets and Halperin (1977), Manski and Lerman (1977), Holt et al. (1980), Cosslett (1981a, 1981b, 1991, 1993), Manski and McFadden (1981), Jewell (1985), Quesenberry and Jewell (1986), Scott and Wild (1986), Kalbfleisch and Lawless (1988), Bickel and Ritov (1991), Imbens (1992), Imbens and Lancaster (1996), Deaton (1997),

[1] If X^* is constant $P^*_{X^*}$-a.s., then there is no conditioning and (1.2.1) reduces to a system of unconditional moment equalities. These models are studied in Tripathi (2011a, 2011b).

[2] Similar notation, but without the '$*$' superscript, applies to the random variables and probability measures in the realised population.

Wooldridge (1999, 2001), Butler (2000), Bhattacharya (2005, 2007), Hirose (2007), Hirose and Lee (2008), Tripathi (2011a, 2011b), and Severini and Tripathi (2013).

Let the support of Z^*, denoted by $\mathrm{supp}(Z^*)$, be partitioned into L non-empty disjoint strata $\mathbb{C}_1, \ldots, \mathbb{C}_L$. In VP sampling, typically used when data is collected by telephone surveys, an observation is first drawn randomly from the target population. If it lies in stratum $\mathbb{C}_l$, it is retained with known probability $p_l \in (0, 1]$. If it is discarded, all information about the observation is lost. Hence, instead of observing a random vector Z^* drawn from the target distribution $P^* \stackrel{\text{def}}{=} \mathrm{Law}(Z^*)$, we observe a random vector Z, with $\mathrm{supp}(Z) = \mathrm{supp}(Z^*)$, drawn from the realised distribution $P \stackrel{\text{def}}{=} \mathrm{Law}(Z)$ given by[3]

$$P(Z \in B) \stackrel{\text{def}}{=} \sum_{l=1}^{L} \frac{p_l}{b^*} \int_B \mathbb{1}_{\mathbb{C}_l}(z)\, dP^*(z), \qquad B \in \mathscr{B}(\mathbb{R}^{\dim Z^*}), \tag{1.2.2}$$

where $\mathscr{B}(\mathbb{R}^{\dim Z^*})$ is the Borel sigma-field of $\mathbb{R}^{\dim Z^*}$, $b^* \stackrel{\text{def}}{=} \sum_{l=1}^{L} p_l Q_l^*$, and $Q_l^* \stackrel{\text{def}}{=} P^*(Z^* \in \mathbb{C}_l) > 0$ denotes the probability that a randomly chosen observation from the target population lies in the lth stratum.

Since Q_l^* represents the probability mass of the lth stratum in the target population, the Q_l^*'s are popularly called 'aggregate shares'. The aggregate shares, which add up to one, i. e. $\sum_{l=1}^{L} Q_l^* = 1$, are unknown parameters of interest to be estimated along with the structural parameter θ^*. The parameter b^* also has a practical interpretation; namely, it is the probability that an observation drawn from the target population during the sampling process is ultimately retained in the sample.

It is immediate from (1.2.2) that the density of P, with respect to any measure on $\mathscr{B}(\mathbb{R}^{\dim Z^*})$ that dominates P^*, is given by

$$\begin{aligned} dP(z) &\stackrel{\text{def}}{=} \sum_{l=1}^{L} \frac{p_l}{b^*} \mathbb{1}_{\mathbb{C}_l}(z) dP^*(z) && (z \in \mathbb{R}^{\dim Z^*}) \\ &= \frac{b(z)}{b^*} dP^*(z), && (1.2.3) \end{aligned}$$

where $b(z) \stackrel{\text{def}}{=} \sum_{l=1}^{L} p_l \mathbb{1}_{\mathbb{C}_l}(z)$. Following Imbens and Lancaster (1996, p. 296), $b(\cdot)/b^*$ is referred to as a bias function because it determines the selection bias due to stratified sampling, i. e. the extent to which P differs from P^*. For instance, it is easy to see that if the sampling probabilities $p_1, \ldots, p_L$ are all equal, then there is no selection bias, i. e. $P = P^*$, because $b(\cdot)/b^* = 1$ irrespective of the values taken by the aggregate shares.

The marginal density of X is given by

$$\begin{aligned} dP_X(x) &\stackrel{\text{def}}{=} \int_{y \in \mathbb{R}^{\dim Y^*}} dP(y, x) && (x \in \mathbb{R}^{\dim X^*}) \\ &= \int_{y \in \mathbb{R}^{\dim Y^*}} \frac{b(y, x)}{b^*}\, dP^*_{Y^*|X^*=x}(y)\, dP^*_{X^*}(x) && ((1.2.3)) \\ &= \frac{\gamma^*(x)}{b^*}\, dP^*_{X^*}(x), && (1.2.4) \end{aligned}$$

where $\gamma^*(x) \stackrel{\text{def}}{=} \mathbb{E}_{P^*_{Y^*|X^*}}[b(Y^*, x) \mid X^* = x]$. Throughout the paper, we maintain the

[3] Cf. Severini and Tripathi (2013, Appendix H) for a short proof of (1.2.2).

assumption that $\gamma^* > 0$ on $\operatorname{supp}(X^*)$.[4] Under this condition, the probability distributions P_X and $P^*_{X^*}$ are mutually absolutely continuous, which we denote by writing $P^*_{X^*} \ll P_X \ll P^*_{X^*}$.

Since $\operatorname{supp}(Y, X) = \operatorname{supp}(Y^*, X^*)$ and $\gamma^* > 0$ on $\operatorname{supp}(X^*)$, the conditional density of $Y \mid X$ is given by

$$\begin{aligned} dP_{Y|X=x}(y) &\stackrel{\text{def}}{=} \frac{dP(y,x)}{dP_X(x)} && ((y,x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)) \\ &= \frac{b(y,x)}{\gamma^*(x)}\, dP^*_{Y^*|X^*=x}(y), && (1.2.5) \end{aligned}$$

where (1.2.5) follows from (1.2.3) and (1.2.4).

By (1.2.5), $dP_{Y|X=x}(y) = dP^*_{Y^*|X^*=x}(y)$ if and only if $b(y,x) = \gamma^*(x)$ for all $(y,x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$. However, as discussed subsequently, the condition $b(y,x) = \gamma^*(x)$ holds only in a special case. Therefore, in general, $dP_{Y|X} \neq dP^*_{Y^*|X^*}$. Consequently, estimating (1.2.1) using the realised sample without accounting for the fact that it was obtained by stratified sampling, i. e. ignoring stratification, will generally not lead to a consistent estimator of θ^*.

1.2.3 Identification

In contrast to some other stratified sampling schemes (Tripathi, 2011b, Sections 3.1 and 4.1), identification, i. e. uniqueness, of θ^* cannot be lost because of VP sampling. To see this, begin by recalling that the assumption that $\gamma^* > 0$ on $\operatorname{supp}(X^*)$ implies that the distributions P_X and $P^*_{X^*}$ are mutually absolutely continuous. Hence,

$$\begin{aligned} (1.2.1) &\iff \mathbb{E}_{P^*_{Y^*|X^*}}[g(Y^*, x, \theta^*) \mid X^* = x] = 0 \quad \text{for } P^*_{X^*}\text{-a.a. } x \in \operatorname{supp}(X^*) \\ &\iff \gamma^*(x)\mathbb{E}_{P_{Y|X}}\Big[\frac{g(Y,x,\theta^*)}{b(Y,x)} \;\Big|\; X = x\Big] = 0 \quad \text{for } P^*_{X^*}\text{-a.a. } x \in \operatorname{supp}(X^*) \quad ((1.2.5)) \\ &\iff P^*_{X^*}\Big\{x \in \operatorname{supp}(X^*) : \mathbb{E}_{P_{Y|X}}\Big[\frac{g(Y,x,\theta^*)}{b(Y,x)} \;\Big|\; X = x\Big] \neq 0\Big\} = 0 \qquad (\gamma^* > 0) \\ &\iff P_X\Big\{x \in \operatorname{supp}(X^*) : \mathbb{E}_{P_{Y|X}}\Big[\frac{g(Y,x,\theta^*)}{b(Y,x)} \;\Big|\; X = x\Big] \neq 0\Big\} = 0 \\ &\qquad\qquad (P^*_{X^*} \ll P_X \ll P^*_{X^*}) \\ &\iff \mathbb{E}_{P_{Y|X}}\Big[\frac{g(Y,x,\theta^*)}{b(Y,x)} \;\Big|\; X = x\Big] = 0 \quad \text{for } P_X\text{-a.a. } x \in \operatorname{supp}(X^*). \end{aligned}$$

Therefore, we have that

$$(1.2.1) \iff \mathbb{E}_{P_{Y|X}}\Big[\frac{g(Z,\theta^*)}{b(Z)} \;\Big|\; X\Big] = 0 \qquad P_X\text{-a.s.}. \tag{1.2.6}$$

Since $b(Z)$ does not depend on θ^*, the equivalence in (1.2.6) reveals that θ^* in (1.2.1) is uniquely defined if and only if θ^* in $\mathbb{E}_{P_{Y|X}}[g(Z,\theta^*)/b(Z) \mid X] = 0$ (P_X-a.s.) is uniquely defined. That is, any condition that leads to the identification of θ^* in (1.2.1) will also ensure identification of θ^* in the right hand side of (1.2.6) and vice-versa. To illustrate this, assume that the columns of the partial derivative $\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*) \mid X^*]$ are

[4] A sufficient condition for this is that $P^*_{Y^*|X^*}((Y^*,x) \in \mathbb{C}_l \mid X^* = x) > 0$ for each l and $x \in \operatorname{supp}(X^*)$.

linearly independent $P^*_{X^*}$-a.s.. As shown in Appendix 3.B (in Chapter 3), this condition is sufficient to ensure that θ^* is locally identified.[5] However, since b does not depend on θ (which implies that γ^* does not depend on θ), we have that

$$\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*) \mid X^* = x] \overset{(1.2.5)}{=} \gamma^*(x)\partial_\theta \mathbb{E}_{P_{Y|X}}\Big[\frac{g(Z,\theta^*)}{b(Z)} \;\Big|\; X = x\Big], \qquad x \in \operatorname{supp}(X^*).$$

Therefore, since $\gamma^* > 0$ on $\operatorname{supp}(X^*)$, the columns of $\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*) \mid X^*]$ are linearly independent $P^*_{X^*}$-a.s. if and only if the columns of $\partial_\theta \mathbb{E}_{P_{Y|X}}[g(Z,\theta^*)/b(Z) \mid X]$ are linearly independent P_X-a.s. (because P_X and $P^*_{X^*}$ are mutually absolutely continuous).

Since the identification of θ^* cannot be lost because of VP sampling, for the remainder of the paper, we maintain that θ^* is identified.

1.2.4 Endogenous and exogenous stratification

As noted by Wooldridge (1999, p. 1385), VP sampling is employed when it is cheaper to obtain information on a subset of variables in the target population. Hence, it may happen that in certain datasets only Y^* is stratified (endogenous stratification),[6] or only X^* is stratified (exogenous stratification), or both Y^* and X^* are stratified. To see how all these cases can be handled in a unified manner in our framework, let the support of Y^* be partitioned into J non-empty disjoint strata $\mathbb{A}_1, \ldots, \mathbb{A}_J$, and the support of X^* be partitioned into M non-empty disjoint strata $\mathbb{B}_1, \ldots, \mathbb{B}_M$. Then, since $\cup_{j=1}^J \mathbb{A}_j \times \cup_{m=1}^M \mathbb{B}_m = \cup_{j=1}^J \cup_{m=1}^M \mathbb{A}_j \times \mathbb{B}_m$,

$$\operatorname{supp}(Y^*, X^*) = \begin{cases} \cup_{j=1}^J \cup_{m=1}^M \mathbb{A}_j \times \mathbb{B}_m & \text{if both } Y^* \text{ and } X^* \text{ are stratified} \\ \cup_{j=1}^J (\mathbb{A}_j \times \operatorname{supp}(X^*)) & \text{if only } Y^* \text{ is stratified} \\ \cup_{m=1}^M (\operatorname{supp}(Y^*) \times \mathbb{B}_m) & \text{if only } X^* \text{ is stratified.} \end{cases}$$

Therefore, if both Y^* and X^* are stratified, then $\operatorname{supp}(Z^*) = \cup_{l=1}^L \mathbb{C}_l$ with $L = JM$ and each $\mathbb{C}_l = \mathbb{A}_j \times \mathbb{B}_m$ for some $(j,m) \in \{1,\ldots,J\} \times \{1,\ldots,M\}$. This is the most general case for which $P_{Y|X}$ is given by (1.2.5).[7]

In contrast, simplifications occur if the stratification is endogenous or exogenous. For instance, if only Y^* is stratified, then $\operatorname{supp}(Z^*) = \cup_{l=1}^L \mathbb{C}_l$ with $L = J$ and $\mathbb{C}_l = \mathbb{A}_l \times \operatorname{supp}(X^*)$, which implies that, for $(y,x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$,

$$b(y,x) = \sum_{l=1}^J p_l \mathbb{1}_{\mathbb{A}_l \times \operatorname{supp}(X^*)}(y,x) = \sum_{l=1}^J p_l \mathbb{1}_{\mathbb{A}_l}(y) =: b_{\text{endog}}(y).$$

Hence, by (1.2.5), we have that, for $(y,x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$,

$$\text{endogenous stratification} \implies dP_{Y|X=x}(y) = \frac{b_{\text{endog}}(y)}{\gamma^*_{\text{endog}}(x)}\, dP^*_{Y^*|X^*=x}(y), \tag{1.2.7}$$

where $\gamma^*_{\text{endog}}(x) \overset{\text{def}}{=} \mathbb{E}_{P^*_{Y^*|X^*}}[b_{\text{endog}}(Y^*) \mid X^* = x]$.

[5] The same condition leads to global identification of θ^* whenever $g(Z^*,\theta^*)$ is linear in θ^*.

[6] In the econometrics literature, stratification based on a finite set of response variables is often referred to as choice-based sampling.

[7] Unless mentioned otherwise, it is assumed throughout the paper that both Y^* and X^* are stratified.

If only X^* is stratified, then $\operatorname{supp}(Z^*) = \cup_{l=1}^{L} \mathbb{C}_l$ with $L = M$ and $\mathbb{C}_l = \operatorname{supp}(Y^*) \times \mathbb{B}_l$. Consequently, for $(y, x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$,

$$b(y,x) = \sum_{l=1}^{M} p_l \mathbb{1}_{\operatorname{supp}(Y^*)\times\mathbb{B}_l}(y,x) = \sum_{l=1}^{M} p_l \mathbb{1}_{\mathbb{B}_l}(x) =: b_{\text{exog}}(x),$$

which implies that $\gamma^*_{\text{exog}}(x) \stackrel{\text{def}}{=} \mathbb{E}_{P^*_{Y^*|X^*}}[b_{\text{exog}}(X^*) \mid X^* = x] = b_{\text{exog}}(x)$. Hence, by (1.2.5),

$$\text{exogenous stratification} \implies dP_{Y|X=x}(y) = dP^*_{Y^*|X^*=x}(y) \tag{1.2.8}$$

for $(y, x) \in \operatorname{supp}(Y^*) \times \operatorname{supp}(X^*)$. Consequently, exogenous stratification can be ignored, at least as far as consistent estimation is concerned. However, as the following example demonstrates, ignoring endogenous stratification does not lead to a consistent estimator.

Example 1.2.1 (Linear regression with exogenous regressors)**.** Consider the linear regression model $Y^* = \tilde{X}^{*\prime}\theta^* + \varepsilon^*$, where $\tilde{X}^* \stackrel{\text{def}}{=} (1, X^*)$. Assume that the regressors are exogenous with respect to the model error in the target population, i. e. $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^* \mid X^*] = 0$ $P^*_{X^*}$-a.s..

Suppose that only Y^* is stratified. If we ignore the fact that the data were collected by VP sampling and simply regress the observed Y on the observed X and the constant regressor, then θ^* cannot be consistently estimated by the least-squares (LS) estimator. Indeed, letting $\hat{\theta}_{\text{LS}}$ denote the LS estimator obtained by regressing Y on $\tilde{X} \stackrel{\text{def}}{=} (1, X)$, we have that

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \hat{\theta}_{\text{LS}} &= \operatorname*{plim}_{n\to\infty} \left(n^{-1} \sum_{j=1}^{n} \tilde{X}_j \tilde{X}_j' \right)^{-1} \left(n^{-1} \sum_{j=1}^{n} \tilde{X}_j Y_j \right) \\
&= (\mathbb{E}_{P_X} \tilde{X}\tilde{X}')^{-1} (\mathbb{E}_P \tilde{X} Y) \\
&= (\mathbb{E}_{P_X} \tilde{X}\tilde{X}')^{-1} \big(\mathbb{E}_{P_X} \tilde{X} \mu(X) \big),
\end{aligned} \tag{1.2.9}$$

where $\mu(X) \stackrel{\text{def}}{=} \mathbb{E}_{P_{Y|X}}[Y \mid X]$. But, $\mathbb{E}_{P_X} \tilde{X}\tilde{X}' \stackrel{(1.2.4)}{=} \mathbb{E}_{P^*_{X^*}} \gamma^*_{\text{endog}}(X^*) \tilde{X}^* \tilde{X}^{*\prime} / b^*$ and

$$\begin{aligned}
\mu(x) &\stackrel{\text{def}}{=} \mathbb{E}_{P_{Y|X}}[Y \mid X = x] && (x \in \operatorname{supp}(X^*)) \\
&= \frac{1}{\gamma^*_{\text{endog}}(x)} \mathbb{E}_{P^*_{Y^*|X^*}}[Y^* b_{\text{endog}}(Y^*) \mid X^* = x] && ((1.2.7)) \\
&= \frac{1}{\gamma^*_{\text{endog}}(x)} \mathbb{E}_{P^*_{Y^*|X^*}}[(\tilde{X}^{*\prime}\theta^* + \varepsilon^*) b_{\text{endog}}(Y^*) \mid X^* = x] \\
&= \tilde{x}'\theta^* + \frac{1}{\gamma^*_{\text{endog}}(x)} \mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^* b_{\text{endog}}(Y^*) \mid X^* = x]. && (1.2.10)
\end{aligned}$$

Hence, writing (1.2.9) in terms of $P^*_{X^*}$,

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty}\hat{\theta}_{\mathrm{LS}} &= \left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{endog}}(X^*)}{b^*}\tilde{X}^*\tilde{X}^{*\prime}\right)^{-1}\left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{endog}}(X^*)}{b^*}\tilde{X}^*\mu(X^*)\right) && \text{((1.2.9) \& (1.2.4))}\\
&\overset{(1.2.10)}{=} \left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{endog}}(X^*)}{b^*}\tilde{X}^*\tilde{X}^{*\prime}\right)^{-1}\\
&\qquad\times\left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{endog}}(X^*)}{b^*}\tilde{X}^*\Big[\tilde{X}^{*\prime}\theta^*+\frac{1}{\gamma^*_{\mathrm{endog}}(X^*)}\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^* b_{\mathrm{endog}}(Y^*)\mid X^*]\Big]\right)\\
&= \theta^*+\big(\mathbb{E}_{P^*_{X^*}}\gamma^*_{\mathrm{endog}}(X^*)\tilde{X}^*\tilde{X}^{*\prime}\big)^{-1}\big(\mathbb{E}_{P^*}\tilde{X}^*\varepsilon^* b_{\mathrm{endog}}(Y^*)\big)\\
&\neq \theta^*,
\end{aligned}$$

because $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^*\mid X^*]=0$ ($P^*_{X^*}$-a.s.) does not imply that $\mathbb{E}_{P^*}\tilde{X}^*\varepsilon^* b_{\mathrm{endog}}(Y^*)=0$.

If, however, stratification is exogenous, then

$$\begin{aligned}
\mu(x)=\mathbb{E}_{P_{Y|X}}[Y\mid X=x] &\overset{(1.2.8)}{=} \mathbb{E}_{P^*_{Y^*|X^*}}[Y^*\mid X^*=x] && (x\in\operatorname{supp}(X^*))\\
&= \mathbb{E}_{P^*_{Y^*|X^*}}[\tilde{X}^{*\prime}\theta^*+\varepsilon^*\mid X^*=x]\\
&= \tilde{x}'\theta^*. && \text{(1.2.11)}
\end{aligned}$$

Hence, ignoring exogenous stratification does not affect the consistency of $\hat{\theta}_{\mathrm{LS}}$ because

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty}\hat{\theta}_{\mathrm{LS}} &= (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}\big(\mathbb{E}_{P_X}\tilde{X}\mu(X)\big) && \text{((1.2.9))}\\
&= \left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{exog}}(X^*)}{b^*}\tilde{X}^*\tilde{X}^{*\prime}\right)^{-1}\left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{exog}}(X^*)}{b^*}\tilde{X}^*\mu(X^*)\right) && \text{((1.2.4))}\\
&= \left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{exog}}(X^*)}{b^*}\tilde{X}^*\tilde{X}^{*\prime}\right)^{-1}\left(\mathbb{E}_{P^*_{X^*}}\frac{\gamma^*_{\mathrm{exog}}(X^*)}{b^*}\tilde{X}^*\tilde{X}^{*\prime}\theta^*\right) && \text{((1.2.11))}\\
&= \theta^*.
\end{aligned}$$

However, as shown subsequently (cf. Example 1.3.1), $\hat{\theta}_{\mathrm{LS}}$ is not asymptotically efficient. Hence, ignoring exogenous stratification does not affect the consistency of the LS estimator,[8] but it does affect its efficiency. □

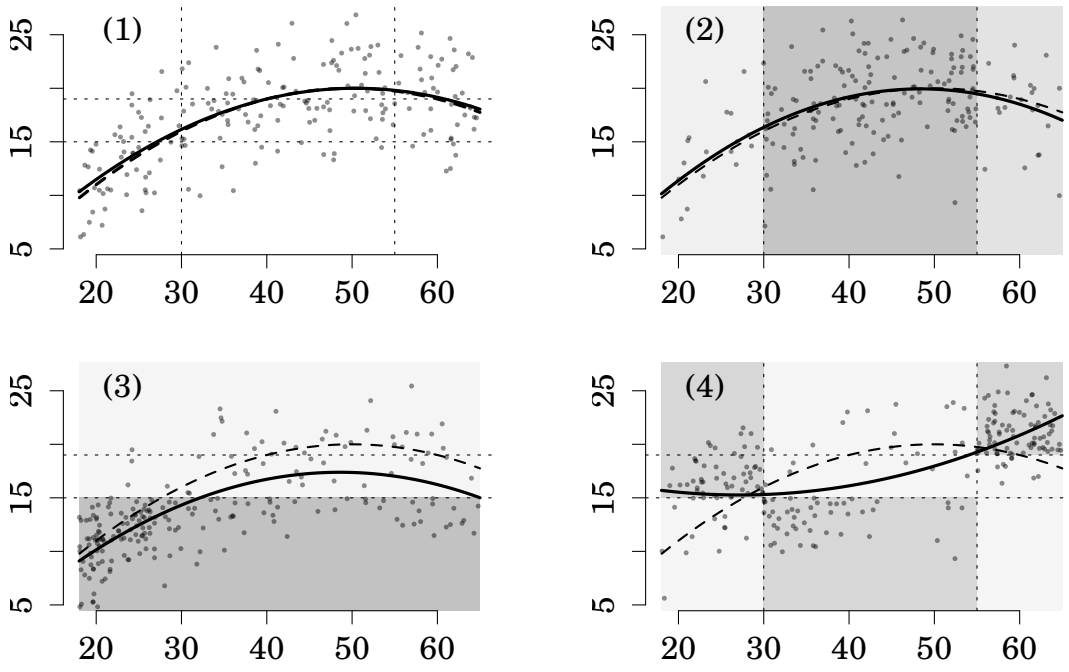
Example 1.2.2 (Implications of stratification, visualised)**.** Consider the structural model $Y^*=\delta_0+\delta_1X^*+\delta_2X^{*2}+U^*$ with $\mathbb{E}(U^*\mid X^*)=0$, where X^* denotes the age of a working individual, and Y^* denotes their monthly wage. This quadratic specification of earnings as a function of time-related variables (such as age or experience) is arguably the most popular version of the Mincer equation accounting for non-linear individual wage change over time (Heckman et al., 2003). Let $\delta^*\overset{\text{def}}{=}(\delta_0,\delta_1,\delta_2)\overset{\text{def}}{=}(-5,1,-0.01)$, let X^* be uniformly distributed between 18 and 65, and let $U^*\mid X^*\sim\mathcal{N}(0,9)$. Let the strata of the endogenous income be defined as follows: $\mathbb{A}_1=(-\infty,15)$, $\mathbb{A}_2=[15,19)$, $\mathbb{A}_3=[19,+\infty)$ (low-, middle-, and high-income individuals). Let the strata of the exogenous age be defined as follows: $\mathbb{B}_1=[18,30)$, $\mathbb{B}_2=[30,55)$, $\mathbb{B}_3=[55,65]$ (younger, middle-aged, and older working individuals). Consider four sampling schemes devised by a data collection agency:

[8] Tripathi (2011b) shows that in unconditional-moment-restriction models, even exogenous stratification cannot be ignored.

1. Random sampling with probability 0.1;
2. VP sampling with middle-aged individuals over-sampled and young individuals under-sampled ($(p_{\mathbb{B}_1}, p_{\mathbb{B}_2}, p_{\mathbb{B}_3}) = (0.0325, 0.1463, 0.0650)$);
3. VP sampling with lower-income individuals over-sampled and middle- and high-income individuals under-sampled ($(p_{\mathbb{A}_1}, p_{\mathbb{A}_2}, p_{\mathbb{A}_3}) = (0.2913, 0.0324, 0.0324)$);
4. VP sampling with rich younger, middle-income younger, rich older, and poor middle-aged individuals over-sampled 10 times compared to the remaining strata ($p_{\mathbb{C}_1} = 0.3160$, $p_{\mathbb{C}_2} = 0.0316$, where $\mathbb{C}_1 \stackrel{\text{def}}{=} (\mathbb{A}_2 \times \mathbb{B}_1) \cup (\mathbb{A}_3 \times \mathbb{B}_1) \cup (\mathbb{A}_1 \times \mathbb{B}_2) \cup (\mathbb{A}_3 \times \mathbb{B}_3)$ and $\mathbb{C}_2 \stackrel{\text{def}}{=} \mathbb{R}^2 \setminus \mathbb{C}_1$).

In all four cases, numerical evaluation shows that $b^* \approx 0.1$ (up to three significant digits), which means that the unconditional retention probability is the same. Thus, on average, every 10^{th} observation from the population is selected into the sample. However, estimation results in these four samples are qualitatively different.

Figure 1.1: Impact of stratification on $\mathrm{Law}(Y \mid X)$



Four synthetic data sets with individuals' ages (horizontal axis) and wages (vertical axis) were generated according to the sampling schemes (1)–(4). Dotted lines delimit the strata. The dashed curve shows the true $\mathrm{Law}(Y^* \mid X^*)$ (the same in all 4 cases). The solid curves shows least-squares model fits. The sample size is approximately 200 in all cases. The darker the shade of a stratum, the higher the retention probability.

Simulated data sets coming from schemes (1)–(4) are shown in Figure 1.1. If the data are collected via (1), then $P = P^*$, and simple least-squares estimation is consistent. In the case of exogenous stratification (2), $P \neq P^*$; however, $dP_{Y|X=x}(y) = dP^*_{Y^*|X^*=x}(y)$, and the probability limit of the LS estimator is the same, which is evidenced by the closeness of the true law and predicted regression line in panel (2). In the case of endogenous stratification (3), due to over-sampling of low-income individuals, the estimated $\mathbb{E}(Y \mid X = x)$ is strictly smaller than $\mathbb{E}(Y^* \mid X^* = x)$, and the point estimates are far away from the true values (the relative absolute bias is 23–25% for δ_1 and δ_2). Finally, case (4) represents the most egregious case of distortion, in which the entire joint distribution of (Y, X) is represented by the regions of $\mathbb{R}^2$ where the centres of

mass of each vertical strip form a parabola branching upwards. Consequently, a model estimated on data collected in this manner predicts the wrong sign of δ_2, which changes its economic interpretation. E. g. the researcher might mistakenly conclude that the expected wage starts increasing after attaining the age of 27 and continues increasing until retirement, instead of realising that the correct inference is 'the estimated age of peak expected wage is 50', which may have implications for the pension scheme.

This example demonstrates that if Y^* or both Y^* and X^* are stratified, then the resulting data set will be primarily shaped by the sampling probabilities. In other words, the distortions appearing in $\mathrm{Law}(Y \mid X)$ may obscure the true $\mathrm{Law}(Y^* \mid X^*)$ due to the regression line gravitating towards over-represented strata depending on Y^*.

1.3 Inference

1.3.1 Related literature and our contribution

There is a large literature on estimation and testing models using data collected by various types of stratified sampling schemes; cf. the papers cited at the beginning of Section 1.2.2, and the references therein. In this section, we briefly describe only some of the works that consider VP sampling.

Earlier papers in the literature on estimating models with conditioning variables assume that $P^*_{Y^*|X^*}$ is known up to a finite-dimensional parameter; only $P^*_{X^*}$ is left completely unspecified. E. g. a well-known application of VP sampling can be found in Hausman and Wise (1981). Imbens and Lancaster (1996) extend the maximum likelihood approach of Hausman and Wise to a moment-based methodology that allows for VP sampling, mixed response variables, and stratification on exogenous covariates. Regression under VP sampling and a parametric $P^*_{Y^*|X^*}$ has also been investigated. E. g. Jewell (1985) and Quesenberry and Jewell (1986) propose iterative estimators of regression coefficients under VP sampling without imposing normality or independence, though they do not provide any asymptotic theory for their estimators.

The papers described above impose strong conditions on the distribution of $Y^* \mid X^*$. Exceptions include Wooldridge (1999) and Tripathi (2011b), who leave both $P^*_{Y^*|X^*}$ and $P^*_{X^*}$ completely unspecified. Wooldridge provides asymptotic theory for M-estimation under VP sampling for a model defined in terms of a set of just-identified unconditional moment equalities, whereas Tripathi considers optimal generalised method of moments (GMM) estimation in unconditional moment restriction models that allow for the parameter of interest to be over-identified. The major difference between (1.2.1) and the models in the papers of Wooldridge and Tripathi is that (1.2.1) is a conditional moment restriction, whereas the moment conditions in the aforementioned papers are all unconditional. Therefore, (1.2.1) nests the moment conditions of Wooldridge and Tripathi as a special case.

In this paper, we show how to efficiently estimate θ^* and the aggregate shares using a smoothed-empirical-likelihood-based approach. The results presented here answer the question posed in Wooldridge (1999, p. 1402) by providing efficiency bounds for models with conditional moment restrictions under VP sampling and showing that these bounds are attainable.

Furthermore, the results in this paper are also directly applicable to a class of 'biased sampling' problems. To see this, recall that the phenomenon where the realised probability distribution P differs from the target probability distribution P^* is

generically referred to as selection bias.[9] It is useful to note that the class of problems that can be handled when selection is modelled using (1.2.3) includes more than just those involving stratified sampling. For instance, consider the so-called 'length biased sampling' problem where the probability of observing a random variable is proportional to its 'size'. E. g. economists are more likely to observe longer unemployment spells if they are sampled at a randomly chosen time. Similarly, as Owen (2001, p. 127) points out, if internet log files are sampled randomly, then longer sessions are likely to be over-represented. It is useful to examine length biased sampling in the context of VP sampling because in length-biased sampling, we have

$$dP(z) \stackrel{\text{def}}{=} \frac{\|z\|}{\mathbb{E}_{P^*}\|Z^*\|}\, dP^*(z), \qquad z \in \mathbb{R}^{\dim Z^*},$$

where $\|\cdot\|$ is the Euclidean norm. That is, length biased sampling can be expressed as (1.2.3) with $b(z) \stackrel{\text{def}}{=} \|z\|$ and $b^* \stackrel{\text{def}}{=} \mathbb{E}_{P^*}\|Z^*\|$. Therefore, with only minor notational changes, the results obtained in this paper can be extended to length biased sampling as well.

Length biased sampling has been extensively studied for the parametric case, i. e. where dP^* is specified up to a finite-dimensional parameter. Cf., e. g. Patil and Rao (1977, p. 1978), Bickel et al. (1993, Section 4.4), and Owen (2001, Chapter 6). As far as a non-parametric treatment of length biased sampling is concerned, Vardi (1982) deals with the case when P^* is unknown. Vardi assumes that both P^* and P can be sampled with positive probability. Using two independent samples (one each from P^* and P), he shows how to construct the non-parametric maximum likelihood estimators (NPMLE) of P^* and P, and also obtains their asymptotic distributions. Vardi (1985) and Gill et al. (1988) provide conditions for the existence and uniqueness of the NPMLE of P^* in a general setup when more than two independent samples from F^* and F are available. These papers concentrate on the distributions P^* and P; there are no other parameters to estimate. Qin (1993) uses the empirical likelihood approach to construct a non-parametric likelihood ratio confidence interval for $\theta^* \stackrel{\text{def}}{=} \mathbb{E}_{P^*}Z^*$, i. e. a just-identified unconditional moment equality, using an independent sample from P^* and P. El-Barmi and Rothmann (1998) generalise Qin's treatment to handle models with over-identified unconditional moment restrictions of the form $\mathbb{E}_{P^*}g(Z, \theta^*) = 0$. They also obtain efficient estimators of P^* and P. However, they do not consider the testing of over-identifying restrictions.

1.3.2 Efficiency bounds

The efficiency bounds for estimating θ^* and related functionals have been derived in Severini and Tripathi (2013, Section 14.3). In this section, we describe some of these bounds and discuss their salient features. Construction of estimators that achieve these bounds is considered in the next section.

For the remainder of the paper, let $\rho_1(Z, \theta) \stackrel{\text{def}}{=} g(Z, \theta)/b(Z)$. Since the right hand side of (1.2.6) is a conditional moment equality with respect to the realised conditional distribution $P_{Y|X}$, the efficiency bound for θ^* follows from Chamberlain (1987). Namely,

[9] Hence, for the LS estimator in Example 1.2.1, one can say that it is inconsistent because of selection bias due to endogenous stratification, whereas exogenous stratification does not lead to any selection bias.

the efficiency bound for estimating θ^* is given by[10]

$$\text{l.b.}(\theta^*) \stackrel{\text{def}}{=} \big(\mathbb{E}_{P_X} D'(X) V_1^{-1}(X) D(X)\big)^{-1}, \tag{1.3.1}$$

where $D(X) \stackrel{\text{def}}{=} \partial_\theta \mathbb{E}_{P_{Y|X}}[\rho_1(Z,\theta^*) \mid X]$ and $V_1(X) \stackrel{\text{def}}{=} \mathbb{E}_{P_{Y|X}}[\rho_1(Z,\theta^*)\rho_1'(Z,\theta^*) \mid X]$.

The efficiency bound in (1.3.1), given as a functional of the realised distribution P, can be used to determine whether an estimator of θ^* is semi-parametrically efficient by comparing its asymptotic variance with $\text{l.b.}(\theta^*)$. However, as the moment condition model (1.2.1) is specified in terms of the target distribution P^*, in order to answer questions such as how the efficiency bound for θ^* changes if stratification is purely endogenous (or purely exogenous) or if the error term in a regression model is conditionally homoskedastic in the target population, it is helpful to rewrite (1.3.1) in terms of P^*. To do so, observe that, by (1.2.5), we have

$$\begin{aligned} D(x) &= \frac{1}{\gamma^*(x)} \partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*) \mid X^* = x], \qquad x \in \text{supp}(X^*), \\ V_1(x) &= \frac{1}{\gamma^*(x)} \mathbb{E}_{P^*_{Y^*|X^*}}\Big[\frac{g(Z^*,\theta^*)g'(Z^*,\theta^*)}{b(Y^*,x)} \;\Big|\; X^* = x\Big]. \end{aligned} \tag{1.3.2}$$

Hence, by (1.2.4) and (1.3.2), the efficiency bound in (1.3.1) can be written as

$$\text{l.b.}(\theta^*) = b^*\Big(\mathbb{E}_{P^*_{X^*}}\big(\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*) \mid X^*]\big)' V_b^{*-1}(X^*)\big(\partial_\theta \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*) \mid X^*]\big)\Big)^{-1}, \tag{1.3.3}$$

where $V_b^*(X^*) \stackrel{\text{def}}{=} \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*)g'(Z^*,\theta^*)/b(Z^*) \mid X^*]$.

We can use (1.3.3) to determine the efficiency bound for estimating θ^* if stratification is purely endogenous or purely exogenous. For instance, the efficiency bound when only Y^* is stratified follows from (1.3.3) on replacing $b(Z^*)$ in the definition of $V_b^*(X^*)$ with $b_{\text{endog}}(Y^*)$. Similarly, the bound when only X^* is stratified follows from (1.3.3) on replacing $b(Z^*)$ in the definition of $V_b^*(X^*)$ with $b_{\text{exog}}(X^*)$.

If there is no conditioning in (1.2.1), i. e. X^* is constant $P^*_{X^*}$-a.s., and $\dim g \geq \dim \theta^*$, then (1.3.3) reduces to the efficiency bound for estimating θ^* in unconditional moment restriction models when observations are collected by VP sampling (Severini & Tripathi, 2013, Section 14.2.1). At the other extreme, if there is no stratification, i. e. $L = 1 = p_1$ and $\mathbb{C}_1 = \text{supp}(Z^*)$, so that $Z^* = Z$ and $P^* = P$, then the efficiency bound in (1.3.3) becomes

$$\Big(\mathbb{E}\big(\partial_\theta \mathbb{E}[g(Z^*,\theta^*) \mid X^*]\big)'\big(\mathbb{E}[g(Z^*,\theta^*)g'(Z^*,\theta^*) \mid X^*]\big)^{-1}\big(\partial_\theta \mathbb{E}[g(Z^*,\theta^*) \mid X^*]\big)\Big)^{-1},$$

which is Chamberlain's 1987 bound for estimating θ^* in the absence of any selection.

The next example uses (1.3.3) to determine the efficiency bound for θ^* under various scenarios.

Example 1.3.1 (Example 1.2.1 contd.)**.** Here, $g(Z^*,\theta) = Y^* - \tilde{X}^{*\prime}\theta$ and the efficiency

[10] The abbreviation 'l.b.' stands for 'lower bound', because the efficiency bound is the greatest lower bound for the asymptotic variance of any $n^{1/2}$-consistent regular estimator.

bound for estimating θ^* is given by

$$\text{l.b.}(\theta^* \text{ in Example 1.2.1}) \overset{(1.3.3)}{=} b^* \Big(\mathbb{E}_{P^*_{X^*}} \frac{\tilde{X}^* \tilde{X}^{*\prime}}{V^*_b(X^*)} \Big)^{-1} \overset{(1.2.4),(1.2.5)}{=} \Big(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{\gamma^{*2}(X) V_1(X)} \Big)^{-1}. \tag{1.3.4}$$

If stratification is endogenous, then

$$\text{l.b.}(\theta^* \text{ in Example 1.2.1})\big|_{\text{endog. strat.}} \overset{(1.3.4)}{=} \Big(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{\gamma^{*2}_{\text{endog}}(X) V_{1,\text{endog}}(X)} \Big)^{-1},$$

where $V_{1,\text{endog}}(X) \overset{\text{def}}{=} \mathbb{E}_{P_{Y|X}}[(Y - \tilde{X}'\theta^*)^2 / b^2_{\text{endog}}(Y) \mid X]$.

In contrast, if stratification is exogenous then

$$\text{l.b.}(\theta^* \text{ in Example 1.2.1})\big|_{\text{exog. strat.}} \overset{(1.3.4)}{=} \Big(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{V_{1,\text{exog}}(X)} \Big)^{-1}, \tag{1.3.5}$$

where $V_{1,\text{exog}}(X) \overset{\text{def}}{=} \mathbb{E}_{P_{Y|X}}[(Y - \tilde{X}'\theta^*)^2 \mid X]$.

Recall from Example 1.2.1 that, under exogenous stratification, the LS estimator consistently estimates θ^*. Since $n^{1/2}(\hat{\theta}_{\text{LS}} - \theta^*)$ is asymptotically (as $n \to \infty$) normal with mean zero and variance $V_{\text{LS,exog}} \overset{\text{def}}{=} (\mathbb{E}_{P_X} \tilde{X}\tilde{X}')^{-1} \big(\mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X) \big) (\mathbb{E}_{P_X} \tilde{X}\tilde{X}')^{-1}$, an application of a matrix version of the Cauchy-Schwarz inequality reveals that

$$\text{l.b.}(\theta^* \text{ in Example 1.2.1})\big|_{\text{exog. strat.}} \leq_L V_{\text{LS,exog}},$$

where $\leq_L$ is the usual (Löwner) order on the set of symmetric matrices.[11] Therefore, under exogenous stratification, the LS estimator is consistent but not asymptotically efficient. However, if stratification is exogenous and ε^* is conditionally homoskedastic in the target population, then (1.3.5) and (1.A.3) imply that the LS estimator is asymptotically efficient.

Even under endogenous stratification, it is not difficult to obtain an estimator of θ^* that is consistent but asymptotically inefficient. To see this, assume that only Y^* is stratified. Then,

$$\begin{aligned} &\mathbb{E}_{P^*_{Y^*|X^*}}[Y^* - \tilde{X}^{*\prime}\theta^* \mid X^*] = 0 \quad P^*_{X^*}\text{-a.s.} \\ &\qquad \iff \mathbb{E}_{P_{Y|X}}\Big[\frac{Y - \tilde{X}'\theta^*}{b_{\text{endog}}(Y)} \;\Big|\; X \Big] = 0 \quad P_X\text{-a.s.} \qquad \text{((1.2.6) \& (1.2.7))} \\ &\qquad \implies \mathbb{E}_P \tilde{X} \Big[\frac{Y - \tilde{X}'\theta^*}{b_{\text{endog}}(Y)} \Big] = 0. \end{aligned}$$

Hence, it is straightforward to show that the GMM estimator

$$\hat{\theta}_{\text{GMM,endog}} \overset{\text{def}}{=} \Big(\sum_{j=1}^n \frac{\tilde{X}_j \tilde{X}_j'}{b_{\text{endog}}(Y_j)} \Big)^{-1} \Big(\sum_{j=1}^n \frac{\tilde{X}_j Y_j}{b_{\text{endog}}(Y_j)} \Big)$$

[11] Namely, $M_1 \leq_L M_2$ for symmetric matrices M_1, M_2 means that $M_1 - M_2$ is negative semidefinite.

is consistent for θ^*.[12] However, $\hat{\theta}_{\text{GMM,endog}}$ is not asymptotically efficient because its asymptotic variance is

$$V_{\text{GMM,endog}} \stackrel{\text{def}}{=} \big(\mathbb{E}_P \tilde{X}\tilde{X}'/b_{\text{endog}}(Y)\big)^{-1}\big(\mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{endog}}(X)\big)\big(\mathbb{E}_P \tilde{X}\tilde{X}'/b_{\text{endog}}(Y)\big)^{-1}$$

but

$$\text{l.b.}(\theta^* \text{ in Example 1.2.1})\big|_{\text{endog. strat.}} \leq_L V_{\text{GMM,endog}}.$$

Analogous to $\hat{\theta}_{\text{GMM,endog}}$, the GMM estimator under exogenous stratification is

$$\hat{\theta}_{\text{GMM,exog}} \stackrel{\text{def}}{=} \bigg(\sum_{j=1}^n \frac{\tilde{X}_j \tilde{X}_j'}{b_{\text{exog}}(X_j)}\bigg)^{-1}\bigg(\sum_{j=1}^n \frac{\tilde{X}_j Y_j}{b_{\text{exog}}(X_j)}\bigg),$$

which is also not asymptotically efficient because its asymptotic variance is $V_{\text{GMM,exog}} \stackrel{\text{def}}{=} \big(\mathbb{E}_{P_X} \tilde{X}\tilde{X}'/b_{\text{exog}}(X)\big)^{-1}\big(\mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X)/b^2_{\text{exog}}(X)\big)\big(\mathbb{E}_{P_X} \tilde{X}\tilde{X}'/b_{\text{exog}}(X)\big)^{-1}$ but

$$\text{l.b.}(\theta^* \text{ in Example 1.2.1})\big|_{\text{exog. strat.}} \leq_L V_{\text{GMM,exog}}.$$

Constructing efficient estimators requires more effort. For instance, suppose that stratification is purely exogenous. Then, following Robinson (1987), it can be shown that the asymptotic variance of $\hat{\theta}_{\text{Robinson}} \stackrel{\text{def}}{=} \big(\sum_{j=1}^n \tilde{X}_j \tilde{X}_j'/\hat{\sigma}^2(X_j)\big)^{-1}\big(\sum_{j=1}^n \tilde{X}_j Y_j/\hat{\sigma}^2(X_j)\big)$ equals (1.3.5), where $\hat{\sigma}^2$ denotes a consistent estimator of $V_{1,\text{exog}}$. Hence, $\hat{\theta}_{\text{Robinson}}$ is an asymptotically efficient estimator of θ^* under exogenous stratification. A general approach, which can be used to construct efficient estimators irrespective of whether stratification is endogenous, exogenous, or both, is discussed in Section 1.3.3. □

Since the aggregate shares add up to one, it suffices to determine the efficiency bound for estimating $Q^*_{-L} \stackrel{\text{def}}{=} (Q^*_1, \ldots, Q^*_{L-1})_{(L-1)\times 1} \in (0,1)^{L-1}$. The aggregate shares are identified in the realised population by the moment condition

$$\mathbb{E}_P\Big[\frac{s(Z) - Q^*_{-L}}{b(Z)}\Big] = 0, \tag{1.3.6}$$

where $s(Z) \stackrel{\text{def}}{=} \big(\mathbb{1}_{\mathbb{C}_1}(Z), \ldots, \mathbb{1}_{\mathbb{C}_{L-1}}(Z)\big)_{(L-1)\times 1}$. The moment conditions in (1.3.6) modify accordingly if stratification is endogenous or exogenous; namely,

$$\begin{aligned}
\text{endog. strat.} \implies &\begin{cases} \mathbb{E}_{P_Y}\Big[\frac{s_{\text{endog}}(Y) - Q^*_{-L}}{b_{\text{endog}}(Y)}\Big] = 0 \\ s_{\text{endog}}(Y) \stackrel{\text{def}}{=} \big(\mathbb{1}_{\mathbb{A}_1}(Y), \ldots, \mathbb{1}_{\mathbb{A}_{J-1}}(Y)\big)_{(J-1)\times 1} \end{cases} \\
\text{exog. strat.} \implies &\begin{cases} \mathbb{E}_{P_X}\Big[\frac{s_{\text{exog}}(X) - Q^*_{-L}}{b_{\text{exog}}(X)}\Big] = 0 \\ s_{\text{exog}}(X) \stackrel{\text{def}}{=} \big(\mathbb{1}_{\mathbb{B}_1}(X), \ldots, \mathbb{1}_{\mathbb{B}_{M-1}}(X)\big)_{(M-1)\times 1}. \end{cases}
\end{aligned} \tag{1.3.7}$$

Let $\rho_2(Z, Q^*_{-L}) \stackrel{\text{def}}{=} \big(s(Z) - Q^*_{-L}\big)/b(Z)$, and $\Sigma_{12}(X) \stackrel{\text{def}}{=} \mathbb{E}_{P_{Y|X}}[\rho_1(Z,\theta^*)\rho_2'(Z, Q^*_{-L}) \mid X]$ be the conditional (on X) covariance between $\rho_1(Z,\theta^*)$ and $\rho_2(Z, Q^*_{-L})$. Then, under (1.2.1),

[12] The estimator $\hat{\theta}_{\text{GMM,endog}}$ is an example of an inverse probability weighted (IPW) estimator, which uses the weights $1/b_{\text{endog}}(Y_1), \ldots, 1/b_{\text{endog}}(Y_n)$ to correct the selection bias due to stratification by downward weighting the strata that are oversampled and upward weighting the strata that are undersampled.

the efficiency bound for estimating Q^*_{-L} is given by

$$\begin{aligned}\text{l.b.}(Q^*_{-L}) \stackrel{\text{def}}{=} b^{*2}\Big[&\operatorname{Var}_P\big(\rho_2(Z,Q^*_{-L})\big) - \big(\mathbb{E}_{P_X}\Sigma_{12}'(X)V_1^{-1}(X)\Sigma_{12}(X)\big)\\ &+ \big(\mathbb{E}_{P_X}\Sigma_{12}'(X)V_1^{-1}(X)D(X)\big)\big(\text{l.b.}(\theta^*)\big)\big(\mathbb{E}_{P_X}D'(X)V_1^{-1}(X)\Sigma_{12}(X)\big)\Big], \quad (1.3.8)\end{aligned}$$

where $\text{l.b.}(\theta^*)$ is the efficiency bound for estimating θ^* given in (1.3.1).

In the absence of (1.2.1), the efficiency bound for Q^*_{-L} is given by $b^{*2}\operatorname{Var}_P\big(\rho_2(Z,Q^*_{-L})\big)$, which follows from standard GMM theory applied to (1.3.6). Hence, estimating the aggregate shares in the presence of (1.2.1) leads to efficiency gains under endogenous stratification. There are no efficiency gains for estimating Q^*_{-L} under exogenous stratification because

$$\begin{aligned}\text{exog. strat.} \implies \Sigma_{12}(X) &= \mathbb{E}_{P_{Y|X}}\Big[\frac{g(Z,\theta^*)}{b_{\text{exog}}(X)}\frac{\big(s_{\text{exog}}(X)-Q^*_{-L}\big)'}{b_{\text{exog}}(X)}\;\Big|\;X\Big]\\ &= \mathbb{E}_{P^*_{Y^*|X^*}}\Big[\frac{g(Z^*,\theta^*)}{b_{\text{exog}}(X^*)}\frac{\big(s_{\text{exog}}(X^*)-Q^*_{-L}\big)'}{b_{\text{exog}}(X^*)}\;\Big|\;X^*\Big] \qquad ((1.2.8))\\ &= \mathbb{E}_{P^*_{Y^*|X^*}}[g(Z^*,\theta^*)\mid X^*]\frac{\big(s_{\text{exog}}(X^*)-Q^*_{-L}\big)'}{b^2_{\text{exog}}(X^*)}\\ &= 0 \qquad P^*_{X^*}\text{-a.s.} \qquad ((1.2.1))\\ &= 0 \qquad P_X\text{-a.s..} \qquad (P^*_{X^*} \ll P_X \ll P^*_{X^*})\end{aligned}$$

1.3.3 Efficient estimation

The estimation and testing techniques demonstrated here extend Kitamura et al. (2004) and Tripathi and Kitamura (2003). These papers, which are based on a generalisation of the empirical likelihood approach of Owen (1988), develop an asymptotically efficient methodology for estimating and testing models with conditional moment restrictions when the data are collected by random sampling.

In the papers of Kitamura, Tripathi and Ahn, and Tripathi and Kitamura, kernel smoothing is used to efficiently incorporate the information implied by conditional moment restrictions into an empirical likelihood, which is henceforth referred to as the 'Smoothed Empirical Likelihood' (SEL). As shown in these papers, maximising the SEL leads to one-step estimators which avoid any preliminary estimation of optimal instruments. It also yields internally studentised likelihood ratio-type statistics for testing H_0 and parametric restrictions on θ^* that do not require preliminary estimation of any variance terms. Moreover, the resulting estimation and testing procedures are invariant to normalisations of H_0. Simulation results presented in the aforementioned papers suggest that the SEL-based approach can work very well in finite samples.

The advantages of the SEL approach described above extend to the case when the observations are collected by VP sampling. Furthermore, it leads to a unified approach of estimating and testing models using stratified samples, which should appeal to applied economists and practitioners in the field. Therefore, we now demonstrate how to use the SEL approach to construct asymptotically efficient estimators, i. e. estimators with asymptotic variance equal to the efficiency bounds in Section 1.3.2.

If the focus is on the efficient estimation of θ^* alone, then the equivalence in (1.2.6) reveals that replacing the moment function in Kitamura, Tripathi, and Ahn (Equa-

tion 2.1) with $\rho_1(Z,\theta^*)$ will deliver an asymptotically efficient estimator of θ^*.

But what about Q^*_{-L}? Although the aggregate shares $Q^*_{-L} \overset{(1.3.6)}{=} \mathbb{E}_P[s(Z)]/\mathbb{E}_P[1/b(Z)]$ can be simply estimated by their sample analogues, this estimator will not be efficient because it does not take (1.2.1) into account; cf. the discussion after (1.3.8). To construct an estimator of Q^*_{-L} that accounts for (1.2.1), we have to jointly estimate θ^* and Q^*_{-L}, which we do using the SEL approach.

For the remainder of the paper, assume that we have independent observations $Z_1,\dots,Z_n$ collected by VP sampling. Hence, these are IID draws from the realised density dP in (1.2.3). Our estimation approach relies on a smoothed version of empirical likelihood. This smoothing, or localisation, is carried out using positive kernel weights $w_{ij} \overset{\text{def}}{=} \dfrac{\mathcal{K}_{b_n}(X_i - X_j)}{\sum_{k=1}^n \mathcal{K}_{b_n}(X_i - X_k)}$, $i,j = 1,\dots,n$, where $\mathcal{K}$ is a second-order kernel, $\mathcal{K}_{b_n}(\cdot) \overset{\text{def}}{=} \mathcal{K}(\cdot/b_n)$, and b_n the bandwidth.

For $i,j = 1,\dots,n$, let p_{ij} denote the probability mass placed at (X_i, Z_j) by a discrete distribution with support $(X_1,\dots,X_n)\times(Z_1,\dots,Z_n)$. The collection of probabilities $(p_{ij})_{i,j=1}^n$ can be thought of as a set of nuisance parameters that includes the empirical distribution of the data. Using the kernel weights (w_{ij}) and the distribution (p_{ij}) construct the smoothed log-likelihood $\sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij}$. Then, given (θ, Q_{-L}), concentrate out (p_{ij}) by solving the following optimisation problem:

$$\begin{aligned} &\max_{(p_{ij})} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log p_{ij} \\ &\text{s.t.} \quad p_{ij} \geq 0 \quad \text{for} \quad i,j = 1,\dots,n, \qquad \sum_{i=1}^n \sum_{j=1}^n p_{ij} = 1, \\ &\sum_{j=1}^n \rho_1(Z_j,\theta) p_{1j} = 0, \ \dots, \ \sum_{j=1}^n \rho_1(Z_j,\theta) p_{nj} = 0, \quad \sum_{i=1}^n \sum_{j=1}^n \rho_2(Z_j, Q_{-L}) p_{ij} = 0. \end{aligned} \tag{1.3.9}$$

If the convex hulls of $\{\rho_1(Z_1,\theta),\dots,\rho_1(Z_n,\theta)\}$ and $\{\rho_2(Z_1,Q_{-L}),\dots,\rho_2(Z_n,Q_{-L})\}$ contain the origin, then (1.3.9) can be solved by using Lagrange multipliers. In this case, it can be verified that the solution to (1.3.9) is given by

$$\hat{p}_{ij}(\theta, Q_{-L}) \overset{\text{def}}{=} \frac{1}{n}\Big(\frac{w_{ij}}{1 + \lambda_i' \rho_1(Z_j,\theta) + \mu' \rho_2(Z_j, Q_{-L})}\Big), \qquad i,j = 1,\dots,n,$$

where the multipliers $\lambda_i \overset{\text{def}}{=} \lambda_i(\theta, Q_{-L})$ and $\mu \overset{\text{def}}{=} \mu(\theta, Q_{-L})$ solve

$$\begin{aligned} \sum_{j=1}^n \frac{w_{ij}\rho_1(Z_j,\theta)}{1 + \lambda_i' \rho_1(Z_j,\theta) + \mu' \rho_2(Z_j, Q_{-L})} &= 0, \qquad i = 1,\dots,n, \\ \sum_{i=1}^n \sum_{j=1}^n \frac{w_{ij}\rho_2(Z_j,Q_{-L})}{1 + \lambda_i' \rho_1(Z_j,\theta) + \mu' \rho_2(Z_j, Q_{-L})} &= 0. \end{aligned} \tag{1.3.10}$$

The smoothed empirical log-likelihood of (θ, Q_{-L}) is given by

$$\begin{aligned}\mathrm{SEL}(\theta, Q_{-L}) &\stackrel{\text{def}}{=} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\hat{p}_{ij}(\theta, Q_{-L}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\Big(\frac{w_{ij}/n}{1+\lambda_i'\rho_1(Z_j,\theta)+\mu'\rho_2(Z_j,Q_{-L})}\Big),\end{aligned} \tag{1.3.11}$$

where the multipliers solve (1.3.10).

The estimators of θ^* and Q^*_{-L} can, in principle, be defined to be the maximisers of $\mathrm{SEL}(\theta, Q_{-L})$. However, this leads to a constrained optimisation problem because the Lagrange multipliers in $\mathrm{SEL}(\theta, Q_{-L})$ have to satisfy (1.3.10). To ease computation, we convert the constrained optimisation problem into an unconstrained optimisation problem as follows. Begin by observing that, by (1.3.11),

$$\mathrm{SEL}(\theta, Q_{-L}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log(w_{ij}/n) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\big(1+\lambda_i'\rho_1(Z_j,\theta)+\mu'\rho_2(Z_j,Q_{-L})\big).$$

Furthermore,[13]

$$\lambda_1,\ldots,\lambda_n,\mu = \operatorname*{argmax}_{\tilde{\lambda}_1,\ldots,\tilde{\lambda}_n,\tilde{\mu}} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\log\big(1+\tilde{\lambda}_i'\rho_1(Z_j,\theta)+\tilde{\mu}'\rho_2(Z_j,Q_{-L})\big). \tag{1.3.12}$$

Therefore, the estimators of θ^* and Q^*_{-L} are defined to be

$$(\hat{\theta}, \hat{Q}_{-L}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta, Q_{-L}} \mathrm{SEL}_{\mathbb{T}}(\theta, Q_{-L}), \tag{1.3.13}$$

where the 'trimmed' SEL objective function

$$\begin{aligned}\mathrm{SEL}_{\mathbb{T}}(\theta, Q_{-L}) &\stackrel{\text{def}}{=} -\max_{\tilde{\lambda}_1,\ldots,\tilde{\lambda}_n,\tilde{\mu}} \sum_{i=1}^{n}\mathbb{T}_{i,n}\sum_{j=1}^{n} w_{ij}\log\big(1+\tilde{\lambda}_i'\rho_1(Z_j,\theta)+\tilde{\mu}'\rho_2(Z_j,Q_{-L})\big) \\ &= -\max_{\tilde{\mu}} \sum_{i=1}^{n}\mathbb{T}_{i,n}\max_{\tilde{\lambda}_i}\sum_{j=1}^{n} w_{ij}\log\big(1+\tilde{\lambda}_i'\rho_1(Z_j,\theta)+\tilde{\mu}'\rho_2(Z_j,Q_{-L})\big).\end{aligned} \tag{1.3.14}$$

The trimming indicator $\mathbb{T}_{i,n} \stackrel{\text{def}}{=} \mathbb{1}\big(\hat{h}(X_i) \geq b_n^{\tau}\big)$, where $\hat{h}(X_i) \stackrel{\text{def}}{=} (nb_n^{\dim X})^{-1}\sum_{j=1}^{n}\mathcal{K}_{b_n}(X_i - X_j)$ and $\tau \in (0,1)$ is a trimming parameter, is incorporated in (1.3.14) to deal with the 'denominator problem', namely, the instability of the local empirical log-likelihood $\sum_{j=1}^{n} w_{ij}\log\big(1+\tilde{\lambda}_i'\rho_1(Z_j,\theta)+\tilde{\mu}'\rho_2(Z_j,Q_{-L})\big)$ caused by the density of the conditioning variables becoming too small in the tails. Since $\mathbb{T}_{i,n} \xrightarrow{\text{p}} 1$ as $n \to \infty$, this trimming scheme ensures that asymptotically no data is lost.

Following Kitamura, Tripathi and Ahn, it can be shown that, under some regularity conditions, $\hat{\theta}$ and $\hat{Q}_{-L}$ are consistent, asymptotically normal, and asymptotically efficient, i. e. their asymptotic variances match the efficiency bounds.

[13] To see this, compare the first-order conditions for (1.3.12) with (1.3.10).

1.3.4 Testing

The empirical likelihood approach provides a convenient unified environment for testing. For instance, suppose we want to test the parametric restriction $\tilde{\mathrm{H}}_0 : R(\theta^*) = 0$ against the alternative that $\tilde{\mathrm{H}}_0$ is false, where R is a vector of twice continuously differentiable functions such that $\operatorname{rank} \partial_\theta R(\theta^*) = \dim R$. Let

$$(\hat{\theta}_R, \hat{Q}_{-L,R}) \stackrel{\text{def}}{=} \underset{\theta, Q_{-L}}{\operatorname{argmax}} \operatorname{SEL}_{\mathbb{T}}(\theta, Q_{-L}) \quad \text{s.t.} \quad R(\theta) = 0.$$

A version of the likelihood ratio statistic for testing $\tilde{\mathrm{H}}_0$ is given by

$$\mathrm{LR} \stackrel{\text{def}}{=} 2[\operatorname{SEL}_{\mathbb{T}}(\hat{\theta}, \hat{Q}_{-L}) - \operatorname{SEL}_{\mathbb{T}}(\hat{\theta}_R, \hat{Q}_{-L,R})].$$

It can be shown that, under some regularity conditions, $\mathrm{LR} \xrightarrow[n\to\infty]{\text{d}} \chi^2_{\dim R}$ whenever $\tilde{\mathrm{H}}_0$ is true. This result can be used to obtain the critical values for LR. Although a Wald statistic can also be constructed, it is less attractive than LR because the latter is internally studentised. As in parametric situations, LR can be inverted to obtain asymptotically valid confidence intervals. A nice property of confidence intervals based on LR is that they are invariant to nonsingular transformations of the moment conditions. They also automatically satisfy natural range restrictions.

Since inference based on $\hat{\theta}$ is sensible only if (1.2.1) is true, it is important to devise a test for H_0 against the alternative that it is false. As we are dealing with conditional moment restrictions, any specification test which first converts (1.2.1) into a finite set of unconditional moment restrictions will not be consistent for testing H_0. However, using the equivalence in (1.2.6), a consistent test of H_0 is easily obtained by replacing the moment function in Tripathi and Kitamura (2003, Equation 1.1) with $\rho_1(Z, \theta^*)$. Note that since (1.3.6) just identifies the aggregate shares, testing the specification of (1.2.1) and (1.3.6) jointly is equivalent to testing (1.2.1).

1.4 Simulation study

We now examine the finite-sample behaviour of the LS, GMM, and SEL estimators to illustrate the effects of estimating a simple linear regression model specified for the target population when data is collected by VP sampling and stratification is either endogenous or exogenous. Code for the simulations is written in **R**, and the SEL estimator of the model parameters and aggregate shares defined in (1.3.13) is implemented using the algorithm in Owen (2013); see Appendix 1.B for details.

1.4.1 Design

We consider the design in Kitamura et al. (Section 5), which has been used earlier by Cragg (1983) and Newey (1993). The model to be estimated is

$$Y^* = \beta_0^* + \beta_1^* X^* + \sigma^*(X^*)\varepsilon^*, \tag{1.4.1}$$

where $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^* \mid X^*] = 0$ $P^*_{X^*}$-a.s., $\theta^* \stackrel{\text{def}}{=} (\beta_0^*, \beta_1^*) = (1, 1)$, and $(\varepsilon^*, \log X^*) \overset{\text{IID}}{\sim} \mathcal{N}(0, 1)$. We consider two specifications for the skedastic function in the target population: a

(conditional) heteroskedastic design, relevant for applications, with $\sigma^*(X^*) \stackrel{\text{def}}{=} (0.1 + 0.2X^* + 0.3X^{*2})^{1/2}$; and a (conditional) homoskedastic design, essentially of theoretical interest, with $\sigma^*(X^*) \stackrel{\text{def}}{=} 1$.

The target population is partitioned into two strata. Under endogenous stratification, $\mathbb{A}_1 = (-\infty, 1.4)$ and $\mathbb{A}_2 = [1.4, \infty)$. Under exogenous stratification, $\mathbb{B}_1 = \mathbb{A}_1$ and $\mathbb{B}_2 = \mathbb{A}_2$. The aggregate shares for the four configurations are given in Table 1.B.1. The VP sampling probabilities are $(p_1, p_2) = (0.9, 0.3)$; i. e. the first stratum is heavily oversampled, irrespective of whether the stratification is endogenous or exogenous. Since it is typically strata with small aggregate shares that are oversampled, this sampling design focuses on endogenous stratification, which is the object of attention in most applications.

Tables 1.B.2 and 1.B.3 reports the summary statistics averaged across 1000 Monte Carlo replications for the LS estimator $\hat{\theta}_{\text{LS}}$, the GMM estimators $\hat{\theta}_{\text{GMM,endog}}$ and $\hat{\theta}_{\text{GMM,exog}}$, and the SEL estimator $\hat{\theta}$.[14] Three sample sizes are considered, namely, $n = 50, 150, 500$. Tables 1.B.4 and 1.B.5, which summarise the simulation results for estimating Q_1^*, compare the GMM estimators based on the moment conditions in (1.3.7) with the SEL estimator $\hat{Q}_1$.

1.4.2 Discussion

Recall that the LS estimator is inconsistent under endogenous stratification and consistent but generally inefficient under exogenous stratification; the GMM estimators are consistent but inefficient under endogenous and exogenous stratification; the SEL estimator is consistent and asymptotically efficient irrespective of whether the stratification is endogenous or exogenous. Tables 1.B.2–1.B.5 largely confirm these results, at least as far as estimating the model parameters is concerned.

The inconsistency of the LS estimator of β_1^* under endogenous stratification is apparent from Tables 1.B.2 and 1.B.3 because the bias of the LS estimator, as a fraction of β_1^*, remains greater than 9% in magnitude under heteroskedasticity, and greater than 6% under homoskedasticity, as the sample size increases from 50 to 500.[15] In contrast, in both designs, the LS and GMM estimators under exogenous stratification are practically unbiased even when $n = 50$. Under exogenous stratification, the LS estimator has a smaller sampling variance than the GMM estimator for each sample size. However, this finding can be mathematically justified only for homoskedastic designs (recall from Example 1.3.1 that the LS estimator is asymptotically efficient when stratification is exogenous and the error term in the regression model is conditionally homoskedastic in the target population). Indeed, as shown in Appendix 1.A

[14] The SEL estimator is implemented with $\mathbb{T}_{i,n} \stackrel{\text{def}}{=} 1$. To the best of our knowledge, how to choose an optimal data-driven bandwidth for the SEL estimator remains an open problem. Consequently, we naïvely chose the bandwidth by repeating the simulation experiment on a grid of bandwidths and picking the one that minimised the average (across the simulation replications) RMSE of the SEL estimator of β_1^*. The naïvely chosen bandwidth, labelled c_n, is reported in Tables 1.B.2–1.B.5. For the sake of comparison, we also report the SEL estimator when the bandwidth is chosen using Silverman's rule of thumb, namely, $b_n = 1.06\,\widehat{\text{sd}}(X)\,n^{-1/5}$. Since $\widehat{\text{sd}}(X)$ depends on the data, the b_n reported in the tables is the value averaged across the simulations.

[15] This is even more so for the LS estimator of the intercept because, under endogenous stratification, the bias of the LS estimator of β_0^* is ≈18% (resp. ≈41%) in magnitude for the heteroskedastic (resp. homoskedastic) design even when $n = 500$. For the remainder of this section, however, we only discuss the simulation results for the slope coefficient because it can be interpreted as an average partial effect. Results for the intercept, which is a pure level effect, are qualitatively very similar.

(cf. Example 1.A.1), counterexamples can be constructed to show that in heteroskedastic designs, the LS estimator can have higher sampling variance than the GMM estimator when stratification is exogenous.[16] Under endogenous stratification, the GMM estimator of the slope coefficient exhibits some bias (≈2–4% in both designs) when $n = 50$, but the bias is very close to zero when $n = 500$. This is true whether the design is homoskedastic or heteroskedastic, although the magnitude of the bias is higher under heteroskedasticity.

Tables 1.B.2–1.B.5 reveal that the SEL estimator using the naïvely chosen bandwidth (c_n), described in Footnote 14, behaves very similarly to the SEL estimator using the Silverman's rule of thumb bandwidth (b_n). Hence, subsequent discussion regarding the SEL estimator is based on its implementation using the naïvely chosen bandwidth.

The SEL estimator of β_1^* is consistent whether stratification is exogenous or endogenous. In the heteroskedastic design, the SEL estimator exhibits some bias (≈1%) under endogenous stratification when $n = 500$, although its bias under exogenous stratification is close to zero. Moreover, in the heteroskedastic design, the SEL estimator beats the GMM estimator in terms of the RMSE under each stratification scheme and for each sample size. Not surprisingly, the contrast between the two is most pronounced when $n = 500$; e. g. irrespective of the stratification scheme, the RMSE of the GMM estimator is at least 65% larger than the RMSE of the SEL estimator.

In the homoskedastic design, even though it exhibits some bias under endogenous and exogenous stratification when $n = 50$, the bias of the SEL estimator is close to zero for $n = 500$. However, its RMSE is larger than that of the GMM estimator even when $n = 500$. This finding, which corroborates the simulation results in Kitamura et al. (p. 1682), is likely due to the fact that the SEL estimator internally estimates the skedastic function non-parametrically to achieve semi-parametric efficiency and is thus unable to take advantage of conditional homoskedasticity in small samples.

Tables 1.B.4 and 1.B.5 reveal that the GMM estimator of Q_1^* is consistent whether stratification is endogenous or exogenous. It exhibits some upward bias (≈1–2%) in both designs and for both types of stratification when $n = 50$, but the bias is very close to zero when $n = 500$.[17] In both designs, the RMSE of the SEL estimator of Q_1^* is always slightly larger than the RMSE of the GMM estimator under endogenous stratification, implying that in small samples, there appears to be no efficiency gain in estimating Q_1^* jointly with the model parameters. As can be seen from Tables 1.B.4 and 1.B.5, the increase in the RMSE of $\hat{Q}_1$ is due to its bias because RMSE ≈ SE whenever the bias is small. This becomes clear on comparing the bias of $\hat{Q}_1$ under endogenous and exogenous stratification: the latter is always larger. The higher bias of $\hat{Q}_1$ under exogenous stratification is likely a design effect.

[16] It is shown in Appendix 1.A, cf. (1.A.1), that $\mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}}-\theta^*)\big)-\mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{LS}}-\theta^*)\big) = A+B$ holds under exogenous stratification, where the matrix A is positive semidefinite and the matrix B is negative semidefinite. Therefore, in general, it is not clear which estimator has a smaller asymptotic variance. However, since $B = 0$ under conditional homoskedasticity, cf. (1.A.4), $\mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{LS}}-\theta^*)\big) \leq_L \mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}}-\theta^*)\big)$ holds under exogenous stratification and conditional homoskedasticity. Alternatively, under conditional homoskedasticity, the Gauss-Markov theorem implies the same result because $\hat{\theta}_{\mathrm{GMM,exog}}$ and $\hat{\theta}_{\mathrm{LS}}$ are both linear and unbiased when stratification is exogenous.

[17] In Tables 1.B.4 and 1.B.5, the results under exogenous stratification are almost identical for the heteroskedastic and homoskedastic designs because $P^*(X^* \in \mathbb{B}_1)$ is not affected by conditional heteroskedasticity in Y^* (cf. Table 1.B.1).

1.5 Conclusion

When estimating or testing economic relationships, economists often discover that the data they plan to use is not drawn randomly from the target population for which they wish to draw an inference. Instead, the observations are found to be sampled from a related but different distribution. If this feature is not taken into account when doing statistical analysis, subsequent inference can be severely biased. In this paper, we show how to use a smoothed empirical likelihood approach to conduct efficient semi-parametric inference in models characterised as conditional moment equalities when data is collected by variable probability sampling. Results from a simulation experiment suggest that the smoothed-empirical-likelihood-based estimator can estimate the model parameters very well in small to moderately sized stratified samples.

Acknowledgements

We thank two anonymous referees and seminar participants at the 2017 'Econometrics of Complex Survey Data: Theory and Applications' workshop organised by the Bank of Canada, Ottawa, Canada, for helpful comments. The simulation experiments reported in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014, https://hpc.uni.lu).

Appendix

1.A Comparing the asymptotic variance of LS and GMM estimators under exogenous stratification

We begin by proving the assertion in Footnote 16, namely, that, under exogenous stratification, $\mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}} - \theta^*)\big) - \mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{LS}} - \theta^*)\big) = A + B$, where the matrix A is positive semidefinite, the matrix B is negative semidefinite, and $B = 0$ under conditional homoskedasticity.

Recalling the expressions for $V_{\mathrm{GMM,exog}}$ and $V_{\mathrm{LS,exog}}$ in Example 1.3.1, we can write

$$\begin{aligned}\mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}} - \theta^*)\big) - \mathrm{AVar}(n^{1/2}(\hat{\theta}_{\mathrm{LS}} - \theta^*)) \\ = \Big(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\Big)^{-1}\Omega\Big(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\Big)^{-1},\end{aligned}$$

where

$$\begin{aligned}\Omega \stackrel{\mathrm{def}}{=} &\Big(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'\frac{V_{1,\mathrm{exog}}(X)}{b^2_{\mathrm{exog}}(X)}\Big) \\ &- \Big(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\Big)(\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}\big(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\mathrm{exog}}(X)\big)(\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}\Big(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\Big).\end{aligned}$$

Next, letting $a_1 \stackrel{\mathrm{def}}{=} \tilde{X}\sqrt{V_{1,\mathrm{exog}}(X)}/b_{\mathrm{exog}}(X)$ and $a_2 \stackrel{\mathrm{def}}{=} (\mathbb{E}_{P_X}\tilde{X}\tilde{X}')^{-1}\tilde{X}/\sqrt{V_{1,\mathrm{exog}}(X)}$, we have

$$\begin{aligned}\Omega &= \mathbb{E}_{P_X}a_1a_1' - (\mathbb{E}_{P_X}a_1a_2')\big(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\mathrm{exog}}(X)\big)(\mathbb{E}_{P_X}a_2a_1') \\ &= \mathbb{E}_{P_X}a_1a_1' - (\mathbb{E}_{P_X}a_1a_2')(\mathbb{E}_{P_X}a_2a_2')^{-1}(\mathbb{E}_{P_X}a_2a_1') \\ &\qquad + (\mathbb{E}_{P_X}a_1a_2')\big[(\mathbb{E}_{P_X}a_2a_2')^{-1} - \big(\mathbb{E}_{P_X}\tilde{X}\tilde{X}'V_{1,\mathrm{exog}}(X)\big)\big](\mathbb{E}_{P_X}a_2a_1').\end{aligned}$$

Consequently, under exogenous stratification we can write

$$\mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{GMM,exog}} - \theta^*)\big) - \mathrm{AVar}\big(n^{1/2}(\hat{\theta}_{\mathrm{LS}} - \theta^*)\big) = A + B, \tag{1.A.1}$$

where

$$A \stackrel{\mathrm{def}}{=} \Big(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\Big)^{-1}[\mathbb{E}_{P_X}a_1a_1' - (\mathbb{E}_{P_X}a_1a_2')(\mathbb{E}_{P_X}a_2a_2')^{-1}(\mathbb{E}_{P_X}a_2a_1')]\Big(\mathbb{E}_{P_X}\frac{\tilde{X}\tilde{X}'}{b_{\mathrm{exog}}(X)}\Big)^{-1}$$

and

$$B \stackrel{\text{def}}{=} \Big(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{b_{\text{exog}}(X)}\Big)^{-1} (\mathbb{E}_{P_X} a_1 a_2') \\ \times \big[(\mathbb{E}_{P_X} a_2 a_2')^{-1} - \big(\mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X)\big)\big] \Big(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{b_{\text{exog}}(X)}\Big)^{-1} (\mathbb{E}_{P_X} a_2 a_1').$$

It remains to show that A is positive semidefinite and B is negative semidefinite. To do so, recall the matrix version of the Cauchy-Schwarz inequality (Tripathi, 1999), namely,

$$(\mathbb{E}GH')(\mathbb{E}HH')^{-1}(\mathbb{E}HG') \leq_L \mathbb{E}GG', \tag{1.A.2}$$

where G and H are random column vectors. Then, letting $G \stackrel{\text{def}}{=} a_1$ and $H \stackrel{\text{def}}{=} a_2$, it is immediate from (1.A.2) that A is positive semidefinite. Next,

$$(\mathbb{E}_{P_X} a_2 a_2')^{-1} = (\mathbb{E}_{P_X} \tilde{X}\tilde{X}') \Big(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{V_{1,\text{exog}}(X)}\Big)^{-1} (\mathbb{E}_{P_X} \tilde{X}\tilde{X}') \\ \leq_L \mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X)$$

follows from (1.A.2) on letting $G \stackrel{\text{def}}{=} \tilde{X}\sqrt{V_{1,\text{exog}}(X)}$ and $H \stackrel{\text{def}}{=} \tilde{X}/\sqrt{V_{1,\text{exog}}(X)}$. Hence, B is negative semidefinite. Consequently, as A is positive semidefinite and B is negative semidefinite, it is not clear from (1.A.1) which estimator has smaller asymptotic variance.

However, if conditional homoskedasticity holds in the target population, then

$$\text{Var}_{P^*}(Y^* \mid X^*) = \sigma^{*2} \qquad P^*_{X^*}\text{-a.s.}$$

for some constant $\sigma^{*2} > 0$. Moreover, under exogenous stratification,

$$\text{Var}_{P^*}(Y^* \mid X^* = x) \stackrel{(1.2.8)}{=} \text{Var}_P(Y \mid X = x), \qquad x \in \text{supp}(X^*).$$

Hence, since $P^*_{X^*} \ll P_X \ll P^*_{X^*}$, conditional homoskedasticity and exogenous stratification together imply that

$$V_{1,\text{exog}}(X) = \text{Var}_P(Y \mid X) = \sigma^{*2} \qquad P_X\text{-a.s.}. \tag{1.A.3}$$

Therefore, under conditional homoskedasticity and exogenous stratification,

$$\begin{aligned} &(\mathbb{E}_{P_X} a_2 a_2')^{-1} - \mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X) \\ &\quad = (\mathbb{E}_{P_X} \tilde{X}\tilde{X}') \Big(\mathbb{E}_{P_X} \frac{\tilde{X}\tilde{X}'}{V_{1,\text{exog}}(X)}\Big)^{-1} (\mathbb{E}_{P_X} \tilde{X}\tilde{X}') - \mathbb{E}_{P_X} \tilde{X}\tilde{X}' V_{1,\text{exog}}(X) \\ &\quad = \sigma^{*2}[\mathbb{E}_{P_X} \tilde{X}\tilde{X}' - \mathbb{E}_{P_X} \tilde{X}\tilde{X}'] \qquad \text{((1.A.3))} \\ &\quad = 0. \end{aligned}$$

It follows from the definition of B that

$$\text{conditional homoskedasticity and exogenous stratification} \implies B = 0. \tag{1.A.4}$$

Hence, $\text{AVar}\big(n^{1/2}(\hat{\theta}_{\text{LS}} - \theta^*)\big) \leq_L \text{AVar}\big(n^{1/2}(\hat{\theta}_{\text{GMM,exog}} - \theta^*)\big)$ holds under exogenous strati-

fication and conditional homoskedasticity.

However, as demonstrated in the following example, this result may not hold under conditional heteroskedasticity.

Example 1.A.1. Consider (1.4.1) with $\beta_0 \stackrel{\text{def}}{=} 0$, i. e. a simple linear regression through the origin. As before, $\mathbb{E}_{P^*_{Y^*|X^*}}[\varepsilon^* \mid X^*] = 0$ $P^*_{X^*}$-a.s.. Assume that only

$$X^* \stackrel{\text{def}}{=} \begin{cases} c & \text{w. p. } 1-r \\ d & \text{w. p. } r \end{cases}$$

is stratified with $L = 2$, where $\mathbb{B}_1 = (-\infty, 0)$ and $\mathbb{B}_2 = [0, +\infty)$.

Under exogenous stratification,

$$\begin{aligned} \operatorname{AVar}\big(n^{1/2}(\hat{\beta}_{1,\text{LS}} - \beta_1^*)\big) &= \frac{\mathbb{E}_{P_X} X^2 V_{1,\text{exog}}(X)}{(\mathbb{E}_{P_X} X^2)^2} = \frac{\mathbb{E}_{P_X} X^2 \sigma^{*2}(X)}{(\mathbb{E}_{P_X} X^2)^2} \\ \operatorname{AVar}\big(n^{1/2}(\hat{\beta}_{1,\text{GMM}} - \beta_1^*)\big) &= \frac{\mathbb{E}_{P_X}[X^2 V_{1,\text{exog}}(X)/b^2_{\text{exog}}(X)]}{(\mathbb{E}_{P_X} X^2/b_{\text{exog}}(X))^2} = \frac{\mathbb{E}_{P_X}[X^2 \sigma^{*2}(X)/b^2_{\text{exog}}(X)]}{(\mathbb{E}_{P_X} X^2/b_{\text{exog}}(X))^2}. \end{aligned}$$

Let $r = 1/3$, $c = -1$, $d = 2$, $\sigma^{*2}(c) = 1$, $\sigma^{*2}(d) = 4$, $p_1 = 0.2$, and $p_2 = 0.8$. Note that $b_{\text{exog}}(c) = p_1 \mathbb{1}_{\mathbb{B}_1}(c) + p_2 \mathbb{1}_{\mathbb{B}_2}(c) = p_1$ because $c < 0$, and $b_{\text{exog}}(d) = p_1 \mathbb{1}_{\mathbb{B}_1}(d) + p_2 \mathbb{1}_{\mathbb{B}_2}(d) = p_2$ because $d > 0$. Then, it can be verified that

$$\mathbb{E}_{P_X} X^2 \sigma^{*2}(X) = 6,\ \mathbb{E}_{P_X} X^2 = 2,\ \mathbb{E}_{P_X}[X^2 \sigma^{*2}(X)/b^2_{\text{exog}}(X)] = 25,\ \mathbb{E}_{P_X}[X^2/b_{\text{exog}}(X)] = 5.$$

Consequently,

$$\operatorname{AVar}\big(n^{1/2}(\hat{\beta}_{1,\text{LS}} - \beta_1^*)\big) = 1.5 \quad > \quad \operatorname{AVar}\big(n^{1/2}(\hat{\beta}_{1,\text{GMM}} - \beta_1^*)\big) = 1.$$

This shows that the LS estimator may be asymptotically inefficient compared to the GMM estimator under conditional heteroskedasticity and exogenous stratification. □

1.B Computation

In this appendix, we describe how the SEL estimator was implemented by adapting the code of Owen (2017). The **R** function `cemplik` in Owen (2017) was originally written for count random variables, allowing for ties in the data. Let $Z_j \stackrel{\text{def}}{=} (Y_j, X_j)$ be IID draws from the realised density dP, and assume that each of the n distinct values of Z_j can be taken by c_j distinct draws, so that the total sample size is $N \stackrel{\text{def}}{=} \sum_{j=1}^n c_j$. If we impose on the data the vector of unconditional moment equalities $\mathbb{E}_P m(Z, \theta) = 0$, then Owen (2017, p. 2) shows that the empirical log-likelihood, as a function of θ, and modulo constants not depending on θ, is obtained by solving (in our notation)

$$-\max_{\tilde{\lambda}} \sum_{j=1}^n c_j \log\big(1 + \tilde{\lambda}' m(Z_j, \theta)\big). \tag{1.B.1}$$

Note how in (1.B.1) the original sample size N has disappeared, and only the number n of distinct values of Z_j remains. The function `cemplik` asks for $\mathbf{m} \stackrel{\text{def}}{=} \big(m(Z_1, \theta), \ldots, m(Z_n, \theta)\big)$ and a vector $\mathbf{c} \stackrel{\text{def}}{=} (c_1, \ldots, c_n)$ as inputs and delivers three outputs:

1. The empirical log-likelihood (EL) for a given value of θ, computed at the vector $\lambda_{(\dim m)\times 1}$ of Lagrange multipliers that maximise (1.B.1), i. e.

$$\mathrm{EL}_{\mathbf{m}}(\theta; \mathbf{c}, \lambda) \stackrel{\text{def}}{=} -\sum_{j=1}^{n} c_j \log(1 + \lambda' m(Z_j, \theta)).$$

2. The vector λ used to compute $\mathrm{EL}_{\mathbf{m}}(\theta; \mathbf{c}, \lambda)$.
3. The unconditional empirical probabilities

$$p_j \stackrel{\text{def}}{=} \frac{c_j}{n} \frac{1}{1 + \lambda' m(Z_j, \theta)}, \qquad j = 1, \ldots, n.$$

We now describe how to compute $\mathrm{SEL}_{\mathbb{T}}(\theta)$ when only the conditional moment restriction $\mathbb{E}_{P_{Y|X}}[\rho_1(Z, \theta) \mid X] = 0$ is imposed on the data. In the following, we do not deal with ties in the data.[18] Instead, we take advantage of the formal resemblance of the optimisation problem in (1.B.1) with the one that leads to the smoothed empirical log-likelihood. Indeed, obtaining $\mathrm{SEL}_{\mathbb{T}}(\theta)$ only under $\mathbb{E}_{P_{Y|X}}[\rho_1(Z, \theta) \mid X] = 0$ is equivalent to solving (1.3.14) with $\rho_2 \stackrel{\text{def}}{=} 0$, i. e.

$$\mathrm{SEL}_{\mathbb{T}}(\theta)\big|_{\rho_2=0} \stackrel{\text{def}}{=} -\max_{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n} \sum_{i=1}^{n} \mathbb{T}_{i,n} \sum_{j=1}^{n} w_{ij} \log\big(1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta)\big). \tag{1.B.2}$$

From the first order conditions, it is clear that the maximisers in (1.B.2) can be recovered as solutions to n independent maximisation problems, namely,

$$\lambda_i \stackrel{\text{def}}{=} \operatorname*{argmax}_{\tilde{\lambda}_i} \sum_{j=1}^{n} w_{ij} \log\big(1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta)\big), \qquad i = 1, \ldots, n. \tag{1.B.3}$$

The elements of $\mathbf{c}$ in (1.B.1) are not constrained to be integers, but are only supposed to be positive. Hence, comparing (1.B.1) with (1.B.3), we can obtain each λ_i by invoking `cemplik` n times with $\mathbf{c}_i \stackrel{\text{def}}{=} (w_{i1}, \ldots, w_{in})$ as the weights and $\mathbf{m}$ replaced with $\rho_1 \stackrel{\text{def}}{=} \big(\rho_1(Z_1, \theta), \ldots, \rho_1(Z_n, \theta)\big)$. Consequently,

$$\mathrm{SEL}_{\mathbb{T}}(\theta)\big|_{\rho_2=0} = \sum_{i=1}^{n} \mathbb{T}_{i,n} \mathrm{EL}_{\rho_1}(\theta; \mathbf{c}_i, \lambda_i) \tag{1.B.4}$$

with $\mathrm{EL}_{\rho_1}(\theta; \mathbf{c}_i, \lambda) \stackrel{\text{def}}{=} \sum_{j=1}^{n} w_{ij} \log\big(1 + \lambda' \rho_1(Z_j, \theta)\big)$. The **R** commands used to implement (1.B.4) are as follows. Let `rho1` denote $\big(\rho_1(Z_1, \theta), \ldots, \rho_1(Z_n, \theta)\big)$, `sel.weights` be the $n \times n$ matrix whose elements are the kernel weights w_{ij}, and `trim` the trimming vector $\mathbb{T}_{i,n}$. Then, $\mathrm{SEL}_{\mathbb{T}}(\theta)\big|_{\rho_2=0}$ is obtained with the following code:

```
emplik.list <- apply(sel.weights, MARGIN = 1, function(w) cemplik(rho1, w))
SEL <- trim %*% unlist(lapply(emplik.list, "[[", "logelr"))
```

Finally, we show how to impose a conditional and an unconditional moment restriction on the data, i. e. compute the objective function $\mathrm{SEL}_{\mathbb{T}_{i,n}}(\theta, Q_{-L})$ defined in (1.3.14).

[18] In our setup, all the components of Z are continuous random variables, so that ties in the data occur with probability (P) zero.

We treat the optimisation problem in (1.3.14) as a two-step maximisation. In the first step, we fix $\bar{\mu}$ and solve the n independent maximisation problems

$$\max_{\tilde{\lambda}_i} \sum_{j=1}^{n} w_{ij} \log\big(1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta) + \bar{\mu}' \rho_2(Z_j, Q_{-L})\big), \qquad i = 1, \ldots, n. \tag{1.B.5}$$

To carry out the maximisations in (1.B.5), we need to slightly modify Owen's `emplik`. We wrote a function `emplik2` which receives an extra argument $\bar{\mu}' \rho_2(Z_j, Q_{-L})$. The new function `emplik2` evaluates the logarithm in (1.B.3) at $1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta) + \bar{\mu}' \rho_2(Z_j, Q_{-L})$ instead of $1 + \tilde{\lambda}_i' \rho_1(Z_j, \theta)$. The second step needed to compute $\mathrm{SEL}_{\mathbb{T}_{i,n}}(\theta, Q_{-L})$ is a maximisation over $\bar{\mu}$ as shown in (1.3.14), which can be carried out by a standard optimisation routine as follows:[19]

```
SmoothEmplik <- function(mu, rho1, rho2, sel.weights) {
  smooth.emplik.list <- apply(sel.weights, MARGIN = 1, function (w)
  ↪  emplik2(rho1, mu*rho2, w))
  SEL <- trim %*% unlist(lapply(smooth.emplik.list, "[[", "logelr"))
  return(SEL)
}
muopt <- optim(0, SmoothEmplik, rho1, rho2, sel.weights, method = "Brent",
↪  lower = -10, upper = 10)$value
```

The finite-sample performance of the SEL estimator, implemented as described above, is discussed in Section 1.4. The simulation experiments in Section 1.4 were carried out on the high-performance computing clusters at the University of Luxembourg. Table 1.B.6 gives some idea about the average time taken to complete one Monte Carlo replication for the heteroskedastic design (the execution times under endogenous and exogenous stratification are very similar).

[19] This is a simplified but working version of the code we actually used. The complete code is available from GitHub at https://github.com/Fifis/SELshares.

Table 1.B.1: Aggregate shares for the simulation study.

Stratification	Design	Q_1^*
endogenous	homoskedastic	0.27
	heteroskedastic	0.28
exogenous	homoskedastic	0.63
	heteroskedastic	0.63

Table 1.B.2: Simulation summary: Estimated β_0^*, β_1^* under heteroskedasticity.

			Intercept			Slope		
Stratification	n	Estimator	Bias	SE	RMSE	Bias	SE	RMSE
endogenous	50	LS	−.1595	.4417	.4694	−.1076	.4868	.4983
		GMM	.0307	.5165	.5171	−.0408	.4757	.4772
		SEL ($c_n = 0.3$)	−.0033	.2909	.2910	−.0329	.3845	.3859
		SEL ($b_n \approx 0.46$)	.0015	.3916	.3916	−.0213	.4140	.4146
	150	LS	−.1714	.3489	.3885	−.1025	.3514	.3658
		GMM	.0234	.3980	.3985	−.0248	.3325	.3332
		SEL ($c_n = 0.4$)	.0202	.1880	.1891	−.0332	.2394	.2417
		SEL ($b_n \approx 0.39$)	.0224	.2313	.2324	−.0296	.2543	.2560
	500	LS	−.1805	.2894	.3410	−.0906	.2641	.2790
		GMM	.0043	.3304	.3302	−.0061	.2456	.2456
		SEL ($c_n = 0.8$)	.0107	.1316	.1321	−.0130	.1486	.1492
		SEL ($b_n \approx 0.31$)	.0096	.1242	.1246	−.0131	.1454	.1460
exogenous	50	LS	.0038	.3435	.3434	−.0032	.4275	.4273
		GMM	.0080	.4863	.4861	−.0041	.4791	.4789
		SEL ($c_n = 0.3$)	−.0161	.2518	.2523	.0250	.3754	.3762
		SEL ($b_n \approx 0.29$)	−.0098	.3547	.3549	.0117	.4225	.4227
	150	LS	.0021	.2609	.2608	−.0063	.3062	.3061
		GMM	.0042	.3838	.3836	−.0070	.3364	.3363
		SEL ($c_n = 0.4$)	.0010	.1562	.1562	−.0026	.2326	.2326
		SEL ($b_n \approx 0.24$)	.0020	.1910	.1910	−.0034	.2472	.2472
	500	LS	−.0012	.2189	.2188	.0014	.2323	.2322
		GMM	−.0023	.3354	.3352	.0012	.2540	.2539
		SEL ($c_n = 0.8$)	.0006	.1200	.1200	.0017	.1530	.1530
		SEL ($b_n \approx 0.19$)	.0003	.0988	.0988	.0017	.1425	.1425

Table 1.B.3: Simulation summary: Estimated β_0^*, β_1^* under homoskedasticity.

			Intercept			Slope		
Stratification	n	Estimator	Bias	SE	RMSE	Bias	SE	RMSE
endogenous	50	LS	−.4576	.2855	.5392	.0991	.2156	.2372
		GMM	−.0431	.3389	.3415	.0158	.2231	.2236
		SEL ($c_n = 0.3$)	−.0790	.4180	.4255	.0379	.3501	.3521
		SEL ($b_n \approx 0.42$)	−.0559	.3766	.3807	.0255	.2793	.2805
	150	LS	−.4273	.1480	.4522	.0708	.0906	.1149
		GMM	−.0053	.1680	.1680	−.0011	.0902	.0902
		SEL ($c_n = 0.4$)	−.0160	.2028	.2034	.0070	.1364	.1366
		SEL ($b_n \approx 0.35$)	−.0135	.1914	.1919	.0069	.1235	.1237
	500	LS	−.4142	.0845	.4227	.0626	.0453	.0772
		GMM	−.0005	.0938	.0938	.0000	.0427	.0427
		SEL ($c_n = 0.8$)	−.0047	.1026	.1027	.0031	.0577	.0578
		SEL ($b_n \approx 0.28$)	−.0061	.1053	.1055	.0037	.0590	.0591
exogenous	50	LS	−.0039	.2432	.2431	.0031	.1984	.1983
		GMM	−.0022	.2622	.2621	.0022	.2084	.2083
		SEL ($c_n = 0.3$)	−.0184	.3153	.3159	.0230	.3214	.3222
		SEL ($b_n \approx 0.29$)	−.0077	.2958	.2959	.0078	.2787	.2788
	150	LS	−.0007	.1260	.1260	−.0022	.0843	.0843
		GMM	.0011	.1314	.1314	−.0024	.0863	.0863
		SEL ($c_n = 0.4$)	.0001	.1502	.1502	−.0007	.1261	.1261
		SEL ($b_n \approx 0.24$)	.0001	.1516	.1516	−.0012	.1199	.1199
	500	LS	.0027	.0703	.0703	−.0006	.0410	.0410
		GMM	.0040	.0755	.0756	−.0008	.0416	.0416
		SEL ($c_n = 0.8$)	.0053	.0785	.0787	−.0029	.0554	.0555
		SEL ($b_n \approx 0.19$)	.0033	.0801	.0802	−.0012	.0595	.0595

Table 1.B.4: Simulation summary: Estimated Q_1^* under heteroskedasticity.

Stratification	Sample size	Estimator	Bias	SE	RMSE
endogenous	50	GMM	.0132	.0890	.0900
		SEL ($c_n = 0.3$)	.0178	.0939	.0956
		SEL ($b_n \approx 0.46$)	.0209	.0940	.0963
	150	GMM	.0047	.0504	.0506
		SEL ($c_n = 0.4$)	.0096	.0532	.0540
		SEL ($b_n \approx 0.39$)	.0126	.0534	.0549
	500	GMM	.0014	.0278	.0278
		SEL ($c_n = 0.8$)	.0106	.0294	.0313
		SEL ($b_n \approx 0.31$)	.0092	.0293	.0307
exogenous	50	GMM	.0133	.1070	.1078
		SEL ($c_n = 0.3$)	.0384	.1102	.1167
		SEL ($b_n \approx 0.29$)	.0550	.1032	.1169
	150	GMM	.0050	.0633	.0635
		SEL ($c_n = 0.4$)	.0414	.0655	.0775
		SEL ($b_n \approx 0.24$)	.0523	.0620	.0811
	500	GMM	.0009	.0347	.0347
		SEL ($c_n = 0.8$)	.0719	.0364	.0806
		SEL ($b_n \approx 0.19$)	.0471	.0344	.0583

Table 1.B.5: Simulation summary: Estimated Q_1^* under homoskedasticity.

Stratification	Sample size	Estimator	Bias	SE	RMSE
endogenous	50	GMM	.0135	.0873	.0883
		SEL ($c_n = 0.3$)	.0204	.0909	.0931
		SEL ($b_n \approx 0.42$)	.0266	.0924	.0962
	150	GMM	.0042	.0492	.0493
		SEL ($c_n = 0.4$)	.0129	.0515	.0531
		SEL ($b_n \approx 0.35$)	.0176	.0519	.0548
	500	GMM	.0013	.0262	.0263
		SEL ($c_n = 0.8$)	.0180	.0284	.0337
		SEL ($b_n \approx 0.28$)	.0150	.0284	.0321
exogenous	50	GMM	.0133	.1070	.1078
		SEL ($c_n = 0.3$)	.0380	.1088	.1153
		SEL ($b_n \approx 0.29$)	.0561	.1020	.1164
	150	GMM	.0050	.0633	.0635
		SEL ($c_n = 0.4$)	.0420	.0654	.0777
		SEL ($b_n \approx 0.24$)	.0530	.0619	.0815
	500	GMM	.0009	.0347	.0347
		SEL ($c_n = 0.8$)	.0720	.0364	.0807
		SEL ($b_n \approx 0.19$)	.0473	.0344	.0585

Table 1.B.6: Running time (in minutes) to estimate the parameters.

n	(β_0^*, β_1^*)	$(\beta_0^*, \beta_1^*, Q_1^*)$
50	0.036	16.17
150	0.129	45.70
500	0.523	149.4

Chapter 2

The good, the bad, and the asymmetric: Evidence from a new conditional density model

This chapter is based on joint work with Dmitry Igorevich Malakhov. A current version is available as a working paper: Kostyrka, A. V. & Malakhov, D. I. (2021). The good, the bad, and the asymmetric: Evidence from a new conditional density model. *DEM Discussion Paper 2021-09*. https://hdl.handle.net/10993/47435

2.1 Introduction

One of the most well-known stylised facts about the stock market is the asymmetric relation between volatility and prices. The most popular explanation for this phenomenon is the so-called leverage effect (Black, 1976; Christie, 1982): a reduction in stock prices affects the debt-to-equity ratio, making a company riskier and thus increasing its volatility.[20] However, the leverage effect *per se* yields a rather weak explanation of the observed asymmetric relation (Bollerslev et al., 2006), and volatility feedback can also play a role: if the relation between volatility and expected returns is positive and volatility is persistent, then positive shocks also create an anticipated rise in volatility that increases the risk and lowers the price, dampening the impact of the original change. Negative shocks following the aforementioned mechanism should amplify the impact of the initial effect (Bekaert & Wu, 2000). Conversely, asymmetry can also be hidden in price movements due to investors' loss aversion: losses cause a more significant change in utility than gains of the same size (Tversky & Kahneman, 1991). In such a setting, 'bad' and 'good' news of the same magnitude have different effects on prices. Anatolyev and Petukhov (2016) show that in addition to volatility, there is also an asymmetric reaction of the third moment, skewness, to shocks in the S&P 500. Additionally, Engle and Manganelli (2004) argue that negative returns affect value-at-risk values more strongly than positive returns. Therefore, both moments and quantiles of return distributions have different empirical reactions to positive and negative shocks. In general, the problem of asymmetry in financial economics and econometrics is of present interest: beginning with the pioneering work of Markowitz (1959), to date, researchers have been trying to introduce different types of asymmetry into their models to make them more realistic and accurate.[21]

In this paper, we consider a novel univariate conditional density model, which considers the heterogeneity of 'bad' and 'good' shocks. We begin the analysis with two simple observations. First, within each trading period, both negative and positive shocks occur, letting one interpret price changes as a sum of positive and negative shocks. Second, agents react differently to positive and negative news. Based on these facts, we decompose returns into a sum of correlated unobserved positive and negative shocks, both continuous and discrete, thus yielding up to 4 distinct shocks. We call such an approach *the opposite-sign-shock model*. For simplicity, later in the text, we use the terms *approach* and *model* interchangeably, and call negative shocks *'bad'* and positive shocks *'good'*. With the opposite-sign-shock model, it is possible to construct a general parametric non-Gaussian distribution of returns by combining different time-varying copulæ (because the shocks can be correlated) and different marginal

[20] Many variants of asymmetric GARCH models have been proposed to account for this asymmetric reaction of volatility to positive and negative shocks (see EGARCH by Nelson (1991), TARCH by Zakoian (1994), GJR-GARCH by Glosten et al. (1993), APARCH by Ding et al. (1993), ATGARCH by Crouhy and Rockinger (1997), β-GARCH by Guégan and Diebolt (1994), ANST-GARCH by Nam et al. (2002), DAGARCH by Caporin and McAleer (2006), ANM-GARCH by Alexander and Lazar (2009) and many others). Comparison and discussion of different asymmetric models are in Engle and Ng (1993), Hentschel (1995), and Alberg et al. (2008).

[21] See, for example, Ang, Chen et al. (2006), Barndorff-Nielsen et al. (2010), Bekaert and Engstrom (2017), Bekaert et al. (2015), Bollerslev et al. (2020), Bollerslev et al. (2019), Bollerslev et al. (2006), Carr and Wu (2007), El Babsiri and Zakoian (2001), Feunou and Okou (2019), Kiliç and Shaliastovich (2018), Palandri (2015), Park (2016) and Patton and Sheppard (2015), Pelagatti (2009), Tauchen and Zhou (2011).

distributions of 'good' and 'bad' shocks.[22] We show that such an approach is empirically reasonable by comparing its out-of-sample (value-at-risk, variance forecasts) and in-sample (information criteria and specification tests) performances with a large set of standard GARCH models. Justifying the accuracy of this approach, we then try to show unobserved characteristics of return behaviour.

It is common knowledge that the classical normality assumption in popular GARCH-like models simplifies estimation and guarantees that, under several assumptions that are somewhat restrictive from the real-world perspective (e. g. correct specification of conditional mean and variance processes), estimators will be consistent (Newey & Steigerwald, 1997). However, it is believed that the true conditional and unconditional distributions of returns are far from normal.[23], and forecasting accuracy is higher for non-Gaussian GARCH extensions due to better finite-sample performance. We take into account these phenomena; however, there are complexity costs: it is not possible to obtain analytical expressions for the conditional density for the proposed model. In such a situation, one has to resort to numerical methods to obtain estimates and compute risk measures. In general, the numerical approach seems reasonable in the context of asset returns because the true data-generating process is extremely sophisticated, and as mentioned by Strebulaev, Whited et al. (2012), '. . . there exists a tension between realism and the sorts of models that can produce closed-form estimating equations. Better models that can explain more phenomena may not lend themselves to closed-form solutions. . . '.[24]

Sophisticated models often suffer from stability issues that may stem from ill-conditioning (e. g. the model Jacobian being nearly rank-deficient) or insufficient numerical precision (e. g. integration, optimisation, and root-finding routines stopping too early). To make the proposed model useful for practical purposes, we propose several tricks to standard optimisation and numerical integration approaches that can be helpful for many other non-trivial conditional density models.[25] These tricks include proper rescaling of all functions to be integrated and ensuring that numerical integration routines do not converge prematurely. In addition, we use a two-step optimisation procedure to minimise the chances of obtaining a local optimum, which employs stochastic and gradient-based optimisers, as well as warm-start rolling re-estimation.

Bekaert et al. (2015) developed a somewhat similar model independently, named 'Good environment, bad environment' (or BEGE), and our framework allows one to obtain it as a special restricted case when the shape parameter is dynamic and the scale is constant. The proposed approach is much more general: correlation between shocks, potentially non-zero means of shocks, and discrete jumps. The last two points are discussed in more detail below. Using the proposed numerical integration techniques

[22] We induce another channel of asymmetry by allowing variances and correlations of shocks to react asymmetrically to bad and good news.

[23] See a discussion of the causes of observed skewness in returns in J. Chen et al. (2001), Engle and Mistry (2014), Epstein and Schneider (2008) and Hong et al. (2007) Bollerslev and Wooldridge (1992) present evidence of leptokurtosis.

[24] A good example would be Bekaert et al. (2015), which is a special case of our approach with centred gamma distributions and without copulæ and jumps: in spite of the fact that an analytical solution involving hypergeometric functions exists, the authors instead use numerical techniques in their code.

[25] Often, complicated models are not seriously discussed in the literature due to some sort of a bias-variance trade-off: although the population version of a more complex model should give better forecasts, the disproportional estimation noise can worsen forecast quality in finite samples; therefore, a misspecified but simple and robust model can be of greater practical usability (Clark & McCracken, 2015).

and optimisation routine, we also obtain a higher log-likelihood value of the BEGE model on a full sample of the original data (an increase from 1724.3, as reported in that paper, to 1917.8). We show that although Bekaert et al. (2015) provide convincing results for monthly data, most BEGE-like model specifications are rejected when considering daily-frequency results based on VaR tests, including generalisations to variants with other marginal distributions and copulæ. However, certain BEGE-like models with copulæ achieve the lowest AICs, while certain models with copulæ have AICs higher than that of the base BEGE model.

We consider two variants of the proposed models: with non-zero-mean and zero-mean shocks. To interpret the shocks and their influence correctly, we first assume that both types of shocks have non-zero means. In the non-zero-mean case, we also consider loss aversion and risk aversion, allowing dynamic parts of the volatilities of the signed shocks to affect expected returns differently. However, the dynamics of returns' mean are complex, and simple constant-mean models with zero-mean shocks approximate the data well in terms of predicting risk measures (Anatolyev & Tarasyuk, 2015). Therefore, the zero-mean approach can be useful in practice. For the zero-mean case, 'good' shocks can have negative values, and 'bad' shocks can have positive values.

Because discrete jumps are important for risk management, option pricing, portfolio construction and asset pricing (Ait-Sahalia, 2004; Andersen et al., 2007; Jorion, 1988; Kapadia & Zekhnini, 2019; Maheu & McCurdy, 2004), we model both continuous and discrete changes in prices. However, in the literature, there is no consensus about the extent of the jump impact on total volatility. Earlier studies found that jumps explain a significant part of the variation (see Christensen et al. (2014) for a review); however, recent data analyses show that discrete changes can only be attributed to a small part of total volatility because sudden bursts of continuous volatility can be misleadingly interpreted as discrete jumps (Bajgrowicz et al., 2015; Christensen et al., 2014). In the proposed model, discrete changes can also be separated into negative and positive components by following the same logic as in the continuous case. Thus, in its most general form, the proposed model in its most general form incorporates the following stylised facts: (1) heterogeneous impact of 'good' and 'bad' shocks on returns; (2) volatility clustering; (3) possibility of a negative risk-return relation in a model with non-zero-mean shocks; (4) discrete jumps in the return process; and (5) asymmetric, highly non-linear and possibly stepped reactions of moments to positive and negative shocks. We find that return dynamics are non-trivial, and asymmetries play a critical role in all studied aspects of return behaviour. Therefore, investors and regulators should consider such patterns to prevent heavy losses.

Our approach has four important advantages compared to realised volatility papers. First, the realised-variance approach, in general, is sensitive to the number of trades and market frictions (Barndorff-Nielsen et al., 2008), which primarily limits its application to popular and liquid assets. Second, in the realised volatility literature, estimation is often multi-step, where the output of one estimation procedure is fed into a different procedure, which leads to less efficient estimators. Third, a continuous component of the realised variance cannot be separated into negative and positive parts (Patton & Sheppard, 2015). Fourth, certain news information can be interpreted in two ways: for example, new excellent macroeconomic statistics increase asset prices but may also indicate economic overheating, which in turn may make market participants believe that the central bank will increase the interest rate, which can cause price drops. Therefore, it is nearly impossible to directly decompose shocks from observed returns even at tick frequency.

Using 19.5 years of daily U.S. market return data for estimating (15 years) and backtesting (4.5 years), we compare specifications with different distributions and copulæ and provide insights into the latent characteristics of market returns. We show that the proposed model without jumps with dynamic scale parameters is generally superior to the 40 well-established GARCH variants (4 distributions and 10 variance dynamics formulæ) in terms of VaR out-of-sample forecasting (based on Christoffersen (1998), Christoffersen and Pelletier (2004) and Engle and Manganelli (2004) tests) and in-sample explanatory power (based on Vuong's test for likelihood and AIC, and generalised residual tests). The proposed opposite-sign-shock model also has comparable variance forecast quality based on QLIKE and RMSE loss functions. Because we use long training and test samples and several distinct criteria for model comparison (e. g. quantile forecast, moment forecast, information criteria), we believe that the results of this study are not spurious. In general, the opposite-sign-shock model with dynamic copulæ, centred shocks and a more heavy-tailed distribution performs better. The model without jumps with a conditional return distribution using the assumption of centred log-logistic shocks and Clayton copula with a dynamic parameter is 'the best-choice model' because it does not fail the VaR tests, achieves a violation ratio near unity (ratio of observed VaR exceedances to their expected value, which asymptotically tends to 1 for the true model), accurate variance forecasts and good in-sample fit according to Vuong's tests for information criteria and tests for generalised residuals.

As a robustness check, we estimate the same set of specifications on IBM stock return data for the same time period (15 years of estimation sample, 4.5 years of test sample). We show that the models that performed well on the S&P 500 data set also performed well on IBM data in terms of VaR and variance forecast quality and in-sample fit quality. This indicates that the set of best-performing models for daily stock return data is likely to perform well on similar data sets.

We also provide results for the proposed model with static unified (arbitrarily signed), dynamic unified, static opposite-sign, and dynamic opposite-sign jumps; jump sizes are assumed to be normally distributed for unified jumps and chi-squared or exponential for opposite-sign jumps. The model structure provides a natural explanation for the following observed empirical fact: upward jumps of the VIX index are more important than downward jumps (Park, 2016). In the proposed framework, negative and positive jumps in returns can create only positive jumps in volatility;[26] therefore, the VIX index should have the same property. As the base model, we select the best-performing specification without jumps, and results show that for the model with unified jumps (both static and dynamic intensity cases), the jumps represent rare negative return changes because their mean is negative, and they appear approximately once per day on average. The model with opposite-sign jumps shows that 'bad' and 'good' jumps have different behaviour. 'Bad' jumps appear more frequently, and their size is greater in absolute value, based on most specifications. We find that the introduction of jumps does not improve even the in-sample performance of models. The AICs in specifications with opposite-sign dynamic jumps are somewhat lower than those in models with unified jumps; therefore, such a sign decomposition of jumps may marginally improve the performance. However, the AIC for the best specification without jumps is lower than the AIC of all specifications with jumps.

The results obtained in this paper can be enumerated as follows:

[26] However, if covariances between shocks are negative and experience discrete changes, then the total variance can also have negative jumps.

1. 'Good' volatility is extremely persistent, and 'bad' volatility has a more variable behaviour. 'Bad' volatility is marginally greater in magnitude than 'good' volatility, and the leverage effect is more pronounced in the dynamics of 'bad' volatility.
2. A sizeable correlation between shocks exists, is time-varying, and has a leverage-like effect. During calm periods, the correlation remains near 0.7, which is high, and decreases to nearly zero during turmoil. Therefore, during normal times, shocks amplify one another, and during crises, there is barely any connection between them. Therefore, we can expect the U.S. market to have, on average, a propensity for bull trends and a lower possibility of bear trends during crisis times. There is a mean reversion in the dynamic parameter of the copula.
3. 'Good' variance, 'bad' variance and covariance (multiplied by 2) form nearly equal shares of total variance; 'bad' variance has the largest share of 38%. Both volatilities and the correlation between shocks also determine the conditional skewness of returns. Thus, a covariance between shocks is critical for correct modelling.
4. In general, specifications with zero-mean shocks are preferred over non-zero-mean versions; therefore, the relation between returns and volatility is either very non-linear, which cannot be caught by our model, or insignificant.
5. The overall volatility depends on total shocks (unexpected part of returns) in a highly non-linear manner: if the total shocks are negative, then the volatility increases steadily; however, with strongly positive returns, it stays constant or can even drop marginally.
6. The dependence of skewness on total shocks is also non-linear: positive total shocks increase skewness, and negative shocks have the opposite effect. However, even small negative shocks drive the skewness downwards at a higher rate than positive shocks drive it upwards. However, for both positive and negative values of shocks, the skewness impact curve is convex.
7. Conditional skewness switches its sign: during crisis periods, it is positive, leading to a higher probability of extreme positive returns, and during normal times, the skewness has a negative sign. Such counter-cyclical behaviour, combined with typical patterns of prices and volatility behaviours during crisis/normal times, hints at investors' proclivity for lottery-like behaviour[27] with stocks during poor times ('too-bad-to-be-true' situation) and unceasing waiting of the end of growth during good times ('too-good-to-be-true' situation), thus showing the naïveté of investors' expectations.
8. Tails of the conditional distribution exhibit asymmetry. During good times, the probability of extreme negative returns is higher than the probability of extreme positive returns with the same absolute value. For crisis periods, the opposite is true. The probability of positive returns is only slightly higher than the probability of negative returns throughout the entire time period. Therefore, asymmetry mostly comes from the tails, not from the centre of the distribution.
9. At least for daily frequency, the inclusion of jumps with normal or exponentially decaying size densities does not improve even the in-sample performance of the model. Therefore, we can conclude that if the model has rich dynamics of continuous shocks, jumps are not so relevant.

Therefore, asymmetry is present nearly everywhere: in the reaction of 'bad' volatility to positive and negative shocks, in the reaction of correlation between 'bad' and 'good' shocks to positive and negative shocks, in volatility and skewness news impact curves,

[27] Stocks with low prices, high volatility and high skewness are often called lotteries (Kumar, 2009).

in the dissimilarity of the dynamics of 'bad' and 'good' volatilities, and in signed jump behaviour.

This paper is organised as follows. Section 2.2 describes the proposed model in its simplest form; in Section 2.3, we describe the full version of the model with jumps; Section 2.4 suggests an estimation procedure of said model; Section 2.5 contains a small simulation study that shows the behaviour of the proposed model under a known data-generating process; Section 2.6 provides a description of market return data and competing GARCH variants that we benchmark the proposed models against; Section 2.7 yields forecasting and inference results; Section 2.3 describes estimation results with jumps; Section 2.9 provides final conclusions. The appendices contain more details on the replication of Bekaert et al. (2015), a brief description of the method we use to improve numerical stability, and a modification of Vuong's test for model selection.

2.2 Opposite-sign-shock model

Consider the logarithmic returns, r_t, for certain assets. The dynamic process for r_t can be naturally described in the following form:

$$r_t = \mu + \psi_t, \quad t = 1, \dots, T, \tag{2.2.1}$$

where μ is the constant part of returns and ψ_t is the dynamic process. We can separate ψ_t into two parts, 'good' and 'bad':

$$\psi_t = \varepsilon_t^+ + \varepsilon_t^-,$$

where ε_t^+ is a random variable bounded from below and ε_t^- is a random variable bounded from above. Assume that both ε_t^+ and ε_t^- have continuous cumulative probability functions from the scale-shape family of distributions, $\tilde{F}_{\varepsilon_t^+}(x \mid \Omega_{t-1})$ and $\tilde{F}_{\varepsilon_t^-}(y \mid \Omega_{t-1})$ (with conditioning on Ω_{t-1}, all returns prior to and including $t-1$). We assume that the scale parameters are dynamic:

$$\varepsilon_t^+ = \sqrt{\sigma^2_{\varepsilon_t^+}} \cdot e_t^+, \quad \varepsilon_t^- = \sqrt{\sigma^2_{\varepsilon_t^-}} \cdot e_t^-,$$

where e_t^+ are IID random variables with scale 1 bounded from below, $\operatorname{supp} e_t^+ = [\underline{e}_t^+, +\infty)$, and e_t^- are IID random variables with scale 1 bounded from above, $\operatorname{supp} e_t^- = (-\infty, \overline{e}_t^-]$. We consider two cases: strictly positive 'good' and negative 'bad' shocks ($\underline{e}_t^+ = \overline{e}_t^- = 0 \; \forall t$) or zero-mean shocks with centred distributions, such that $\underline{e}_t^+ < 0$ and $\overline{e}_t^- > 0$.

The conditional scale parameters $\sigma_{\varepsilon_t^+}$ and $\sigma_{\varepsilon_t^-}$ are defined in a GJR-GARCH-like manner (Glosten et al., 1993) (see formula (2.2.6) below). The baseline shocks e_t^+ and e_t^- have continuous CDFs $\tilde{F}_{e_t^+}(x)$ and $\tilde{F}_{e_t^-}(y)$. We name e_t^+ *'good' shocks* and e_t^- *'bad' shocks*, while $\sqrt{\operatorname{Var} \varepsilon_t^+}$ and $\sqrt{\operatorname{Var} \varepsilon_t^-}$ represent *'good' volatility* and *'bad' volatility*, respectively. Using such a decomposition, we can consider r_t as a weighted mixture of 'good' and 'bad' stochastic shocks with volatilities as weights. These shock weights depend on the information from the previous periods, creating non-linear inertia in returns. If conditional scales have disparate dynamics, then shocks of different signs have different impacts, and the model can exhibit asymmetric properties. To replicate the results of Bekaert et al. (2015), we also estimate specifications with dynamic shape and constant

scale parameters in the same manner: formula (2.A.1) (identical to Bekaert et al., 2015, eq. (3) up to scaling) generates dynamic shape series. If one considers such a specification with conditional shape, then the aforementioned shock weights will be constant; however, the shocks themselves will have time-varying properties. We refer to the dynamic-scale version as the default version due to its better tractability. Additionally, if one assumes non-zero-mean shocks, then such a decomposition obtains an intuitive and straightforward interpretation because each of the shocks can have only one sign.

The conditional expectation of returns follows:

$$\mathbb{E}_{t-1} r_t = \mu + \sqrt{\sigma^2_{\varepsilon^+_t}}\mu_{e^+} + \sqrt{\sigma^2_{\varepsilon^-_t}}\mu_{e^-},$$

where μ_{e^+} is the unconditional expectation of 'good' shocks and μ_{e^-} is the unconditional expectation of 'bad' shocks. We do not provide a deep qualitative interpretation of e^+ and e^- because certain news information (e. g. news about better economic prospects) can negatively affect the price of counter-cyclical companies because stock buyers and sellers have directly opposite preferences for the future performance of firms. Bartram et al. (2012) provide a discussion of the 'bad' and 'good' parts of the idiosyncratic volatility of American firms. The higher idiosyncratic volatility of returns of American firms compared to foreign firms is associated with factors positively affecting welfare: investor protection, stock market development, innovation, and growth opportunities. Therefore, volatility is not ultimately a bad thing, but can also represent growth potential and entrepreneurial inventiveness. Therefore, we simply treat e^+ as the shocks that drive returns upwards and e^- as the shocks that decrease returns, and their fundamental interpretation strongly depends on the situation. Because μ_{e^-} and μ_{e^+} can take values of arbitrary magnitude, such models combine loss-aversion and risk-aversion effects because the dynamic scale parameters of the signed shocks, which approximate risk, have a different impact on expected returns. Therefore, the market interprets the downward and upward risk in different ways, and the dynamics of expected returns become richer.

If shocks have non-zero means, their unexpected parts determine corresponding volatilities, and their means relate expected returns to these volatilities. Negative values of μ_{e^-} lead to the possibility of a negative risk-return trade-off, diminishing the effect of volatility feedback and assisting in the creation of bear trends. Beginning with the ICAPM of Merton (1973), it is assumed that expected returns have a positive connection with variance; however, modern literature often finds a negative risk-return relation (Ang, Hodrick et al., 2006, 2009; Atilgan et al., 2019; Babenko et al., 2016; Campbell et al., 2008; Ghysels et al., 2014; Glosten et al., 1993; Hou & Loh, 2016; Stambaugh et al., 2015). As can be inferred from recent empirical findings (Adrian et al., 2019; Ghysels et al., 2014), the risk-return relation can be negative during crisis periods and positive during good times. In the proposed case, we obtain a similar picture: when a crisis begins, the influence (volatility) of bad news is heavier; therefore, an overall negative risk-return relation can occur. In the proposed models, the sign of the impacts of 'bad' and 'good' volatilities on expected returns also typically coincides with the findings of Kiliç and Shaliastovich (2018, Table 3), who show that 'bad' realised variance has a negative effect on future returns, and 'good' variance has a positive effect. Therefore, the ambiguity with the coefficient sign in classical GARCH-in-mean models (Engle et al., 1987) can be caused by the aggregation of shocks.

In the standard approach, expected returns are restricted to be positive, and rational

agents should expect an upward trend in prices even during crises when the bear trend on the price graph is clear. Over the course of the last 25 years, approximately 46% of S&P 500 daily returns have been negative. In this case, investors should admit the possibility of negative returns. Finally, with strictly positive expected returns, it is unclear why investors would use short selling; our framework yields a natural statistical argumentation why the conditional expectation of returns can be negative.

The zero-mean assumption for shocks is common in GARCH modelling due to the problem of correct specification of the mean process and insensitivity of variance forecasting accuracy to the mean process specification (Anatolyev & Tarasyuk, 2015). A zero-mean approach can overcome the problems associated with situations where complex specifications turn out to be incorrect and cause more problems than explicitly incorrect but simple specifications. However, in the case of zero-mean shocks, the interpretation of the decomposition is less trivial because a large portion of the probability mass of positive (negative) shocks will be shifted to negative (positive) values; therefore, both shocks can have positive and negative values. In such cases, the marginal distribution of zero-mean 'bad' shocks does not fully determine the shape of the left part of the joint distribution, and 'good' shocks do not fully determine the right part.

It is reasonable to assume that 'good' and 'bad' shocks correlate. To model this assumption, one should use a copula function (Sklar, 1959) to connect the marginal distribution functions of the shocks. The bivariate conditional joint cumulative distribution function of shocks can be written as

$$F_{\varepsilon_t^+,\varepsilon_t^-}(x, y \mid \Omega_{t-1}) = S\big(\tilde{F}_{\varepsilon_t^+}(x \mid \Omega_{t-1}), \tilde{F}_{\varepsilon_t^-}(y \mid \Omega_{t-1}) \;\big|\; \Omega_{t-1}\big), \tag{2.2.2}$$

where $\tilde{F}_{\varepsilon_t^+}(x)$ is the marginal cumulative distribution function of weighted 'good' shocks, $\tilde{F}_{\varepsilon_t^-}(y)$ is the marginal cumulative distribution function of weighted 'bad' shocks, and $S(\cdot \mid \Omega_{t-1})$ is the conditional copula function (Patton, 2006). Because we use continuous marginal distributions, the copula function is unique for certain joint distributions (Sklar, 1959). This structure of the joint distribution function allows for greater flexibility in parameterisation because different cumulative probability functions for 'bad' and 'good' shocks can be chosen; and using various conditional copulæ, one can account for non-trivial dependence and construct and test a general and flexible model (2.2.1) in the manner conditional density models are designed (Anatolyev & Petukhov, 2016; Hansen, 1994; Harvey & Siddique, 1999; Rockinger & Jondeau, 2002).

Because shocks can be centred (zero-mean shocks) or non-centred (non-zero-mean shocks), we use $\tilde{f}_{\varepsilon_t^+}(x)$ to denote the PDF of weighted 'good' shocks,ε_t^+, in the general case:

$$\tilde{f}_{\varepsilon_t^+}(x) \stackrel{\text{def}}{=} \begin{cases} f_{\varepsilon_t^+}(x), & \text{no centring,} \\ f_{\varepsilon_t^+}(x - \sigma_{\varepsilon_t^+}\mu_{e^+}), & \text{centring.} \end{cases}$$

The PDF of weighted 'bad' shocks is defined similarly.

Let $S^{ab}(x, y) \stackrel{\text{def}}{=} \frac{\partial^{a+b}}{\partial x^a \partial y^b} S(x, y)$ denote the mixed partial derivative of the copula function. Using (2.2.2), one obtains the following conditional joint probability density

function:

$$f_{\varepsilon_t^+,\varepsilon_t^-}(x,y \mid \Omega_{t-1}) = \frac{\partial^2}{\partial x \partial y} F_{\varepsilon_t^+,\varepsilon_t^-}(x,y \mid \Omega_{t-1}) \\ = \tilde{f}_{\varepsilon_t^+}(x \mid \Omega_{t-1}) \cdot \tilde{f}_{\varepsilon_t^-}(y \mid \Omega_{t-1}) \cdot S^{11}\big(\tilde{F}_{\varepsilon_t^+}(x \mid \Omega_{t-1}), \tilde{F}_{\varepsilon_t^-}(y \mid \Omega_{t-1}) \,\big|\, \Omega_{t-1}\big). \tag{2.2.3}$$

The probability density function of ψ_t follows from the formula for the density of a sum of two random variables:

$$f_{\psi_t}(z \mid \Omega_{t-1}) = \int_{\operatorname{supp} \varepsilon_t^-} f_{\varepsilon_t^+,\varepsilon_t^-}(z-v, v \mid \Omega_{t-1}) \,\mathrm{d}v. \tag{2.2.4}$$

The density of returns themselves can be formulated trivially via the density of shocks:

$$f_{r_t}(z \mid \Omega_{t-1}) = f_{\psi_t}(z - \mu \mid \Omega_{t-1}). \tag{2.2.5}$$

Now we specify the particular distributions, copulæ, and dynamics of parameters. We use the following GJR-GARCH formula for dynamic scale parameters because they can capture volatility clustering and asymmetry effects:

$$\begin{cases} \sigma^2_{\varepsilon_t^+} = \alpha_0 + \alpha_1 \sigma^2_{\varepsilon_{t-1}^+} + \alpha_2 \tilde{r}^2_{t-1} + \alpha_2^- \tilde{r}^2_{t-1} \mathbb{I}^-_{t-1}, \\ \sigma^2_{\varepsilon_t^-} = \beta_0 + \beta_1 \sigma^2_{\varepsilon_{t-1}^-} + \beta_2 \tilde{r}^2_{t-1} + \beta_2^- \tilde{r}^2_{t-1} \mathbb{I}^-_{t-1}, \end{cases} \tag{2.2.6}$$

where $\tilde{r}_t \overset{\text{def}}{=} r_t$ if there is no de-meaning or $\tilde{r}_t \overset{\text{def}}{=} r_t - \mathbb{E}_{t-1} r_t$ if the returns are de-meaned (in this case, $\tilde{r}_t$ can be interpreted as the 'unexpected' part of returns), and $\mathbb{I}^-_{t-1} \overset{\text{def}}{=} \mathbb{I}(\tilde{r}_{t-1} < 0)$ is the indicator function equal to 1 if the returns on the previous day (original or de-meaned) were negative and 0 otherwise. To save space, we present results only for specifications with de-meaning; those without de-meaning yielded similar or marginally worse results. We also used VAR-type specifications of volatility and included shock scale parameters in the dynamics of the copula parameter. However, those specifications yielded nearly identical results; thus, we also omit them from the discussion.

We choose two distributions for e_t^+ and e_t^-, gamma and log-logistic, each of which is either centred or not centred.[28] The gamma distribution exhibits the property of an exponential law of tail decay; thus, both tails of the modelled distribution will be relatively light, and for large values of θ, it will be similar to a Gaussian distribution. The tail of the log-logistic distribution, however, is extremely heavy for small θ, exhibiting a power law of decay. Log-logistic distribution with shape θ has finite moments of order $k < \theta$. With large values of θ, the log-logistic distribution can also be used to model the return distribution with light tails because its excess kurtosis tends to 6/5 as $\theta \to \infty$, and the distribution tends to simple logistic with infinitesimal variance. Estimation of the shapes of underlying log-logistic distributions will thus show whether the tails of the return distribution are heavy or not. The distributions of 'good' and 'bad' shocks have different shape parameters, θ^+ and θ^-, which allows the distribution of

[28] The PDF of the gamma distribution is $x^{\theta-1} \exp\big(-\frac{x}{\sigma}\big)/(\Gamma(\theta)\sigma^\theta)$ for $x \geq 0$. The PDF of log-logistic distribution is $\frac{\theta}{\sigma}(x/\sigma)^{\theta-1}/(1+(x/\sigma)^\theta)^2$ for $x \geq 0$, where θ denotes the shape parameter, σ denotes the scale parameter, and Γ denotes the gamma function.

'bad' shocks to have a tail thickness different from that of 'good' shocks.[29] For simplicity, we assume that both shocks come from the same family of densities.

To make the proposed model parsimonious, we consider copula functions S with only one parameter governing the strength and direction of dependence (see Table 2.2.1 for the list). Using the results from Anatolyev and Petukhov (2016), it is natural to assume that the copula dependence parameter κ_t, which determines the correlation of weighted shocks and affects higher moments of the joint distribution, follows a specification similar to that of the dynamic scale parameters:

$$\kappa_t = \gamma_0 + \gamma_1 \kappa_{t-1} + \gamma_2 \tilde{r}^p_{t-1} + \gamma_2^- \tilde{r}^p_{t-1} \mathbb{I}^-_{t-1}. \tag{2.2.7}$$

In this study, the returns are raised to powers $p = 2$ or 3 because the latter transformation preserves the sign of the shocks while amplifying the differences between shocks with large and small absolute values. Naturally, specifications with a static copula (where $\kappa_t = \gamma_0$) can also be considered.

Table 2.2.1: Copula functions used in this paper

Copula	$C(G, H)$	Range
Independence	$G \cdot H$	—
Plackett	$\frac{1+(\kappa-1)(G+H)-\sqrt{[1+(\kappa-1)(G+H)]^2-4\kappa(\kappa-1)F\cdot G}}{2(\kappa-1)}$	$\kappa > 0$
Cubic	$G \cdot H[1 + \kappa(G-1)(H-1)(2G-1)(2H-1)]$	$-1 \leq \kappa \leq 2$
AMH	$\frac{G \cdot H}{1-\kappa(1-G)(1-H)}$	$-1 \leq \kappa \leq 1$
Clayton	$(G^{-\kappa} + H^{-\kappa} - 1)^{-1/\kappa}$	$\kappa > 0$

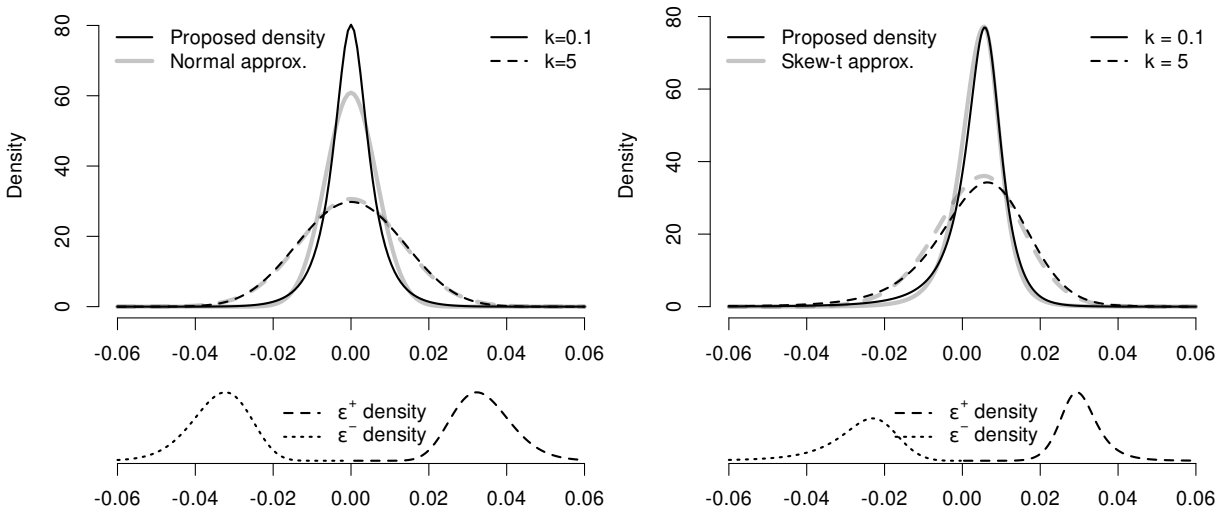
G and H denote marginal distribution functions.

As mentioned previously, the return density function in the proposed model is quite general due to the choice of underlying distribution functions. However, shocks with gamma or log-logistic distributions with large shape parameters can potentially generate densities that are similar to Gaussian densities. The copula function and the differences in shape and scale parameters determine the skewness in return distribution; thus, one can directly infer that volatilities and correlation between the shocks affect the skewness of returns.

We assume that the model is correctly specified (the proposed density is the true density of returns) and that the standard conditions for weak consistency and normality of the maximum likelihood estimator hold (Wooldridge, 1994, Theorems 5.1 and 5.2). The presence of copulæ does not have any implications for identification because we assume that the conditional joint density in Eq. 2.2.3 containing the chosen copula is the true density; therefore, there are no partial-identification-related problems, such as those described in Fan et al. (2014), stemming from the lack of information on the joint distribution.

[29] For the log-logistic distribution, the moments of order k exist only if the shape parameter is greater than k; however, in our estimation, they turned out to be greater than 4. Thus, we obtained finite values of conditional volatility, skewness, and kurtosis.

Figure 2.2.1: Conditional density of the shock sum used in this paper



The left panel shows how the proposed model with a non-centred gamma distribution of signed shocks can produce return distributions similar to a Gaussian distribution. Clayton copula is used. The parameter $\kappa = 0.1$ corresponds to weak positive dependence, and $\kappa = 5$ to strong positive dependence. Shape parameters: $\theta^+ = \theta^- = 20$, scale parameters: $\sigma_{\varepsilon^+} = \sigma_{\varepsilon^-} = 0.0017$.
The right panel shows how the proposed model with a non-centred log-logistic distribution of signed shocks can produce asymmetric heavy-tailed return distributions. Plackett copula is used. The parameter $\kappa = 0.1$ corresponds to strong negative dependence, and $\kappa = 5$ to strong positive dependence. Shape parameters: $\theta^+ = 5$, $\theta^- = 10$. Scale parameters: $\sigma_{\varepsilon^+} = 0.030$, $\sigma_{\varepsilon^-} = 0.025$.
Grey lines correspond to the Gaussian (left) and skew-t (right) distributions with densities closest to the proposed ones. Skewness is introduced into the t distribution according to formula (2.6.1).

2.3 Adding jumps to the model

Because discrete changes are important for return dynamics (Ait-Sahalia, 2004; Andersen et al., 2007; Bollerslev et al., 2016; Bollerslev et al., 2019; Kapadia & Zekhnini, 2019; Maheu & McCurdy, 2004; Maheu et al., 2013; Patton & Sheppard, 2015), we introduce arbitrarily signed (unified) and opposite-signed (separate positive and negative) jumps into the model.

2.3.1 Unified jumps

We re-specify (2.2.1) as follows:

$$r_t = \mu + \psi_t, \quad \psi_t = \varepsilon_t^+ + \varepsilon_t^- + \nu_t,$$

$$\nu_t = \begin{cases} 0, & \mathbb{P}(n_t = 0), \quad n_t \sim \mathrm{Pois}(\lambda_t), \\ \sum_{i=1}^{n_t} \xi_{i;t}, & \mathbb{P}(n_t = i), \quad \xi_{1;t}, \dots, \xi_{i;t} \overset{\text{IID}}{\sim} \mathcal{N}(\mu_\xi, \sigma_\xi^2), \end{cases}$$

where $\xi_{i;t}$ are the magnitudes of jumps (continuous IID random variables), and n_t is their number (conditionally dependent Poisson random variables). For simplicity, the following notation will be used: $\nu_{i;t} \overset{\text{def}}{=} \sum_{k=1}^{i} \xi_{k;t}$. We thus use the normality assumption to simplify the model and be consistent with the existing literature (e. g. Maheu and McCurdy (2004)). Additionally, such a setting is important because we can analyse whether the model can separate continuous fat-tailed shocks from discrete jumps from thin-tailed distributions. In many cases, it is difficult to say whether an abrupt price

change is caused by a discontinuous jump or by a realisation from the tail region of a fat-tailed continuous shock.

Due to a property of Gaussian random variables, $\nu_{i;t} \sim \mathcal{N}(i \cdot \mu_\xi, i \cdot \sigma_\xi^2)$. Therefore, jumps in return dynamics create discrete changes in volatility, making the volatility process less smooth, which corresponds with empirical findings (Todorov & Tauchen, 2011). Although centred or non-centred jumps can be used for modelling these discrete changes, the case of non-centred jumps is more important because jumps affect the risk premium, and the expected jump size determines the sign of the return-jump relation. As mentioned in Bollerslev et al. (2016), investors can treat jump risk in a different manner compared to smooth risk because it is more difficult to hedge from jumps. Compared to the approach in the literature, where jumps must be rare and large (e. g. Tauchen and Zhou (2011)), such strong assumptions are unnecessary with the proposed parametric settings, where both types of shocks exhibit a different nature.

Following the approach from the previous section, we use copula functions to model the dependence between 'bad' and 'good' shocks and between continuous shocks and jumps. To avoid considering models with sums of many random variables requiring the computation of quadruple, quintuple, and more multiple integrals, we use nested copulæ. We model the continuous shocks with copula S and then connect the copula S and jumps with the outer copula C. The joint cumulative distribution function becomes a weighted sum, and we drop the '$\mid \Omega_{t-1}$' conditioning notation, which is implied:

$$\begin{aligned} F_{\varepsilon_t^+,\varepsilon_t^-}(x, y, u \mid \Omega_{t-1}) = {} & S\big(F_{\varepsilon_t^+}(x), F_{\varepsilon_t^-}(y)\big) \cdot \mathbb{P}(n_t = 0) \\ & + \sum_{i=1}^{\infty} C\Big(S\big(F_{\varepsilon_t^+}(x), F_{\varepsilon_t^-}(y)\big), F_{\nu_{i;t}}(u)\Big) \cdot \mathbb{P}(n_t = i), \end{aligned} \tag{2.3.1}$$

where C is the outer bivariate copula for connecting jumps and continuous shocks and S is the bivariate copula connecting shocks.

Let S denote the copula for shocks; C the outer copula; $\mathcal{S} \overset{\text{def}}{=} S\big(F_{\varepsilon_t^+}(\cdot), F_{\varepsilon_t^-}(\cdot\cdot)\big)$, where the arguments $(\cdot)$ and $(\cdot\cdot)$ are the arguments of $f_{\varepsilon_t^+}$ and $f_{\varepsilon_t^-}$ preceding in the same product, respectively; and $\mathcal{C}_i \overset{\text{def}}{=} C(\mathcal{S}, F_{\nu_{i;t}}(\cdots))$, where $(\cdots)$ is the argument of $f_{\nu_{i;t}}$ preceding in the same product. The arguments of $\mathcal{S}$ and $\mathcal{C}$ are omitted to maintain the expression concise. With these settings, we have a copula between continuous shocks and the overall jump change in the returns. Like previously, let the superscript of the copula function denote the order of the mixed derivative. The density function for the overall stochastic part is

$$\begin{aligned} f_{\psi_t}(z) = {} & \int_{\operatorname{supp} \varepsilon_t^-} f_{\varepsilon_t^+}(z - y) \cdot f_{\varepsilon_t^-}(y) \cdot \mathcal{S}^{11} \,\mathrm{d}y \cdot \mathbb{P}(n_t = 0) \\ & + \sum_{i \in \mathbb{N}} \int_{\mathbb{R}} \int_{\operatorname{supp} \varepsilon_t^-} f_{\varepsilon_t^+}(z - y - w) \cdot f_{\varepsilon_t^-}(y) \cdot f_{\nu_{i;t}}(w) \cdot \\ & \qquad \cdot (\mathcal{S}^{11}\mathcal{C}_i^{11} + \mathcal{S}^{01}\mathcal{S}^{10}\mathcal{C}_i^{21}) \,\mathrm{d}y \,\mathrm{d}w \cdot \mathbb{P}(n_t = i), \end{aligned} \tag{2.3.2}$$

where $\mathbb{P}(n_t = k)$ is the probability that a Poisson random variable equals k (i. e. $\lambda_t^k \mathrm{e}^{-\lambda_t}/k!$). We assume the same distributions and copulæ as in the model without jumps.

The first apparent problem is the fact that n_t can take arbitrarily large integer values; therefore, the number of summation terms in (2.3.2) will be infinite. For

practical purposes, we assume that the market cannot experience more than m jumps, where m is sufficiently high, at any point in time. We thus restrict the summation limit in (2.3.2) to a finite number.[30] Because the distribution of the number of jumps is truncated at m, the probabilities $\mathbb{P}(n_t = 0), \ldots, \mathbb{P}(n_t = m)$ are re-normalised so that they add up to one:

$$\tilde{\mathbb{P}}_m(n_t = i) = \frac{\mathbb{P}(n_t = i)}{\mathbb{P}(n_t \leq m)}.$$

Finally, we introduce dynamics into the parameter of the number of jumps distribution:

$$\lambda_t = \delta_0 + \delta_1 \lambda_{t-1} + \delta_2 \pi_{t-1},$$

where π_{t-1} is the *intensity residual* as defined by Maheu and McCurdy (2004):

$$\pi_{t-1} = \mathbb{E}(n_{t-1} \mid \Omega_{t-1}) - \lambda_{t-1} = \sum_{j=0}^{\infty} j \mathbb{P}(n_{t-1} = j \mid \Omega_{t-1}) - \lambda_{t-1}. \tag{2.3.3}$$

The probability in the formula above is equal to

$$\mathbb{P}(n_t = j \mid \Omega_t) = \frac{f_{r_t}(r_t \mid n_t = j, \Omega_{t-1}) \mathbb{P}(n_t = j \mid \Omega_{t-1})}{\sum_j f_{r_t}(r_t \mid n_t = j, \Omega_{t-1}) \mathbb{P}(n_t = j \mid \Omega_{t-1})}. \tag{2.3.4}$$

2.3.2 Opposite-sign jumps

Jumps of opposite signs can have a different influence on returns and volatility dynamics (Park, 2016; Patton & Sheppard, 2015; Tauchen & Zhou, 2011). Therefore, to take this phenomenon into account, we consider a specification with distinct 'good' and 'bad' jumps, which yield up to 4 latent variables in the return dynamics:

$$r_t = \mu + \psi_t, \quad \psi_t = \varepsilon_t^+ + \varepsilon_t^- + \nu_t^+ + \nu_t^-,$$

$$\nu_t^+ = \begin{cases} 0, & \mathbb{P}(n_t = 0), \\ \sum_{i=1}^{n_t} \xi_{i;t}^+, & \mathbb{P}(n_t^+ = i), \end{cases} \qquad \nu_t^- = \begin{cases} 0, & \mathbb{P}(n_t = 0), \\ \sum_{i=1}^{n_t} \xi_{i;t}^-, & \mathbb{P}(n_t^- = i), \end{cases} \qquad \begin{cases} n_t^+ \sim \text{Pois}(\lambda_t^+), \\ n_t^- \sim \text{Pois}(\lambda_t^-), \end{cases}$$

where $\xi_i^+ \sim \text{IID}$ and $\xi_i^- \sim \text{IID}$ are the 'good' and 'bad' jumps that are bounded from below and from above, respectively. We consider exponential and Rayleigh distributions[31] for the intensity of signed jumps to maintain the number of parameters low and possibly account for the presence of zero (exponential) or non-zero (Rayleigh) modes in the jump intensity. We do not centre these distributions, so the expectations of jump sizes determine the sign of the return-jump relationship. We thus do not make assumptions similar to those in Tauchen and Zhou (2011), where a jump dominates the overall daily price change; therefore, the jump sign determines the return sign. This model variant is extremely sophisticated, and the dynamics of this model are rich and can consider many stylised facts about asset returns.

Again, we omit all conditioning from the following notation for simplicity. In principle, 'good' and 'bad' jumps can be connected by completely different copulæ; thus, the

[30] Following the results from Maheu and McCurdy (2004), we use $m = 6$.

[31] The PDF of exponential distribution with rate $1/\sigma$ (inverse scale) is $\exp(-x/\sigma)/\sigma$. The PDF of Rayleigh distribution with scale σ is $x \cdot \exp[-x^2/(2\sigma^2)]/\sigma^2$.

joint distribution function can be written as

$$\begin{cases} S\big(F_{\varepsilon_t^+}(x), F_{\varepsilon_t^-}(y)\big) & \text{w. p. } \mathbb{P}(n_t^+ = 0, n_t^- = 0), \\ C\big(S\big(F_{\varepsilon_t^+}(x), F_{\varepsilon_t^-}(y)\big), F_{\nu_{i;t}^+}(u)\big) & \text{w. p. } \mathbb{P}(n_t^+ = i, n_t^- = 0), \\ C\big(S\big(F_{\varepsilon_t^+}(x), F_{\varepsilon_t^-}(y)\big), F_{\nu_{j;t}^-}(v)\big) & \text{w. p. } \mathbb{P}(n_t^+ = 0, n_t^- = j), \\ C\big(S\big(F_{\varepsilon_t^+}(x), F_{\varepsilon_t^-}(y)\big), J\big(F_{\nu_{i;t}^+}(u), F_{\nu_{j;t}^-}(v)\big)\big) & \text{w. p. } \mathbb{P}(n_t^+ = i, n_t^- = j), \end{cases}$$

where $i, j \in \mathbb{N}$.

We obtain the density of the sum of random variables by integrating the joint density function. Similar to the previous section, let $\mathcal{S} \stackrel{\text{def}}{=} S\big(F_{\varepsilon_t^+}(\cdot), F_{\varepsilon_t^-}(\cdot\cdot)\big)$, where the arguments $(\cdot)$ and $(\cdot\cdot)$ are the arguments of $f_{\varepsilon_t^+}$ and $f_{\varepsilon_t^-}$ preceding in the same product, respectively; let $\mathcal{J} \stackrel{\text{def}}{=} J\big(F_{\nu_t^+}(\cdots), F_{\nu_t^-}(\cdots\cdot)\big)$, where the arguments $(\cdots)$ and $(\cdots\cdot)$ are the arguments of $f_{\nu_{i;t}^+}$ and $f_{\nu_{j;t}^-}$ preceding in the same product, respectively; and finally, let $\mathcal{C} \stackrel{\text{def}}{=} C(\mathcal{S}, \mathcal{J})$, $\mathcal{C}_{i\cdot} \stackrel{\text{def}}{=} C\big(\mathcal{S}, F_{\nu_{i;t}^+}(\cdots)\big)$, $\mathcal{C}_{\cdot j} \stackrel{\text{def}}{=} C\big(\mathcal{S}, F_{\nu_{j;t}^-}(\cdots\cdot)\big)$. Then,

$$\begin{aligned} f_{\psi_t}(z) = &\int_{\operatorname{supp}\varepsilon_t^-} f_{\varepsilon_t^+}(z-y)\cdot f_{\varepsilon_t^-}(y)\cdot \mathcal{S}^{11}\,\mathrm{d}y\cdot\mathbb{P}(n_t^+=0,n_t^-=0) \\ &+\sum_{i\in\mathbb{N}}\int_{\operatorname{supp}\nu_t^+}\int_{\operatorname{supp}\varepsilon_t^-} f_{\varepsilon_t^+}(z-y-u)\cdot f_{\varepsilon_t^-}(y)\cdot f_{\nu_{i;t}^+}(u)\cdot \\ &\cdot\left(\mathcal{S}^{11}\mathcal{C}_{i\cdot}^{11}+\mathcal{S}^{01}\mathcal{S}^{10}\mathcal{C}_{i\cdot}^{21}\right)\,\mathrm{d}y\,\mathrm{d}u\cdot\mathbb{P}(n_t^+=i,n_t^-=0) \\ &+\sum_{j\in\mathbb{N}}\int_{\operatorname{supp}\nu_t^-}\int_{\operatorname{supp}\varepsilon_t^-} f_{\varepsilon_t^+}(z-y-v)\cdot f_{\varepsilon_t^-}(y)\cdot f_{\nu_{j;t}^-}(v)\cdot \\ &\cdot\left(\mathcal{S}^{11}\mathcal{C}_{\cdot j}^{11}+\mathcal{S}^{01}\mathcal{S}^{10}\mathcal{C}_{\cdot j}^{21}\right)\,\mathrm{d}y\,\mathrm{d}v\cdot\mathbb{P}(n_t^+=0,n_t^-=j) \\ &+\sum_{i\in\mathbb{N}}\sum_{j\in\mathbb{N}}\int_{\operatorname{supp}\nu_t^-}\int_{\operatorname{supp}\nu_t^+}\int_{\operatorname{supp}\varepsilon_t^-} f_{\varepsilon_t^+}(z-y-u-v)\cdot f_{\varepsilon_t^-}(y)\cdot f_{\nu_{i;t}^+}(u)\cdot f_{\nu_{i;t}^-}(v)\cdot \\ &\cdot\left[\mathcal{J}^{11}(\mathcal{C}^{11}\mathcal{S}^{11}+\mathcal{S}^{01}\mathcal{S}^{10}\mathcal{C}^{21})+\mathcal{J}^{01}\mathcal{J}^{10}(\mathcal{S}^{11}\mathcal{C}^{12}+\mathcal{S}^{01}\mathcal{S}^{10}\mathcal{C}^{22})\right] \\ &\quad\mathrm{d}y\,\mathrm{d}u\,\mathrm{d}v\cdot\mathbb{P}(n_t^+=i,n_t^-=j). \end{aligned} \tag{2.3.5}$$

Because i and j take integer values up to infinity because n_t^+ and n_t^- are Poisson random variables, we must apply truncation and re-normalisation of probabilities.[32]

We again assume that the parameter of the Poisson distribution of the number of jumps might vary over time for both 'good' and 'bad' jumps, and dynamics similar to those of unified jumps can be introduced:

$$\begin{cases} \lambda_t^+ = \delta_0^+ + \delta_1^+\lambda_{t-1}^+ + \delta_2^+\pi_{t-1}^+, \\ \lambda_t^- = \delta_0^- + \delta_1^-\lambda_{t-1}^- + \delta_2^-\pi_{t-1}^-, \end{cases}$$

where the intensity residuals π_{t-1}^+ and π_{t-1}^- are computed separately based on (2.3.3) and probabilities $\mathbb{P}(n_t^+ = i \mid \Omega_t)$ and $\mathbb{P}(n_t^- = j \mid \Omega_t)$ based on (2.3.4) with parameters governing positive and negative dynamics separately, assuming independence of n_t^+

[32] This version of the model is computationally intensive; thus, we use $i = 0, \ldots, 3$ and $j = 0, \ldots, 3$. The resulting expression for $f_{\psi_t}(z)$ in this case contains 1 single, 6 double, and 9 triple integrals.

and n_t^- for computational simplicity.

2.4 Estimation

In the most general model without jumps, the parameters to be estimated are the 4 of the 'good' scale dynamics (α_0, α_1, α_2, α_2^-); 4 of the 'bad' scale dynamics (β_0, β_1, β_2, β_2^-); 2 shapes (θ^+, θ^-); 1 mean (μ), and the copula parameters, of which there is just 1 in the static case, γ_0, and 3 more, γ_1, γ_2, and γ_2^-, in the dynamic case;[33] thus, there are 15 parameters in total to estimate. Following Bekaert et al. (2015), if a dynamic shape is used instead of a dynamic scale, nothing changes fundamentally.

The addition of jumps also does not change the following estimation procedure if the researcher adjusts the definition of f_{r_t}. However, the number of parameters increases dramatically, while, due to the recurrence relation in (2.3.3), it becomes impossible to parallelise computations efficiently, so estimation routines for models with jumps (including inference) take more time.

Because the likelihood function is a product of conditional densities, its maximisation problem is

$$\max_{\substack{\{\alpha\},\{\beta\},\{\gamma\},\\ \theta^+,\theta^-,\mu}} \sum_{t=1}^{T} \log f_{r_t}(z \mid \Omega_{t-1}) = \max \mathcal{L}(\alpha_0, \ldots, \beta_0, \ldots, \gamma_0, \ldots, \theta^+, \theta^-, \mu), \tag{2.4.1}$$

where the expression for the conditional density at time t is that in (2.2.5). The number of integrals being evaluated for one value of the parameter vector is equal to the number of time periods. Such likelihood functions require certain precautions during optimisation, which we discuss below. As mentioned earlier, Bekaert et al. (2015) provide an analytical solution but still use numerical integration techniques; therefore, the proposed approaches are also comparable in terms of estimation accuracy, and in Appendix 2.A, we show the replication and improvement of original estimation results from Bekaert et al. (2015). We do not use data other than the return series, such as realised volatility measures, in model estimation.

For certain copula functions, the parameter κ_t must belong to a specific range. However, large absolute values of r_t severely limit the admissible range of γ_2 and γ_2^-, and values of κ_t near the boundary make the joint density near-degenerate. This problem can be addressed by applying a transformation (e. g. a sigmoid or a strictly increasing non-negative function) to the series from Equation (2.2.7). We use a dampened version of the κ_t series obtained after applying slow-growing Lipschitz-continuous functions, even if the copula allows any parameter from $\mathbb{R}$ (see Appendix 2.B in Table 2.B.1).

We directly integrate densities that appear in the likelihood function. At this step, we use four techniques to improve numerical stability. First, we normalise the integrand by $\sqrt{\sigma^2_{\varepsilon_t^+} + \sigma^2_{\varepsilon_t^-}}$ to ensure that the region where the function should be evaluated is

[33] As mentioned earlier, all distributions in this paper come from a scale-shape family. Should a baseline distribution that does not have an explicit scale parameter governing the variance be chosen, the volatility can be incorporated into it via multiplication: $\tilde{f}_{\varepsilon_t^+}(x) = \tilde{f}_{e_t^+}(x/\sigma_{\varepsilon_t^+})/\sigma_{\varepsilon_t^+}$, and similarly for $\tilde{f}_{\varepsilon_t^-}(x)$.

always captured by the built-in quadrature.[34] Second, when the integrand has limited support, we set the limits explicitly instead of relying on integration from $-\infty$ to $+\infty$, which has good coverage only in a narrow region (e. g. the integral in (2.2.4) is evaluated from $-\infty$ to $\sigma_{\varepsilon_t^-}\overline{e}_t^-$). Third, we use a small relative tolerance (10^{-10} for specifications without jumps and 10^{-7} for specifications with jumps) as the minimum requested accuracy of the Gauss-Kronrod quadrature to eliminate the chance of premature convergence. Finally, during the computation of log-likelihood derivatives, we compute the entire series of numerical derivatives for individual $\log f_{r_t}$ prior to summation in Eq. (2.4.1) *for multiple difference steps*, and then take the median numerical derivative for each t. We use medians of two-sided numerical derivatives across 5 difference steps equally spread on a logarithmic scale (e. g. $10^{-6} \cdot (0.25, 0.5, 1, 2, 4)$).

Derivative-based optimisation techniques can yield accurate results but require a good starting point. Because the target function is highly non-linear in parameters and might have multiple local optima, we use the following approach. In the first step, we use a method of derivative-free meta-heuristics called differential evolution global optimisation. The implementation of differential evolution is based on Ardia et al. (2011), is reasonably fast, and in simulations, outperforms other methods, such as particle swarm and generalised simulated annealing, in terms of speed and convergence. Differential evolution is a smart brute-force approach that randomly generates an initial population of parameter values inside a multidimensional hypercube with sufficiently wide boundaries in each dimension and produces further parameter populations based on the function values that target the global optimum using stochastic merging. This procedure is robust to the existence of multiple local optima. We choose appropriate hyperparameters of the differential evolution, such as strategy, crossover probability, and differential weighting, based on simulation results with a known DGP. After a good initial value has been found, we search for the optimum using the BFGS method.[35] All coefficients should be properly scaled during optimisation to make numerical derivatives more reliable. For forecasting purposes, we use rolling-window re-estimation. In the first sub-sample, we maximise the likelihood in two steps. For subsequent sub-samples, we use the warm start of the BFGS algorithm from the optimum in the previous window and update the estimates. Re-estimation from the previous optimum is rather fast and can be run on a daily basis on a standard computer.

For inference, we compute the QML standard errors numerically (White, 1982). We use the first technique described above to evaluate the Hessian, using 3 differences at each step of repeated differencing, which indicates that the final result is the median difference of median differences; thus, the breakdown point of this approach is high.

[34] If the computed value of an individual likelihood function is below the machine epsilon and there is possible premature convergence ($f_{r_t} < 1.49 \cdot 10^{-8}$), then we make two more attempts of integration, with scaling factors equal to unity and the median of all $(\sigma_{\varepsilon_t^+}, \sigma_{\varepsilon_t^-})$—and retain the maximum value of the three integration procedures.

[35] Constraints do not need to be applied as long as the likelihood function is defined to be zero if any of the shape or scale parameters is negative or if the copula parameter lies outside the admissible range because no damping is applied.

2.5 Simulation

To check the finite-sample performance of the proposed model and ensure that the global optimum can be attained and the true values recovered, we simulate a known DGP and then estimate the structural parameters of the model and check the distribution of the estimates. We simulate a process with independent shocks with non-centred gamma distribution, dynamic scale, and returns with de-meaning. The parameters of the true DGP are based on the estimates of one of the specifications, specifically, from the one with non-centred gamma shocks:

$$\begin{cases} \sigma^2_{\varepsilon^+_t} = 2.5 \cdot 10^{-7} + 0.9\sigma^2_{\varepsilon^+_{t-1}} - 0.004\tilde{r}^2_{t-1} + 0.03\tilde{r}^2_{t-1}\mathbb{I}^-_{t-1}, \\ \sigma^2_{\varepsilon^-_t} = 6.0 \cdot 10^{-7} + 0.9\sigma^2_{\varepsilon^-_{t-1}} - 0.008\tilde{r}^2_{t-1} + 0.05\tilde{r}^2_{t-1}\mathbb{I}^-_{t-1}, \end{cases} \quad \theta^+ = 2.5,\ \theta^- = 2,\ \mu = 0.0004. \tag{2.5.1}$$

In this simulation, we generate independent shocks but estimate a specification with a static AMH copula, for which the independence corresponds to $\gamma_0 = 0$; thus, the distribution of $\hat{\gamma}_0$ should be centred around zero. We show that even if one estimates a more general model, it recovers the parameters of the simpler underlying model, which verifies its robustness. However, in real estimation, the copula matters in most cases; thus, the probability of a false positive is low.

We generate synthetic returns per the following algorithm:

1. Initialise the series: draw $e^+_1 \sim \Gamma(\theta^+, 1)$ and $e^-_1 \sim -\Gamma(\theta^-, 1)$, $r_1 = \mu$, $\sigma^2_{\varepsilon^+_1} = 2e^+_1\mu^2$, $\sigma^2_{\varepsilon^-_1} = -3e^-_1\mu^2$.
2. In a loop for $t = 2, \dots, \lfloor 1.1T \rfloor$, at every step, compute $\sigma^2_{\varepsilon^+_t}$ and $\sigma^2_{\varepsilon^-_t}$ based on (2.5.1), draw $e^+_t \sim \Gamma(\theta^+, 1)$ and $e^-_t \sim -\Gamma(\theta^-, 1)$, compute $\varepsilon^+_t = \sqrt{\sigma^2_{\varepsilon^+_t}}e^+_t$ and $\varepsilon^-_t = \sqrt{\sigma^2_{\varepsilon^-_t}}e^-_t$, and compute $r_t = \mu + \varepsilon^+_t + \varepsilon^-_t$.
3. Discard the first 10% of observations to eliminate the initial-value effect.

We generate 100 samples of length $T = 10{,}000$ based on the aforementioned DGP and estimate the model using the BFGS optimiser.[36] We use the true family of densities in estimation, and the distribution of estimates is shown in Figure 2.5.1.

The plots show that the estimator is centred around the true parameter values, the median bias is near zero, and its distribution is near normal, which indicates that the asymptotic properties of the ML estimator hold for $n = 10{,}000$. Therefore, we conclude that it is possible to recover the true parameters using a large sample, even if a more general model was estimated.

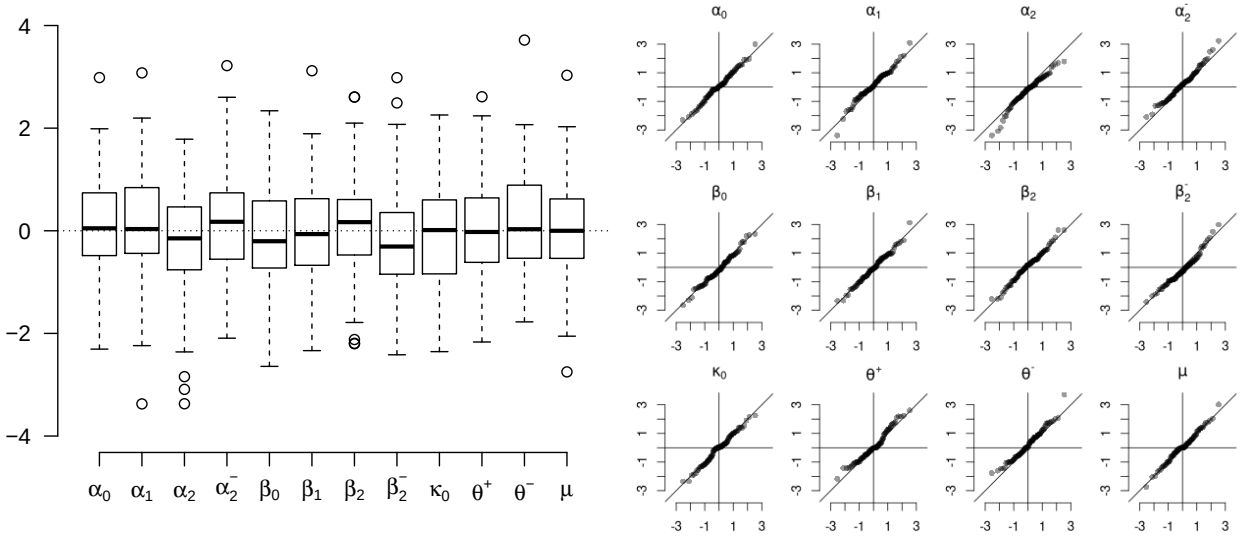
2.6 Data and competing models

2.6.1 Data

We use daily price data of ETF on S&P 500 (SPDR) from Bloomberg (2000-01-03 to 2019-05-31). The models are estimated using the data from between 2000-01-03 and 2014-12-31 (3773 observations, 15 years) and tested using data from between 2015-01-02 and 2019-05-31 (1110 observations, 4.5 years). We use a rolling-window forecasting procedure and re-estimate the parameters every two observations. We select an ETF

[36] We use the true parameter vector as the starting value.

Figure 2.5.1: Simulation results



The box plot on the left shows the distribution of centred (i. e. with the true value subtracted) ML estimates divided by their Monte Carlo standard deviations. The bottom, middle, and top lines of each box denote the 25%, 50%, 75% percentiles of the distribution. The whiskers show the last value within 1.5 interquartile ranges of the box edges. The circles show any values more extreme than the endpoints of the whiskers. Parameter names correspond to formulæ (2.2.6)–(2.2.7).
The Q-Q plot on the right compares the quantiles of the standard normal distribution (horizontal axis) with the quantiles of the centred ML estimator divided by its Monte Carlo standard deviations (vertical axis). The diagonal line has a slope of 45° and goes through the origin.

on the index because its volatility approximates systematic risk; therefore, we can better understand the behaviour of one of the primary driving forces of returns. We use a rather large estimation sample to obtain more accurate estimates of the deep parameters, such as those in copula dynamics. Because many specifications are present, false positives can occur. It is difficult to control the type I error probability using the procedure from Romano and Wolf (2005) because the amount of time required for bootstrapping of the proposed model is large; however, we believe that the proposed results are accurate and reliable due to both different in-sample and out-of-sample comparisons and large sample size.

2.6.2 Competing models

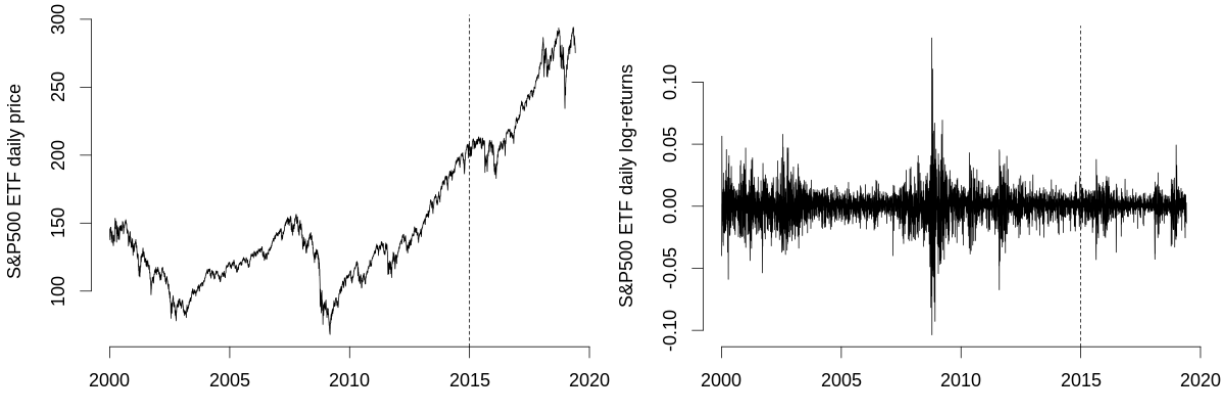
We compare the proposed model with 40 popular symmetric and asymmetric GARCH models that can be written using a general all-in-family GARCH formula. We use specifications without external regressors, as in Ghalanos (2020), and a constant mean of returns:

$$\sigma_t^\lambda = \omega + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^\lambda + \sum_{j=1}^{q} \alpha_j \sigma_{t-j}^\lambda \big(|z_{t-j} - \eta_{2j}| - \eta_{1j}(z_{t-j} - \eta_{2j}) \big)^\delta,$$

where $z_t \overset{\text{def}}{=} (r_t - \mu)/\sigma_t$ are normalised returns, α_j are the ARCH parameters, β_i are the GARCH parameters, and other Greek letters correspond to the following variants:

1. The standard GARCH: $\lambda = \delta = 2$, $\eta_{1j} = \eta_{2j} = 0$;
2. GJR-GARCH: $\lambda = \delta = 2$, $\eta_{2j} = 0$;

Figure 2.6.1: U.S. market data series used for the main analysis



The left panel shows daily closing prices for the SPDR S&P 500 ETF. The right panel shows corresponding logarithmic returns. The vertical lines denote the split between the train and test samples.

3. EGARCH: $\delta = 1$, $\lambda = 0$, $\eta_{2j} = 0$;
4. APARCH: $\delta = \lambda$, $\eta_{2j} = 0$, $|\eta_{1j}| \leq 1$;
5. TGARCH: $\lambda = \delta = 1$, $\eta_{2j} = 0$, $|\eta_{1j}| \leq 1$;
6. AVGARCH: $\lambda = \delta = 1$, $|\eta_{1j}| \leq 1$;
7. NGARCH: $\delta = \lambda$, $\eta_{1j} = \eta_{2j} = 0$;
8. NAGARCH: $\lambda = \delta = 2$, $\eta_{1j} = 0$;
9. Full family GARCH, or ALLGARCH: $\lambda = \delta$.

Another model that we also use for benchmarking that is not a case of the general formula above is component GARCH (CSGARCH): $\sigma_t^2 = q_t + \sum_{i=1}^{q} \alpha_i(\varepsilon_{t-i}^2 - q_{t-i}) + \sum_{i=1}^{p} \beta_i(\sigma_{t-i}^2 - q_{t-i})$, $q_t = \omega + \rho q_{t-1} + \varphi(\varepsilon_{t-i}^2 - \sigma_{t-i}^2)$.

We investigate four distributions for the density of baseline shocks: normal, skew-normal, Student, and skew-Student. Skewness is introduced via a technique described in Fernández and Steel (1998): if a density $f_X(x)$ is symmetric and unimodal around zero, then its skewed generalisation is defined piecewise with one side stretched and the other side shrunk by a factor s:

$$f_X^{\text{skew}}(x) = \frac{2}{s + 1/s}\big(f_X(x/s)\mathbb{I}(x \geq 0) + f_X(sx)\mathbb{I}(x < 0)\big). \tag{2.6.1}$$

We use the implementation of various GARCH estimators due to Ghalanos (2020) and specify the same rolling estimation windows and warm-start method as we do in the proposed model.

2.7 Results for models without jumps

2.7.1 Out-of-sample and in-sample performance on S&P 500 data

We begin with out-of-sample forecasting as the most important performance measure of GARCH-like models and compare the specifications by the accuracy of out-of-sample predictions of two primary risk measures: VaR and variance. As an initial check, we calculate the violation ratio, which is the number of observed VaR exceedances divided by their expected number; for the true model, this proportion is equal to 1. Additionally, we use three statistical tests for VaR. The conditional coverage (CC) test (Christoffersen,

1998) simultaneously checks the equality of a proportion of VaR exceedances among all testing periods to a confidence level and the clustering of VaR exceedances by testing the equality of conditional probabilities of a two-state Markov chain to unconditional probabilities. The dynamic quantile (DQ) test of Engle and Manganelli (2004) checks the same properties using external predictors. Lastly, the duration test (Christoffersen & Pelletier, 2004) is based on the idea that the number of days between VaR violations should not be clustered.

Next, we compare the proposed models with well-established GARCH variants in terms of variance forecast quality. To evaluate the relative performance of the forecasts between the two groups of models, we use the Diebold-Mariano test (Diebold & Mariano, 2002) and report percentages of models better, equivalent, or worse in terms of forecast quality in Table 2.7.2. We use two popular loss functions for this test, RMSE and QLIKE, to compare predicted variances with realised variances from the 'Oxford Realised Library' (Heber et al., 2009). We choose these two functions because Patton (2011) shows that these loss functions are robust to noise in the variance proxy. Additionally, the QLIKE loss is more robust to extreme observations in the sample. The results are shown in Table 2.7.2 (left panel). Then, we compare all specifications using the Vuong test (Vuong (1989) with a HAC correction based on Calvet and Fisher (2004)) to compare differences in likelihood and Akaike criteria (Table 2.7.2, right panel). For details of the modified Vuong test, see Appendix 2.C.

Table 2.7.1: Out-of-sample model tests

Model	Distrib.	VR	p_{CC}	p_{DQ}	p_{dur}
CSGARCH	skew-$\mathcal{N}$	0.855	0.260	0.061	0.079
GJR-GARCH	skew-t	0.818	0.102	0.107	0.737
APARCH	skew-$\mathcal{N}$	0.782	0.326	0.117	0.404
APARCH	skew-t	0.764	0.263	0.070	0.292
TGARCH	skew-t	0.745	0.263	0.070	0.294
ALLGARCH	skew-$\mathcal{N}$	0.727	0.395	0.196	0.741
TGARCH	skew-$\mathcal{N}$	0.727	0.360	0.118	0.186
EGARCH	skew-t	0.727	0.354	0.398	0.335
GJR-GARCH	skew-$\mathcal{N}$	0.727	0.263	0.109	0.743
AVGARCH	skew-t	0.673	0.236	0.098	0.239
EGARCH	skew-$\mathcal{N}$	0.673	0.610	0.498	0.274
NAGARCH	skew-$\mathcal{N}$	0.655	0.294	0.109	0.884
AVGARCH	skew-$\mathcal{N}$	0.655	0.600	0.314	0.149

Distrib.	Copula	Dyn.	VR	p_{CC}	p_{DQ}	p_{dur}
c-log-log	Clayt.	$\tilde{r}^3$	0.964	0.628	0.233	0.807
log-log	Clayt.	$\tilde{r}^3$	0.855	0.080	0.185	0.465
c-log-log	Plack.	$\tilde{r}^3$	0.855	0.203	0.059	0.833
c-log-log	Plack.	stat.	0.818	0.112	0.136	0.822
c-log-log	Plack.	$\tilde{r}^2$	0.818	0.112	0.093	0.469
gamma	Plack.	stat.	0.782	0.126	0.345	0.321
c-gamma	AMH	$\tilde{r}^3$	0.782	0.055	0.062	0.733
log-log	Plack.	stat.	0.782	0.126	0.263	0.884
log-log	Plack.	$\tilde{r}^3$	0.782	0.126	0.346	0.776
log-log	cubic	$\tilde{r}^3$	0.782	0.126	0.252	0.884
gamma	Plack.	$\tilde{r}^2$	0.764	0.088	0.282	0.382
logl-log	indep.	—	0.764	0.088	0.198	0.591
logl-log	cubic	stat.	0.764	0.088	0.199	0.591

The left table represents all well-established GARCH models (13 specifications) that do not fail any of the three VaR out-of-sample tests. The right table represents the top 13 opposite-sign-shock model specifications. The results are sorted based on violation ratio values.
Distrib.: 'c-' is used for centred distributions, 'log-log' for log-logistic, t for Student's t, and $\mathcal{N}$ for Gaussian.
Dyn.: variable used in the dynamics of the copula parameter ($\tilde{r}^3$ for cubes of centred returns, $\tilde{r}^2$ for squares of centred returns).
VR: violation ratio (the ratio of exceedances and their expected number).
p_{CC}: p value of the conditional coverage test (Christoffersen, 1998).
p_{DQ} p value of the dynamic quantile test (Engle & Manganelli, 2004) with 4 lags of VaR exceedances, lag of squared returns, and predicted VaR as regressors.
p_{dur}: p value of the no-hit duration test (Christoffersen & Pelletier, 2004).

The out-of-sample performance is discussed next (Table 2.7.1). To save space, we present only the top 13 of the proposed specifications that did not fail any of the three VaR tests and all well-established models that did not fail the same tests (13 specifications), ranked based on the violation ratio. The specifications that have

violation ratios greater than 1 fail all three VaR out-of-sample tests. The full results of Christoffersen (1998) and Engle and Manganelli (2004) tests are shown in Figure 2.A.2 in Appendix 2.A and provide the full picture for the relative performance of opposite-sign-shock models and well-established models. We should mention that specifications with dynamic shapes (BEGE models) have poor out-of-sample performance: only 2 out of 39 specifications have p values for all VaR tests greater than 0.05. Consequently, they are not shown in Table 2.7.1.

Regarding the VaR tests among the 40 well-established GARCH models, 43% do not fail the conditional coverage test, 38% do not fail the dynamic quantile test, and 98% do not fail the duration test at the 5% significance level. For the proposed model, the fractions of all specifications (out of 50 variants) that do not fail those tests at the same level of significance are 46% (CC), 54% (DQ), and 96% (duration). Therefore, the opposite-sign-shock model has higher rates of non-rejection on average, and the violation ratios are also much closer to 1. For the well-established GARCH models, the skewed distributions explicitly dominate symmetric distributions; however, heavy tails are not important for the S&P 500 data (which is rather expected for the market returns). There is no clear pattern in the preferred variance processes; it is impossible to select the best variance specification for the S&P 500 data. However, the overwhelming majority of well-established models have an asymmetric reaction of variance to the previous shocks; therefore, the leverage effect is critical for the correct modelling of S&P 500 variance.

Among the specifications of the opposite-sign-shock model, the most striking pattern is the fact that only models containing a copula take the top positions, which indicates that the independence assumption is not correct for the U.S. market. The Clayton and Plackett copulæ (particularly the latter) provide more accurate results than other ones. The Clayton copula has stronger left-tail dependence for marginal probabilities and assumes only positive correlations; however, the Plackett copula has a symmetric dependence and yields correlations of any sign. Therefore, it is difficult to determine the most appropriate dependence structure for shocks in the U.S. market. However, dynamic copula parameter specifications are preferred over static ones; therefore, the correlation between shocks is likely to be time-dependent. Specifications with a heavy-tailed (log-logistic) distribution are marginally better than those with gamma distribution, which can be explained by the fact that in opposite-sign-shock models, the right and left tails have distinct behaviour. Therefore, such a discrepancy between the results is a signal that the left and right parts of the return distribution are not equal. There are fewer opposite-sign-shock specifications with centred distributions; however, they take the top positions in the table. Therefore, the appropriate risk-return relationship is either highly non-linear and complex or simply insignificant. However, these results also indicate that in any case, the influence of volatility on returns is rather heterogeneous.

Regarding variance forecasting, we compare the top 13 proposed models sorted according to the violation ratio based on the QLIKE criterion in Table 2.7.2. Specifications with centred log-logistic distributional assumptions outperform other models in terms of loss functions in general. Specifications with dynamic Plackett and Clayton copulæ provide the most accurate results. Model performance based on RMSE loss function has similar patterns: the performance of our models is marginally worse according to this criterion. However, the best model for VaR forecasts also has the best variance forecasting accuracy here.

From the out-of-sample results, we can determine a set of the most accurate specific-

Table 2.7.2: Percentage of competing models beaten in terms of quality indicators by opposite-sign-shock models

Distrib.	Copula	Dyn.	RMSE			QLIKE		
			+	=	−	+	=	−
c-log-log	Clayton	$\tilde{r}^3$	24	76	0	88	12	0
log-log	Clayton	$\tilde{r}^3$	0	35	65	6	41	53
c-log-log	Plackett	$\tilde{r}^3$	0	100	0	94	6	0
c-log-log	Plackett	static	6	94	0	94	6	0
c-log-log	Plackett	$\tilde{r}^2$	0	88	12	47	53	0
gamma	Plackett	static	0	53	47	6	0	94
c-gamma	AMH	$\tilde{r}^3$	0	88	12	29	65	6
log-log	Plackett	static	0	71	29	29	53	18
log-log	Plackett	$\tilde{r}^3$	0	88	12	47	53	0
log-log	cubic	$\tilde{r}^3$	0	53	47	18	29	53
gamma	Plackett	$\tilde{r}^3$	0	59	41	18	29	53
log-log	indep.	—	0	71	29	29	53	18
log-log	cubic	static	0	88	12	29	53	18

Distrib.	Copula	Dyn.	Vuong (LL)			Vuong (AIC)		
			+	=	−	+	=	−
c-log-log	Clayton	$\tilde{r}^3$	100	0	0	100	0	0
log-log	Clayton	$\tilde{r}^3$	94	0	6	88	0	12
c-log-log	Plackett	$\tilde{r}^3$	94	0	6	94	0	6
c-log-log	Plackett	static	94	0	6	94	0	6
c-log-log	Plackett	$\tilde{r}^2$	94	0	6	94	0	6
gamma	Plackett	static	65	0	35	65	0	35
c-gamma	AMH	$\tilde{r}^3$	100	0	0	100	0	0
log-log	Plackett	static	88	0	12	82	0	18
log-log	Plackett	r^3	100	0	0	94	0	6
log-log	cubic	$\tilde{r}^3$	82	6	12	82	0	18
gamma	Plackett	$\tilde{r}^3$	100	0	0	94	0	6
log-log	indep.	—	88	0	12	82	0	18
log-log	cubic	static	88	0	12	88	0	12

In these tables, we compare the top 13 opposite-sign-shock model specifications that do not fail any VaR tests, sorted based on violation ratio in descending order to the pool of well-established GARCH models that did not fail the same tests. The numbers show a percentage (%) of competing GARCH variants beaten (+), matched (=), or outperformed (−) by opposite-sign-shock models. The left sub-table compares out-of-sample performance with Diebold-Mariano tests for variance forecast quality using the RMSE and QLIKE loss functions. The right sub-table compares in-sample performance using Vuong's (1989) LR-based test for the equivalence of Kullback-Leibler (or pure log-likelihood, denoted by LL) and AIC criteria of non-nested models with Calvet-Fisher (2004) HAC with Bartlett kernel and Newey-West lags. For a rigorous definition of the Vuong test statistic, see Appendix 2.C.
We use annualised realised variances for comparison, as in Heber et al. (2009), with 252 days in a year. These variances are equal to 252 times the daily variance.
Only models with dynamic scales are presented.
Distrib.: 'c-' for centred distributions, 'log-log' for log-logistic.
Dyn.: variable used in the dynamics of the copula parameter ($\tilde{r}^3$ for cubes of centred returns, $\tilde{r}^2$ for squares of centred returns).

ations. From the opposite-sign-shock models, the best-performing model is that with a centred log-logistic distribution and a Clayton copula with cubed de-meaned returns in parameter dynamics. This specification has consistent results, high p-values for all VaR tests, and a violation ratio near one, which is arguably the critical result because the true model should generate 5-percent VaR values that would be exceeded on average 5% of the time, and clustering of the VaR can occur purely by chance. This model accurate variance forecasting according to QLIKE and RMSE loss functions. It is more difficult to select the best specification from the well-established GARCH family due to the instability of their performance. Those models with a high violation ratio (CSG-ARCH with skew-normal distribution, GJR-GARCH with skew-Student distribution) or high p-values for VaR tests (EGARCH with skew-normal distribution, EGARCH with skew-Student distribution, AVGARCH with skew-normal distribution, ALLGARCH with skew-normal distribution, NAGARCH with skew-normal distribution, etc.) do not provide precise variance forecasts in either metric. An ALLGARCH specification with skew-normal distribution is that with the most stable results, perhaps because that this model is the most general.

Now, we turn to in-sample performance (we use the same models as before). We use the Vuong test based on likelihood and AICs for comparisons.[37] In terms of in-sample performance, opposite-sign-shock models clearly dominate well-established GARCH models according to both tests.[38] The centred log-logistic distribution model with the Clayton copula with a dynamic parameter containing cubed de-meaned returns, while being the most accurate one in the out-of-sample results, has the best in-sample performance too. Therefore, our results hold both in-sample and out-of-sample. Small modifications of this specification, such as assuming non-centred log-logistic distribution instead of centred log-logistic, or changing Clayton copula to Plackett, does not significantly affect the fit quality or out-of-sample forecast quality.

To test the correctness of the specification of the best opposite-sign-shocks model mentioned above, we use a test due to González-Rivera and Sun (2015). It is based on the idea from Diebold et al. (1998) that the probability integral transform (PIT) of the data-generating process realisations with respect to the predictive density is uniformly distributed under the null hypothesis that the predictive density coincides with the true density. These predicted values of the conditional distribution function are called *generalised residuals*. As the authors notice, a rejection of the null based on a simple Kolmogorov-Smirnov test for the uniformity of $\hat{F}_{r_t}(r_t \mid \Omega_{t-1})$ is not informative because it can occur due to either dependent observations or non-uniform observations (or even both), which is further elaborated on by Kheifets (2015). González-Rivera and Sun (2015) propose a generalised-autocontour-based (G-ACR) test for generalised residuals that verifies that $(\hat{F}_{r_t}(r_t \mid \Omega_{t-1}), \hat{F}_{r_{t-k}}(r_{t-k} \mid \Omega_{t-k-1}))$ are bi-variate uniform by constructing generalised autocontours ($[0, \sqrt{\alpha}]^2$-squares, where $0 < \alpha < 1$) and checking if the empirical share of observations $(\hat{F}_{r_t}(r_t \mid \Omega_{t-1}), \hat{F}_{r_{t-k}}(r_{t-k} \mid \Omega_{t-k-1}))_{t=k+1}^{T}$ falling within the area defined by the autocontour is close to α. This should hold for any integer value of the lag k. Under the null hypothesis, the difference of sample

[37] We also tested the equivalence of Bayesian information criteria for the subsets of opposite-sign-shock and well-established models but are not including it as extra columns of Table 2.7.2; the test statistic is described in Appendix 2.C. On average, opposite-sign-shock models dominate 90% of the well-established models by log-likelihood, 87% by AIC and 72% by BIC. These results demonstrate that even with a heavy penalisation for over-parametrisation, the proposed models yield a better in-sample fit on average.

[38] In Table 2.A.2, we provide results for Bekaert et al. (2015)-like models, and some of these specifications have excellent in-sample performance.

proportions and theoretical square size is normally distributed with mean 0 and a known variance. Since this statistic depends on two parameters (lag k and square size α), the authors propose two more statistics for testing joint hypotheses: multiple lags at once (denoted by $P_2(1,\ldots,k;\alpha) \sim \chi^2_k$ since it comes from Proposition 2 in the article) or multiple square sizes at once (denoted by $P_3(k;\alpha_1,\ldots,\alpha_c) \sim \chi^2_c$) in order to better distinguish between the IID assumption violation and uniformity assumption violation. The results for uniformity testing of $\hat{F}_{r_t}(r_t \mid \Omega_{t-1})$ are given in Table 2.7.3. In multiple tests, there were no rejections of the null hypothesis even at 10% level, which backs up the idea that the proposed density is rather close to the true density. This result indicates that there is no significant correlation or orders 1 and 2 in the generalised residuals, implying no persistence in conditional distribution forecast errors and no significant deviation from uniformity, implying no substantial structural change in the data-generating process.

Table 2.7.3: Specification tests based on predictive densities

Test	KS	G-ACR P_2		G-ACR P_3	
Lag used (k)	—	1	1 and 2	1	2
In sample ($t = 1,\ldots,3773$)	0.480	0.718	0.455	0.361	0.658
Out of sample ($t = 3774,\ldots,4883$)	0.102	0.137	0.327	0.108	0.287
Full sample ($t = 1,\ldots,4883$)	0.171	0.300	0.320	0.332	0.890

This table contains the p-values of several tests of the null hypothesis: the PIT of r_t is IID uniform on $[0,1]$. KS: Kolmogorov-Smirnov test. G-ACR: generalised-autocontour-based test (P_2: multiple lags at once, P_3: multiple autocontour sizes at once). P_2 uses $\alpha = 0.5$, as in González-Rivera and Sun (2015). P_3 uses $\alpha = \{0.05, 0.10, \ldots, 0.95\}$ (19 autocontours). The results of the G-ACR test are similar when a different set of contours is used: we tried $\{0.15, 0.20, \ldots, 0.85\}$ and $\{0.3, 0.4, \ldots, 0.7\}$, and all p-values were greater than 0.1 in these cases.

2.7.2 Estimation results on IBM data

To verify that the best models described above perform well for other stock returns and our results are not false-positive given the high time necessary for the procedure of Romano and Wolf (2005), we used a different data set for estimation and testing as a robustness check. We chose 15 years of daily IBM stock returns from between 2000-01-04 and 2014-12-31 for estimation (3772 observations) and from between 2015-01-02 and 2019-05-31 for testing (1110 observations) because it is one of the most popular stocks in such studies, and the time frames are the same as those of our S&P 500 ETF data. We estimated the same family of models and conducted the same battery of out-of-sample tests to compare opposite-sign-shock models and 40 well-established models.

As in the case with S&P 500 ETF data, the opposite-sign-shock specifications yielding violation ratios near one and not failing conditional coverage and dynamic quantile tests at the 5% level are the specifications with log-logistic and centred log-logistic distributions, and Clayton, Plackett, and Frank copulæ with $\tilde{r}^2$ and $\tilde{r}^3$ in dynamics, or static copulæ. Their violation ratios range from 0.982 to 1.018. The specification with centred log-logistic distribution and static Clayton copula yielded the best in-sample fits (lowest AIC values), and the specifications with centred log-logistic distribution, Clayton copula and dynamic copula parameter (with $\tilde{r}^2$ in its

dynamics) provided variance forecasts with the lowest QLIKE and RMSE criterion values. Therefore, our results hold for this asset as well.

We conclude that applied researchers should consider the log-logistic distribution (centred or non-centred) of unobserved 'good' and 'bad' shocks, and assume that they are connected with Clayton copula with squares or cubes of de-meaned returns in dynamics because the results for this family of specifications seem to be stable in many aspects of out-of-sample forecasting.

2.7.3 Results on the return characteristics

In this section, we examine the proposed best specification in greater detail: centred log-logistic distribution of shocks with a dynamic Clayton copula.

We begin by drawing inferences for the model coefficients (Table 2.7.4). Due to the model structure, these parameters are not directly comparable to those from standard GARCH models. Additionally, insignificance at a given level should not be interpreted as equality of the true parameter to zero, and rather describes estimation uncertainty (high estimate/standard error ratio) because the support of certain parameters, such as distribution shape, is bounded and often in a non-trivial manner. The conditions generating strictly positive $\sigma^2_{\varepsilon^+_t}$ and $\sigma^2_{\varepsilon^-_t}$ given $\{r_t\}_{t=1}^T$ are difficult to verify.

In the equation for 'good' variance, only its lag, $\sigma^2_{\varepsilon^+_{t-1}}$, is significant at the 5% level. However, in the 'bad' variance equation, all coefficients except the constant are significant. Inertia in the 'good' variance is extremely high (i. e. the estimate of α_1 is near 1); however, the autocorrelation of the 'bad' variance, β_1, is lower. Both volatilities are positively related to the asymmetric term; however, only the 'bad' variance has a significance dependence (β_2^-). Even small negative total shocks, which is another name for de-meaned returns, can lead to serious variance increases because both volatilities have high coefficients on asymmetric terms compared to coefficients on ARCH terms. However, positive shocks can decrease volatilities due to the positive coefficients on ARCH terms. This phenomenon becomes more apparent when one compares the magnitude of the coefficients on the asymmetric term: β_2^- is nearly 6 times larger than α_2^-, which indicates that 'bad' variance reacts in a much more sensitive manner to poor news than 'good' variance does. In such a case, the standard GARCH models with the leverage effect can misrepresent the asymmetric dependence between volatility and previous shocks due to the oversimplified structure.

Both α_2 and α_2^- appear individually insignificant according to the p-values. However, the standard errors as a Gaussian asymptotic approximation may be inaccurate or accurate only in the epsilon-neighbourhood around the optimum, and conclusions about insignificance should not be made based solely on those standard errors. It is common in non-linear models that estimation uncertainty about parameter values is higher in one direction than the other. A Wald test of the joint hypothesis $(\alpha_2, \alpha_2^-) = (0, 0)$ is not suitable in this case because if both parameters are equal to zero, then there are no dynamics in the equation for positive variance, and α_1 is not identified. Hansen (1996) gives the null hypothesis of no ARCH effect as an example since under this null, the variance process becomes static, depending on one parameter, and the GARCH effect is not identified, and standard inference fails. Therefore, it is more reasonable to conduct LR tests in order to evaluate significance because they do not require standard errors (Pawitan, 2001, Section 2.7). We test the hypothesis $\alpha_2^- = 0$, which is the most natural simplification of the model (no asymmetry in the positive scale dynamics). The

LR statistic for this hypothesis is equal to 11.62 ($p = 0.0007$); therefore, it should be rejected, and we make a conclusion that omission of the ostensibly insignificant α_2^-, i. e. ignoring the asymmetry in the 'good' variance dynamics, yields a substantially inferior fit. Under this restriction, the estimate of α_2 is equal to 0.233, which corresponds to the logic that both positive and negative returns increase the volatility by a small amount. The LR-based 95% confidence interval for α_2 under the null is $[0.035, 1.27]$, which is highly asymmetrical, which illustrates the point that using standard errors for conclusions about insignificance can be misleading.

From the plot of the volatility dynamics (Figure 2.7.1),[39] the 'bad' volatility is marginally greater than the 'good' volatility. However, a comparison of crisis periods versus normal periods shows that, as can be expected, the 'bad' volatility is much higher during crises than the 'good' volatility. The shape parameter of 'good' shocks is marginally greater than that of 'bad' shocks; therefore, the left tail of the distribution has a lower rate of decay. This fact can be interpreted as extreme negative events occurring with a higher probability compared to positive events. Additionally, 'good' variance constitutes 30% of the total variance on average, and 'bad' variance makes up 38%; thus, 32% of the total variance is explained by the covariance (multiplied by 2) between shocks, meaning that the connection between shocks should be modelled as well. 'Bad' volatility seems to have the strongest influence on total volatility, which supports the findings from (Patton & Sheppard, 2015). The shape parameter for 'good' shocks, θ^+, has higher standard errors, which implies more uncertainty about the exact parameter value.

Table 2.7.4: Estimates and standard errors for the best-performing specification

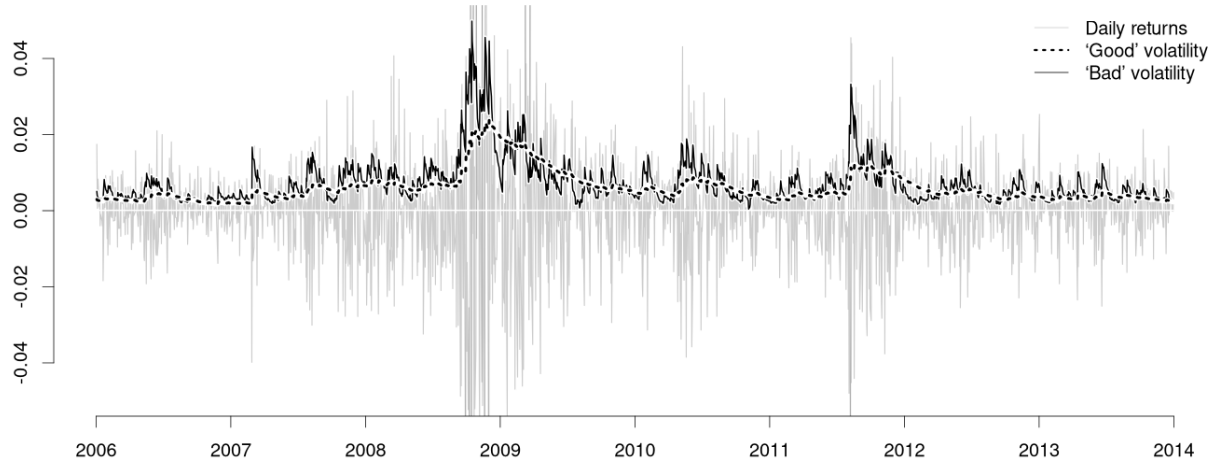
Coefficient	Estimate	t-stat	p	QML t-stat	QML p
α_0	$-1.12 \cdot 10^{-6}$	-0.30	0.762	-0.35	0.730
α_1	0.978	> 10	0.000	> 10	0.000
α_2	-0.201	-0.74	0.461	-0.86	0.391
α_2^-	1.28	1.05	0.292	1.29	0.198
β_0	$3.75 \cdot 10^{-5}$	1.92	0.054	1.92	0.054
β_1	0.810	> 10	0.000	> 10	0.000
β_2	-0.744	< -10	0.000	< -10	0.000
β_2^-	7.79	> 10	0.000	> 10	0.000
γ_0	3.35	1.77	0.076	1.97	0.048
γ_1	-0.310	< -10	0.000	< -10	0.000
γ_2	$-1.36 \cdot 10^6$	-1.95	0.051	-2.10	0.036
γ_2^-	$3.27 \cdot 10^6$	1.96	0.050	2.5126	0.033
θ^+	16.53	2.67	0.008	3.53	0.000
θ^-	11.81	> 10	0.000	> 10	0.000
μ	$9.38 \cdot 10^{-5}$	0.70	0.482	0.70	0.483

Specification: centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$. Estimation sample: 2000-01-03 to 2014-12-31.

The constant parameter in the mean process, μ, is insignificant, leading to a zero-mean process in returns. For the correlation between shocks, the Clayton copula has

[39] We present only 8 years of data to highlight the most tumultuous period while keeping the lines legible.

Figure 2.7.1: 'Good' and 'bad' volatility visualisation



Specification: centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$.

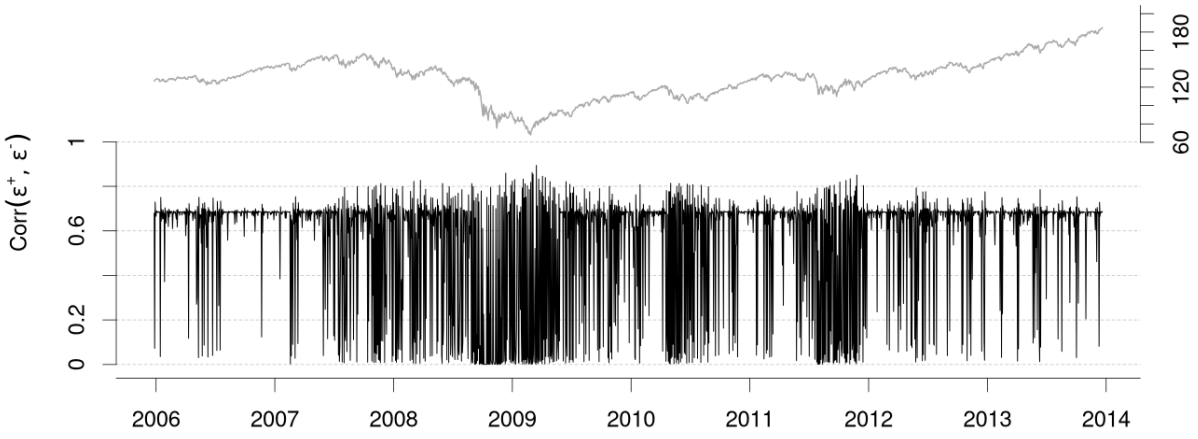
stronger dependence for negative realisations of both shocks; therefore, during a bear market, the model with such copulæ may fit the data rather well. In the dynamics of the copula parameter, only γ_1 is significant and negative according to Fisher-information-based standard errors, showing some sort of mean reversion in the covariance between shocks, which indicates that a higher correlation today implies a lower correlation tomorrow on average. According to QML standard errors, both the ARCH effect and the asymmetric term are significant at the 5% level, which justifies the inclusion of the dynamic part in the copula specification. However, in the correlation plot (Figure 2.7.2), the average level of correlation is approximately 0.7, which is rather strong, and there is pronounced asymmetry. Due to the correlation being positive for this type of copula, if 'good' shocks go arbitrarily high, then 'bad' shocks also become greater. Therefore, U.S. market returns have a propensity for bull trends if we assume that good news dominates bad news on average. After strong negative shocks, the correlation nearly drops to zero, leading to no linear dependence between shocks. Thus, during crises, returns have a lower possibility of a bear trend. Regardless of the type of standard errors used (simple or QML), the coefficients seem to have concordant significance in most cases, which indicates that potential misspecification does not have a large impact on the interpretation of uncertainty about the model.

Now, we analyse how previous shocks affect the volatility and conditional asymmetry of the return distribution. Because it is cumbersome to get closed-form solutions for the second and third conditional moments, we use numerical integration to obtain them. In the opposite-sign-shock model, both types of shocks can affect future moments; therefore, we can plot a news impact surface. Analysing the volatility surface (Figure 2.7.3, left panel),[40] we see that if both 'good' and 'bad' shocks are in the negative zone, then the volatility rises dramatically, which matches the previous results on the volatility behaviour. However, if both shocks have positive values, then the conditional volatility is rather low.[41] Shifts between extreme empirical quantiles of unexpected part of returns or total shocks, e. g. from 50% to 1%, tend to lead to much more dramatic changes in volatility levels compared to the case when we move from 25% to 75%; thus,

[40] We choose the conditional return distribution on the 19th of January, 2010, as an example.

[41] The total volatility in this model includes the covariance between ε_t^+ and ε_t^-.

Figure 2.7.2: Dynamics of the conditional correlation in the best specification



Specification: centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$. The top grey line shows the dynamics of the S&P 500 ETF price.

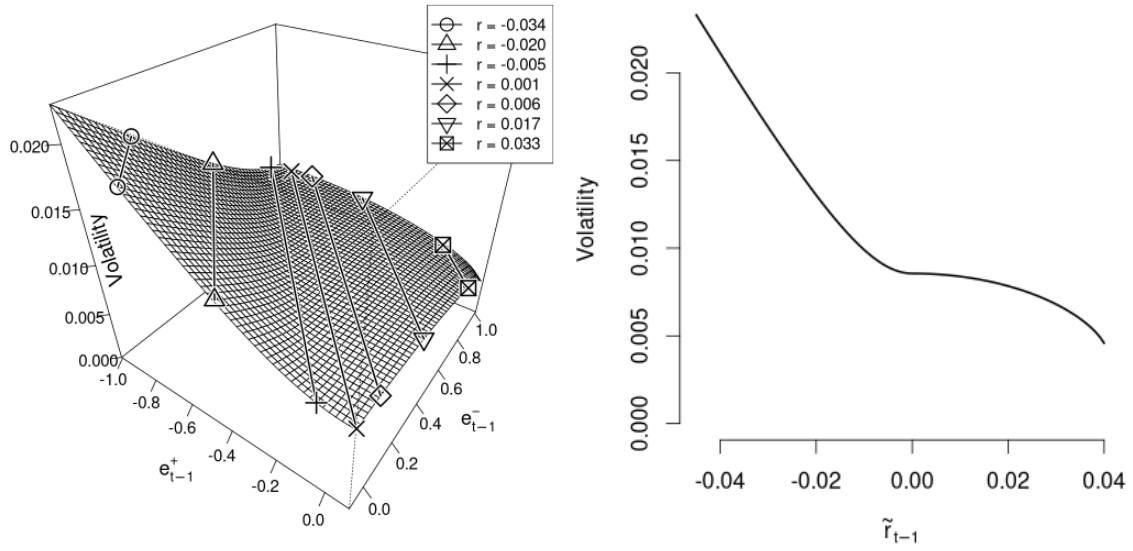
extreme events have significant and prolonged effects on future volatility dynamics. If one plots the dependence of volatility on the total shock (unexpected part of returns), which is the standard news impact curve (Figure 2.7.3, right panel), we clearly see that if the total shock has a negative sign, then volatility will monotonically increase, while if the overall shock is positive, the volatility does not react or can even decrease; therefore, good news can even decrease uncertainty in the market. Thus, volatility behaviour is extremely complex and is typically not captured by many standard models.

In Figure 2.7.4 in the left panel for the impact surface for conditional skewness, we see that the skewness is almost always negative for the chosen parameters, meaning that strong negative shocks are more common than positive shocks. If 'good' and 'bad' shocks have negative values, then the skewness is the largest in absolute value. If both shocks have unusually high positive values, then the skewness is the closest to zero. We see that dependence between empirical quantiles and skewness is far from linear and has a pronounced asymmetry. The right panel of Figure 2.7.4 shows that the negative overall shock yields stronger negative future conditional skewness; however, the positive overall shock clearly causes an increase in conditional skewness. Asymmetry in the skewness news impact curve follows the logic that a strong prior negative shock increases the probability of rare negative events, and vice versa, which agrees with the general logic behind the proposed model that 'bad' shocks occurring due to increased 'bad' volatility increase the weight of the left tail of the distribution, bearing greater skewness. Therefore, the asymmetrical behaviour of 'good' and 'bad' shocks and their connectedness explain such behaviour of skewness.

The volatility news impacts curve is similar to Bekaert et al. (2015). The skewness news impact curve is similar to Bekaert et al. (2015) for positive values of shocks and has distinct behaviour for negative news because in the proposed case, negative shocks lead to lower values of skewness. The proposed skewness news impact curve is similar to Anatolyev and Petukhov (2016); however, in Anatolyev and Petukhov (2016), this graph is convex on the negative part and concave on the positive part, while for the proposed model, the graph is piecewise convex for the positive and the negative parts.

Figure 2.7.5 shows the dynamics of volatility and conditional skewness. The left panel shows the typical process of volatility that increases during crisis periods (the

Figure 2.7.3: News impact surface and curve for volatility

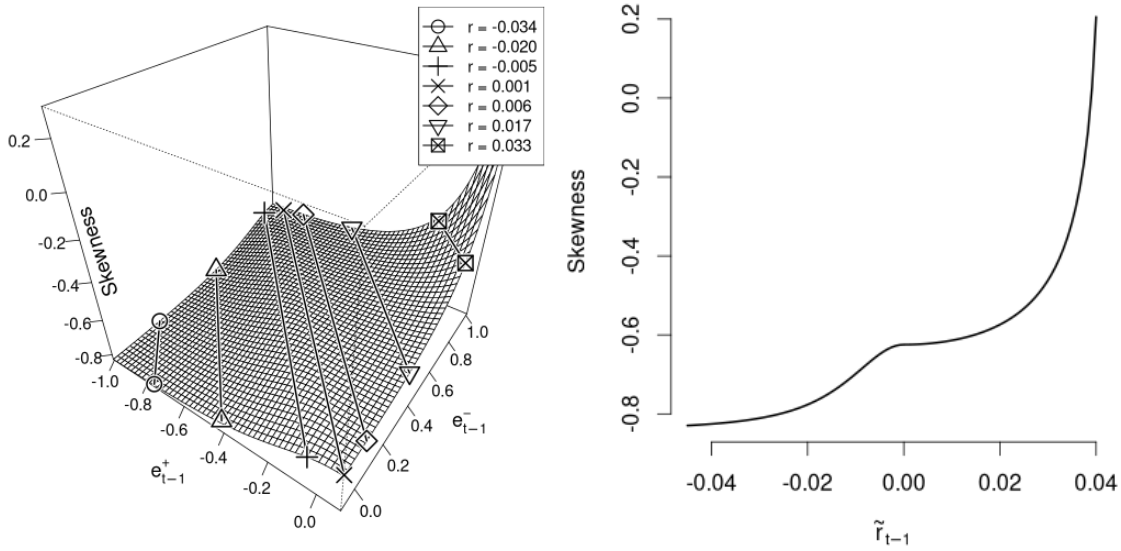


The left panel shows the volatility news impact surface for the best specification (centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$). The right panel represents the news impact curve for volatility. Both plots rely on the parameter values for the date 2010-01-19. The lines in the 3D plot correspond to the empirical (0.01, 0.05, 0.25, 0.50, 0.75, 0.95, 0.99)-quantiles of the returns. The ranges for e_t^- and e_t^+ are $[-1.01, 0.09]$ and $[-0.09, 1.01]$, respectively, and they produce values of total shocks $\tilde{r}_t = \sigma_t^+ e_t^+ + \sigma_t^- e_t^- \in [-0.044, 0.041]$, i. e. the theoretical returns correspond to the middle 99% of the values observed in S&P 500 data.

dot-com bubble, the Great Recession, and the European debt crises) and decreases during normal times, highlighting the well-known asymmetric dependence between price and volatility. The right panel displays more interesting patterns in conditional skewness. During the same crisis periods, conditional skewness becomes positive, which is rather unintuitive at first glance because positive skewness indicates a higher probability of extreme positive events, which is not expected during turmoil. However, considered from a different perspective, skewness becomes positive not immediately after the beginning of crises, but closer to its middle, and in such situations, investors can view stocks as 'cheap beats' (Kumar, 2009): in the middle of a crisis, the market has lower prices, higher volatility, and yields negative average returns. Therefore, if investors expect the market to bounce, which should be due to the mean-reversion pattern in prices and economic recovery, then positive skewness may arise. Additionally, investors can also put more money in stocks with lottery-like behaviour wishing to recoup; thus, the prices of such assets affect the index to a greater extent. Conditional skewness becomes negative when prices approximately return to their pre-crisis levels; therefore, during times of stability, traders expect rare negative returns, although not as strong, compared to the magnitude of the positive returns during crisis periods because volatility is much lower. Such counter-cyclical, sign-switching behaviour of the conditional skewness shows investors' naïve expectations: during good times, they are expecting the end of growth ('too-good-to-be-true' situation), and during bad times, they are expecting recovery ('too-bad-to-be-true' situation).

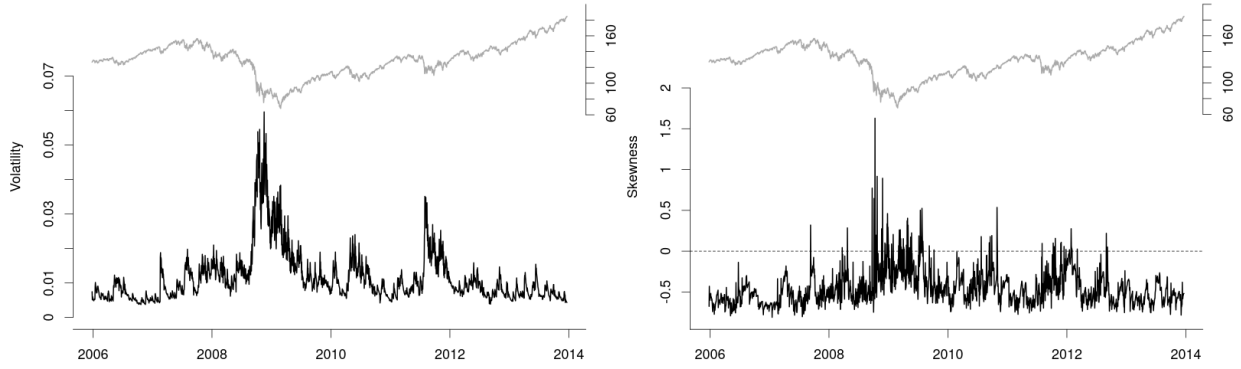
Since we forecast the conditional density of returns for the next day, it is possible to compute certain probabilities for decision-making. For example, the probability of positive and negative returns on the next day can be used to decide whether to buy or sell on the next day or not. However, due to the market efficiency, it is almost impossible to

Figure 2.7.4: News impact surface and curve for skewness



The left panel shows the skewness news impact surface for the best specification (centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$). The right panel represents the news impact curve for skewness. Both plots rely on the parameter values for the date 2010-01-19. The lines in the 3D plot correspond to the empirical (0.01, 0.05, 0.25, 0.50, 0.75, 0.95, 0.99)-quantiles of the returns. The ranges for e_t^- and e_t^+ are $[-1.01, 0.09]$ and $[-0.09, 1.01]$, respectively, and they produce values of total shocks $\tilde{r}_t = \sigma_t^+ e_t^+ + \sigma_t^- e_t^- \in [-0.044, 0.041]$, i. e. the theoretical returns correspond to the middle 99% of the values observed in S&P 500 data.

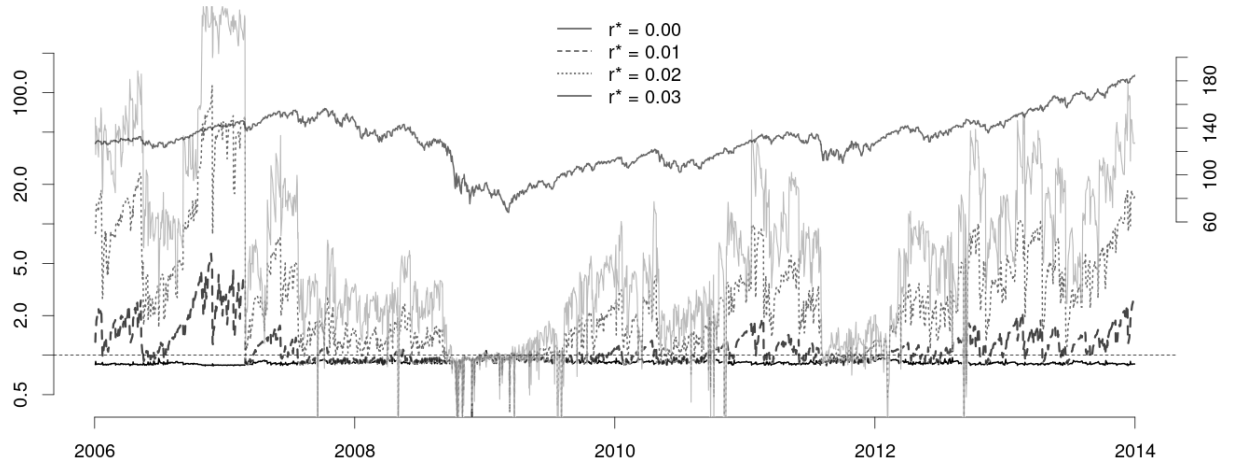
Figure 2.7.5: Dynamics of the volatility and skewness



These two panels show the dynamics of volatility (on the left) and conditional skewness (on the right) for the best specification (centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$). The grey line shows the dynamics of the S&P 500 ETF price.

reliably predict market movements: the middle 95% range for the predicted probability of negative returns (2.5- and 97.5-percentiles of the dynamic series) is 0.453–0.492, which implies that the markets are expected to slowly grow every day, and agrees with the fact that on 46% of trading days, the returns were negative. On the other hand, it is more useful to examine the ratio of the left and the right tail as the odds ratio of large negative and large positive returns. Since the left tail of the distribution has a slower decay rate ($\hat{\theta}^+ \approx 16.5$, $\hat{\theta}^- \approx 11.8$, and the baseline distribution is centred log-logistic), $\hat{\mathbb{P}}(r_t < -r^*)/\hat{\mathbb{P}}(r_t > r^*) \to \infty$ as $r^* \to \infty$. For reasonable values of r^*, however, this ratio shows how relatively more likely substantial losses are, compared to substantial gains. For every trading day, we predicted the ratio of left- and right-tail probabilities of the returns for cut-offs $r^* \in \{0, 0.01, 0.02, 0.03\}$, which is shown in Figure 2.7.6. The median ratios for the chosen tail cut-offs are 0.87, 1.02, 2.17, and 4.90, and the middle 95% ranges of these predicted ratios are $[0.83, 0.97]$, $[0.86, 3.3]$, $[0.91, 41]$ and $[0.90, 202]$, respectively. On average, per 1 day with returns greater than 0.02, 2 days with returns less than -0.02 are expected, and per 1 day with returns greater than 0.03, 5 days with returns less than -0.03 are expected. During good times, the ratio is greater than 1, and, at the bottom of crises, it is near one or even smaller. This matches with the 'too-good/bad-to-be-true' logic discussed above. During good times, investors expect that a trend can change and a crisis can begin; therefore, the probability of extreme negative returns is higher than the probability of large positive ones. At the deepest points of crises, investors expect that the markets should bounce back. Therefore, the practical conclusion following from this model is slightly larger probabilities of positive returns every day with high left-to-right-tail ratios, and advisability of becoming a long-term investor in this asset.

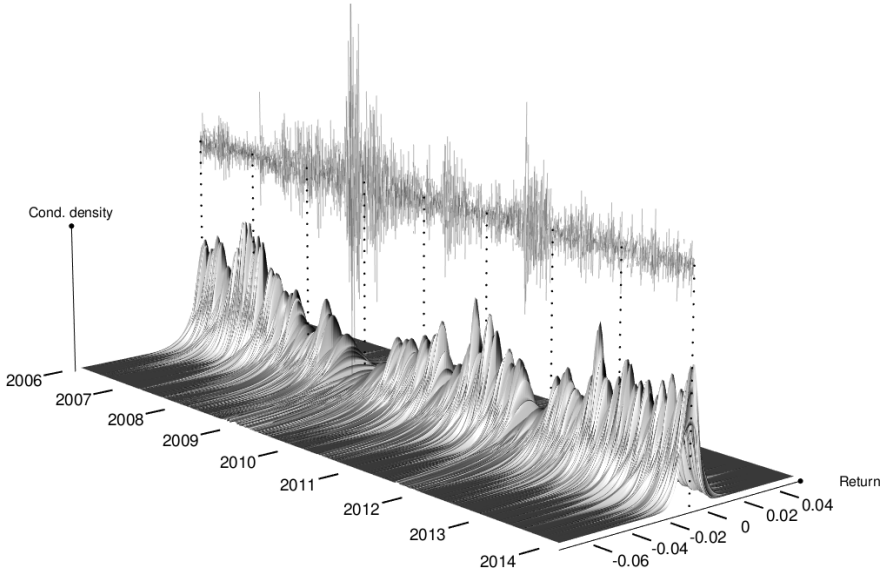
Figure 2.7.6: Dynamics of the left-to-right-tail ratio in the best specification



Specification: centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$. The top grey line shows the dynamics of the S&P 500 ETF price.

Because the conditional density itself provides many useful insights, we show its evolution over time in Figure 2.7.7. The conditional density shown in the plot has rich dynamics: during crises, the distribution becomes flatter; however, during good times, it is concentrated around zero more tightly. Our model also generates tail behaviour with different rates of decay; however, asymmetry in the centre of distribution is not very pronounced; therefore, asymmetry of the distribution is mostly caused by the tail difference.

Figure 2.7.7: Evolution of conditional return distribution in time



Specification: centred log-logistic distribution with Clayton copula with dynamic parameter containing $\tilde{r}_t^3$. The surface conditional return density in different time periods. The line above the surface represents daily returns.

2.8 Results for models with jumps

As it was mentioned, jumps can be important for the return process. For models with jumps, we take the best-performing opposite-sign-shock specification without jumps for S&P 500 data (centred log-logistic shocks connected by a Clayton copula with $\tilde{r}^3$ in dynamics) and introduce jumps with different forms. First, we use S&P 500 ETF data to estimate specifications with static unified jumps and a static outer copula.[42] Then, we consider specifications with dynamic unified jumps (Equation (2.3.3)) with static outer copula. Finally, we analyse static and dynamic opposite-sign jumps with static jumps and outer copula. The results for these models are shown in Table 2.8.1.

We begin with the model with static unified jumps. The expected value of jump size μ_ξ is negative and large in absolute value (Table 2.8.1, upper section), which indicates that primarily discrete downward changes are anticipated. Moreover, approximately one jump is occurring per day on average. The same situation occurs with jumps with dynamic unified jumps (Table 2.8.1, bottom section): the expected jump size μ_ξ is negative, and approximately less than 1 jump per day occurs. For these specifications, the expected number of jumps is calculated as the expectation of an AR(1) process (Maheu & McCurdy, 2004). Based on the in-sample AICs for the last estimation period (from 2005-09-28 to 2019-05-31, which contained more periods of turmoil with potentially more jumps than the first estimation sample, 2000-01-03 to 2014-12-31), the specification without jumps has the lowest values. Therefore, if continuous shocks have a fat-tailed distribution, then jumps are not so critical because many return changes that are thought to be jumps, in reality, turn out to be bursts in continuous volatility, which coincides with the results from Bajgrowicz et al. (2015)

[42] Introducing dynamics into the outer copula parameter leads to severe over-parametrisation. Certain models with a static outer copula connecting the jumps and the shock copula exhibited convergence problems, which restricts the degree of flexibility a researcher has in modelling the dynamic correlation between shocks and jumps.

and Christensen et al. (2014). Additionally, we highlight the differences between the estimates from the models with and without jumps. If we analyse a model with a static distribution of the number of jumps, the parameters of 'good' volatility are affected by the introduction of jumps, typically leading to an increase in the intercept, ARCH coefficient (in absolute values), and leverage coefficient. 'Bad' volatility is, in fact, not so much affected by the introduction of jumps, as well as the shock copula. The most noticeable change is that of the constant mean parameter, which is likely caused by non-zero-mean jumps. The copula parameters for the jump copula and outer copula lie in the range corresponding to medium correlation. The models with no copula between discrete jumps and continuous shocks and with cubic copula give on average results different from all other models.

For the model with opposite-sign jumps (Table 2.8.2), we use two distributions of jump sizes: exponential and Rayleigh. Both for dynamic and static cases with both distributions, results are concordant with one another: the intensity of 'bad' jumps $\mathbb{E}n^-$ is higher than the intensity of positive jumps $\mathbb{E}n^-$. In most models, the average size of 'bad' jumps $\mathbb{E}\xi^-$ is marginally greater in magnitude than the average size of positive jumps $\mathbb{E}\xi^-$. Therefore, 'bad' jumps occur more often and are stronger than positive jumps. Specifications with Rayleigh jump size have marginally lower AICs than those with exponential jumps except for the FGM copula, where both distributions yield nearly identical values. Dynamic opposite-sign jump models, which are the most detailed specification, have marginally lower AICs than models with unified jumps; therefore, there is marginal evidence that real-world jumps might have 'good' and 'bad' parts, and for static jumps, the effect is opposite. Adding dynamics of jump intensity into the specification did not yield any substantial improvement.

In several models with dynamic unified and opposite-sign jumps, the outer copula parameter estimates were so large in absolute value (for copulæ with $\kappa \in \mathbb{R}$) or close to the boundary that it caused degeneracy of the joint distribution, indicating that the most general specification suffers from over-parametrisation and more parsimonious models are preferred. Because the tails of exponential and Rayleigh distributions decay more slowly than those of Gaussian distribution, the poor in-sample performance of models with jumps cannot be explained by the normality assumption for unified jump sizes. Other parameters of the models show, in fact, a similar change as in the case with unified jumps (the AMH copula for the static case also shows some substantial changes). Therefore, the inclusion of jumps, unified or opposite-signed, static or dynamic, does not improve even the in-sample performance, and it is impractical to analyse the out-of-sample predictive power of the models with jumps.

Our model explains the results of Park (2016), where it is shown that upward jumps in the VIX index price are more important than downward jumps. As the VIX index is, roughly speaking, the expected level of volatility for the S&P 500 index, the fact that upward changes in volatility are more important can be explained by the structure of the proposed model, where jumps in returns can cause only upward discrete changes in volatility (it is only the correlations between elements that can be negative, which leads to lower overall volatility). Because the negative jumps are stronger and more frequent, they contribute more to discrete changes in overall volatility.

Table 2.8.1: Estimates of specifications with unified jumps

Static	α_0	α_1	α_2	α_2^-	β_0	β_1	β_2	β_2^-	γ_0	γ_1	γ_2	γ_2^-	θ^+	θ^-	μ	κ_{jump}	μ_ξ	σ_ξ	$\mathbb{E}n$	AIC
Magnitude	10^{-4}	1	10^{-1}	1	10^{-4}	1	10^{-1}	1	1	1	10^6	10^6	1	1	10^{-3}	1	10^{-1}	10^{-1}	1	
No jumps	−0.01	0.98	−2.05	1.26	0.37	0.81	−7.25	7.48	3.34	−0.31	−1.34	3.20	15.84	11.75	0.09					−264.08
Indep.	−0.13	0.94	−11.33	13.02	0.99	0.93	−14.65	11.45	14.71	0.32	−1.91	1.81	20.38	14.30	−0.02		−0.94	0.18	1.09	−193.90
AMH	0.38	0.92	−5.53	3.34	0.72	0.96	−5.64	5.63	14.57	0.16	−1.72	1.55	15.46	15.07	2.16	0.52	−0.13	0.19	0.87	−248.81
Clayt.	−0.22	0.97	−1.86	1.74	0.60	0.85	−7.02	7.32	3.97	−0.15	−2.71	2.69	17.41	15.59	1.88	0.62	−0.22	0.06	0.61	−257.18
Cubic	0.11	0.90	−12.79	11.33	0.54	0.96	−2.41	2.89	−0.07	0.03	6.56	12.23	16.95	16.38	2.40	0.38	−0.45	0.12	1.16	−235.45
FGM	−0.15	0.96	−5.74	4.15	0.74	0.85	−7.56	6.98	0.63	0.19	−2.51	2.97	16.48	15.27	2.04	0.65	0.23	0.32	1.20	−258.30

Dynamic	α_0	α_1	α_2	α_2^-	β_0	β_1	β_2	β_2^-	γ_0	γ_1	γ_2	γ_2^-	θ^+	θ^-	μ	κ_{jump}	μ_ξ	σ_ξ	$\mathbb{E}n_t$	AIC
Magnitude	10^{-4}	1	10^{-1}	1	10^{-4}	1	10^{-1}	1	1	1	10^6	10^6	1	1	10^{-3}	1	10^{-1}	10^{-1}	1	
Indep.	−0.04	0.88	−12.82	12.24	0.72	0.94	−12.65	11.35	14.17	0.02	−1.19	1.48	21.46	14.23	−0.35		−0.70	0.29	3.22	−158.96
AMH	−0.30	0.99	−1.50	0.31	1.33	0.85	−15.06	11.38	8.18	0.09	−0.51	0.65	16.39	15.53	2.12	0.36	−1.03	0.82	0.26	−231.90
Clayt.	0.02	0.98	−2.52	0.71	0.20	0.82	−7.81	8.61	3.50	−0.26	−2.91	3.32	16.01	16.32	1.91	0.45	−0.06	0.14	0.27	−250.43
FGM	−0.34	0.97	−5.01	2.36	0.48	0.85	−7.33	5.19	0.35	0.17	−1.79	2.31	16.95	14.62	1.99	0.62	0.63	0.04	0.91	−241.55

In these tables, we compare the coefficients for the best-performing specification without jumps to its static and dynamic unified jump extensions. The copula connecting the shock copula and the jump distribution (with parameter κ_{jump}) is shown on the left. μ_ξ and σ_ξ are the mean and the standard deviation of the jump size (assumed to be normally distributed), respectively; λ governs jump intensity in static jump specifications; and $\lambda_0, \lambda_1, \lambda_2$ govern the jump intensity in dynamic jump models. The coefficients are given for the period from between 2005-09-28 and 2019-05-31 (re-estimation after every 20 points). Because these estimates span multiple orders of magnitude for various parameters, the reported values must be multiplied by the magnitude. The dynamic specification with cubic jump copula exhibited convergence problems and is omitted from the second table. Standard errors are not reported because, as we showed in Section 2.7.3, they are a poor approximation of parameter uncertainty in these models. To save space, we report AIC values plus 24 000.

2.9 Conclusion

In this paper, we propose a new conditional density model with copula-connected 'good' and 'bad' shocks. The proposed model considers potentially different tail behaviours and asymmetries of stock return distributions. Due to its generality, the model unveils additional inner workings of an asset pricing process. Out-of-sample and in-sample comparisons on the S&P 500 data show that a sub-set of specifications of the proposed model outperforms the entire pool of 40 well-established GARCH models analysed. The correlation between 'good' and 'bad' shocks appears to be important for model performance, is time-variant, positive, and has a leverage-like effect. 'Bad' and 'good' volatilities have rather distinct dynamics, and the 'bad' volatility is greater than the 'good' volatility. The reaction of the total volatility to shocks is extremely asymmetric: negative total shocks can only increase volatility; however, positive shocks can even marginally decrease it. Positive total shocks increase the conditional skewness, although negative shocks have the opposite effect on it. Conditional skewness has positive/negative signs during bad/good regimes and, coupled with other stylised facts about prices and volatility, shows that investors have naïve expectations: during crises, they wish for recovery, and during good times, they invariably expect losses. During good times, extreme negative returns are more probable than extreme positive returns. During crisis peaks, strong positive and negative returns are approximately equally likely. Throughout the entire sample period, the probability of positive daily returns is marginally greater than 0.5, without substantial fluctuations.

These findings indicate that the behaviour of returns is asymmetric and more complex than previously surmised. In addition, we found that the same family of specifications provides the best in-sample and out-of-sample fits on S&P 500 and IBM stock data.

Table 2.8.2: Estimates of models with opposite-sign jumps

Static	α_0	α_1	α_2	α_2^-	β_0	β_1	β_2	β_2^-	γ_0	γ_1	γ_2	γ_2^-	θ^+	θ^-	μ	κ_{jump}	κ_{outer}	$\mathbb{E}\xi^+$	$\mathbb{E}\xi^-$	$\mathbb{E}n^+$	$\mathbb{E}n^-$	AIC
Magnitude	10^{-4}	1	10^{-1}	1	10^{-4}	1	10^{-1}	1	1	1	10^6	10^6	1	1	10^{-3}	1	1	10^{-1}	10^{-1}	1	1	
No jumps	−0.01	0.98	−2.05	1.26	0.37	0.81	−7.25	7.48	3.34	−0.31	−1.34	3.20	15.84	11.75	0.09							−264.08
Exponential jump distribution																						
Indep.	0.24	0.87	−13.14	12.86	1.21	0.89	−13.37	11.29	14.79	0.00	−1.39	0.91	30.41	15.78	−0.01			0.44	−0.57	0.17	0.64	−192.85
AMH	0.15	0.98	−0.97	0.33	0.96	0.85	−14.91	11.10	8.33	0.12	0.42	1.97	16.08	14.49	2.21	0.43	0.35	0.30	−0.44	0.20	0.39	−235.85
Clayt.	−0.37	1.00	−1.94	2.62	0.45	0.80	−7.13	6.50	2.66	−0.23	−1.40	3.12	18.00	14.81	2.02	0.48	1.66	0.42	−0.41	0.25	2.40	−252.75
Cubic	0.31	0.86	−13.42	10.93	0.01	0.96	−2.15	2.15	−0.66	0.18	5.30	43.62	17.16	16.82	1.91	0.63	0.53	0.36	−0.50	0.37	0.57	−226.63
FGM	0.00	0.97	−4.14	3.49	0.38	0.83	−6.95	4.90	0.90	0.21	−2.40	3.42	16.90	15.48	2.07	0.69	0.75	0.32	−0.56	0.21	0.62	−247.40
Rayleigh jump distribution																						
Indep.	−0.08	0.84	−13.19	11.75	0.86	0.90	−13.19	11.20	13.03	0.18	−2.14	1.08	21.89	14.13	0.35			0.08	−0.01	0.35	0.62	−196.26
AMH	−0.21	0.92	−0.84	0.32	1.10	0.85	−16.03	11.10	8.74	0.02	−1.26	2.08	16.56	15.16	2.06	0.47	0.77	0.02	−0.16	0.51	0.76	−235.92
Clayt.	−0.06	0.95	−1.92	1.70	0.10	0.79	−6.85	5.83	3.53	−0.34	−0.50	3.24	15.94	13.69	2.08	0.66	2.69	0.20	−0.49	1.37	2.20	−254.03
Cubic	0.45	0.87	−12.57	11.19	0.01	0.95	−4.19	3.40	0.31	0.11	6.96	43.47	15.92	14.92	2.15	0.72	0.58	0.19	−0.40	0.11	0.42	−232.23
FGM	−0.24	0.95	−5.73	3.97	0.53	0.87	−6.74	4.91	0.74	0.21	−3.12	1.17	15.47	14.96	2.15	0.50	0.31	0.22	0.07	0.34	0.54	−247.37

Dynamic	α_0	α_1	α_2	α_2^-	β_0	β_1	β_2	β_2^-	γ_0	γ_1	γ_2	γ_2^-	θ^+	θ^-	μ	κ_{jump}	κ_{outer}	$\mathbb{E}\xi^+$	$\mathbb{E}\xi^-$	$\mathbb{E}n^+$	$\mathbb{E}n^-$	AIC
Magnitude	10^{-4}	1	10^{-1}	1	10^{-4}	1	10^{-1}	1	1	1	10^6	10^6	1	1	10^{-3}	1	10^{-1}	10^{-1}	1	1	1	
Exponential jump distribution																						
Indep.	0.25	0.88	−12.81	11.69	1.19	0.90	−13.18	11.65	14.18	0.01	−1.66	0.79	31.45	16.56	−0.41	0.03	0.02	0.48	−0.65	0.96	1.94	−189.22
Clayt.	−0.19	0.99	−1.57	1.89	0.43	0.81	−6.96	6.43	4.28	−0.25	−0.14	4.03	18.00	15.38	1.93	0.51	1.24	0.60	−0.64	0.28	0.54	−249.94
FGM	0.08	0.97	−4.52	5.01	0.46	0.84	−6.53	4.95	1.44	0.22	−2.80	2.52	16.05	15.47	1.57	0.72	0.75	0.40	−0.65	1.39	0.70	−244.09
Rayleigh jump distribution																						
Indep.	−0.15	0.86	−14.07	12.67	0.89	0.91	−12.90	12.29	14.37	0.18	−1.69	1.46	23.03	14.48	−0.50	0.02	−0.10	0.22	−0.14	1.20	1.84	−192.38
FGM	−0.13	0.95	−4.63	4.15	0.57	0.84	−6.85	5.65	1.82	0.25	−2.63	2.27	14.80	14.93	1.94	0.61	0.05	0.39	0.00	1.18	1.63	−245.10

In these tables, we compare the coefficients for the best-performing specification without jumps to its static and dynamic opposite-sign jump extensions by assuming different distributions of signed jumps: exponential (top) and Rayleigh (bottom). λ^+ and λ^- govern jump intensity in static jump specifications; $\lambda_0^+, \lambda_1^+, \lambda_2^+, \lambda_0^-, \lambda_1^-$ and λ_2^- govern jump intensity in dynamic jump models. The coefficients are given for the period from between 2005-09-28 and 2019-05-31 (re-estimation after every 20 points). Because these estimates span multiple orders of magnitude for various parameters, the reported values must be multiplied by the magnitude. Several specifications with copulæ exhibited convergence problems and are omitted from the second table. Standard errors are not reported because, as we showed in Section 2.7.3, they are a poor approximation of parameter uncertainty in these models. To save space, we report AIC values plus 24 000.

Specification of the proposed model with dynamic scales performs better than their counterparts with dynamic shapes (BEGE models). Despite the solid in-sample goodness of fit and small Akaike criterion values, and their previously good performance on monthly data in the existing literature, the latter did not perform well on daily return data. Adding copulæ also did not improve those results much, which is why we believe that models with dynamic scale and copulæ should be used.

The results for models with jumps indicate that unified jumps have a negative mean and intensity below one per day. Models with opposite-sign jumps show that 'bad' jumps occur more often and have a greater mean in absolute value than 'good' jumps. However, the introduction of jumps, both unified and opposite-sign, does not improve even in-sample performance; therefore, jumps are not very important at daily frequencies, and continuous shocks that come from a tail of their distribution tail can explain the observed phenomena better. The proposed framework also provides a natural explanation for the observed phenomenon of upward jumps in the VIX index.

Acknowledgements

We would like to thank Alexei Boulatov, Daniil Yesaulov, Antonio Cosma, Gautam Tripathi, Sergey Gelman, and Benjamin Holcblat for their suggestions and comments that greatly improved this paper. We would also like to thank Thierry Magnac; Shin Kanaya; Dennis Kristensen; Diego Ronchetti; Rustam Ibragimov; all participants and discussants of the 13th International Conference on Computational and Financial Econometrics, the 5th Econometric Research in Finance Workshop, the 6th International Symposium in Computational Economics and Finance; and all participants of research seminars at the University of Luxembourg, the HSE University, the Central Economic Mathematical Institute, and the Centre for Econometrics and Business Analytics for many valuable ideas and insightful questions. Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged. Andreï V. Kostyrka gratefully acknowledges financial support from FNR-Luxembourg through a PRIDE grant for the Migration and Labour (MINLAB) doctoral training unit. The estimation and simulation results reported in this paper were obtained thanks to the HPC facilities of the University of Luxembourg (Varrette et al., 2014).

Appendix

2.A Replication and improvement of Bekaert et al. (2015)

The numerical accuracy of the original Bekaert et al. (2015) BEGE model can be improved using original densities, as in equation (2.2.4), and applying integration with stabilising techniques, thus increasing the required precision at the cost of a longer computation time. Since the BEGE model is a particular case of the proposed model, its parameters can be converted into equivalent parameters used in this paper. We ensure that the series-generating function returns similar dynamic shape series for the values from the article (Figure 2.A.1, dotted lines).[43] However, the differences shown in Table 2.A.1 are too large to attribute them to rounding or small data discrepancies.

We begin by observing that if the 'good' dynamic shape in Bekaert et al. (2015) follows the expression

$$\theta_t^+ = p_0 + \rho_p \theta_{t-1}^+ + \frac{\phi_p^+}{2\sigma_p^2}\tilde{r}_{t-1}^2 \mathbb{I}_{t-1}^+ + \frac{\phi_p^-}{2\sigma_p^2}\tilde{r}_{t-1}^2 \mathbb{I}_{t-1}^-, \tag{2.A.1}$$

and in this article:

$$\theta_t^+ = \alpha_0 + \alpha_1 \theta_{t-1}^+ + \alpha_2 \tilde{r}_{t-1}^2 + \alpha_2^- \tilde{r}_{t-1}^2 \mathbb{I}_{t-1}^-.$$

We thus find equivalence between parameters:

$$\alpha_0 \equiv p_0, \quad \alpha_1 \equiv \rho_p, \quad \alpha_2 \equiv \frac{\phi_p^+}{2\sigma_p^2}, \quad \alpha_2^- \equiv \frac{\phi_p^- - \phi_p^+}{2\sigma_p^2}.$$

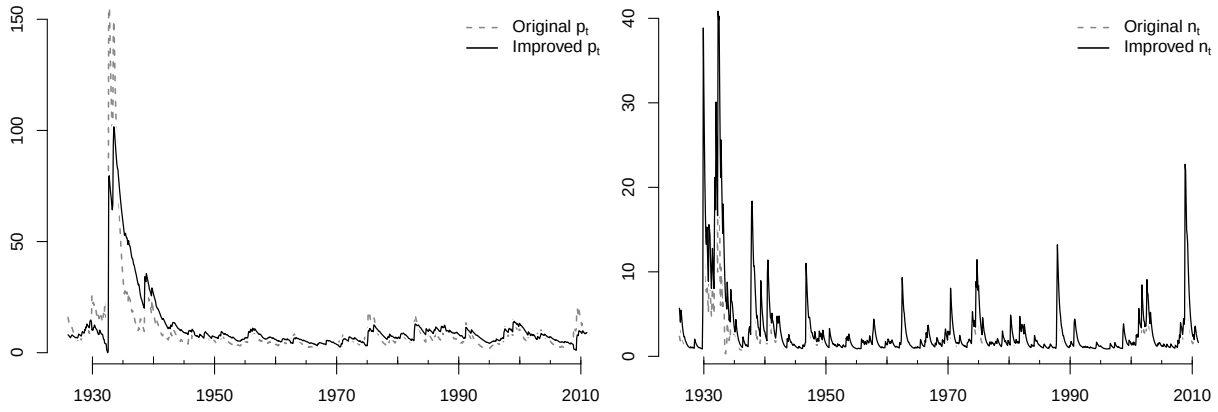
The same applies to the β parameters and θ_t^- for the 'bad' shape dynamics.

Table 2.A.1 shows the optima and corresponding likelihood values from Bekaert et al. (2015, Table 2, corresponding to their original monthly data). We convert their parameter values into ours using the equivalence formula above. The original log-likelihood value of 1724.26 in Bekaert et al. (2015) appears markedly different due to the rough numerical approximation of the density.[44] We obtain a much higher likelihood value when using their estimates and then apply a gradient-based improvement to the first stage of all models (BFGS with 10^{-8} relative tolerance as a stopping criterion).

[43] The series look similar, but not the same due to rounding errors in the parameter estimates or the fact that there might be slight discrepancies between the monthly return data we use (Shiller, 2015) and those which Bekaert et al. (2015) use (CRSP).

[44] Bekaert et al. (2015) use only 100 grid points for integration, whereas we rely on adaptive Gauss-Kronrod cubature with thousands of points where the function is evaluated until the relative difference is smaller than 10^{-10}.

Figure 2.A.1: Replicated and improved dynamic shape series (Bekaert et al., 2015, Figure 3, p. 266)



The values before and after refinement are similar for most parameter values; however, certain parameters, such as α_0, α_2, β_2, or σ_-^2, change by more than 40%, which explains the substantial likelihood gains. Finally, the null hypothesis that the 11 parameters of the model are equal to those reported in Bekaert et al. (2015) is rejected based on the LR test (the LR statistic is equal to 58.1, $p(\chi^2_{11} < 58.1) \approx 2 \cdot 10^{-8}$).

Table 2.A.1: Bekaert et al. (2015) result improvement

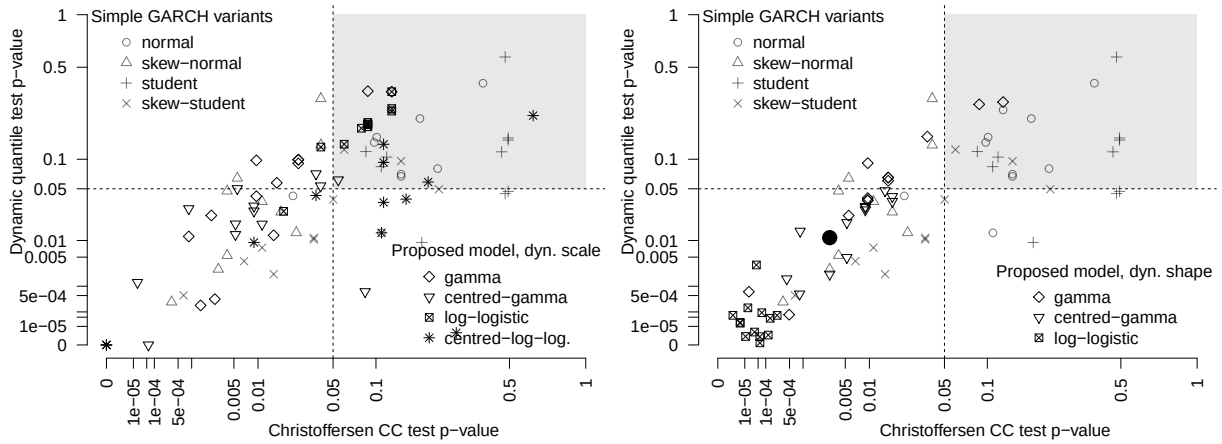
	Article (reported values)	Article (our evaluation)	BFGS optimised	Magnitude multiplier
α_0	0.0890	0.0890	0.0325	1
α_1	0.9099	0.9099	0.9665	1
α_2	0.9298	0.9298	0.4827	10^3
α_2^-	-0.8063	-0.8063	-0.5146	10^3
β_0	0.2204	0.2204	0.1614	1
β_1	0.7822	0.7822	0.8110	1
β_2	-0.0495	-0.0495	-0.0052	10^3
β_2^-	0.2721	0.2721	0.3513	10^3
σ_+^2	0.0518	0.0518	0.0435	10^{-3}
σ_-^2	0.7969	0.7969	0.4669	10^{-3}
μ	1.0000	1.0000	0.6252	10^{-2}
Log-lik.	1724.3	1888.7	1917.8	

The first column represents estimates and log-likelihood value from the original Bekaert et al. (2015) model. The second column represents the case when estimates from the original article converted into their exact equivalents from the proposed specification and inserted into the log-likelihood function. The third column describes the case when BFGS optimiser uses the estimates of Bekaert et al. (2015) as initial values. The last column shows the multipliers by which all estimates should be multiplied to have the original magnitude.

Then, we estimate the original model and its variants with copulæ with the S&P 500 ETF daily data and report the results in Table 2.A.2. All of these specifications fail the conditional coverage and dynamic quantile tests. For a more general picture, the p-values of these two tests are plotted against one another in Figure 2.A.2 with opposite-sign specifications with dynamic scales on the left and dynamic shapes on the right. The latter do not pass conditional coverage and dynamic quantile tests

regardless of the copula and its dynamics. Therefore, the dynamic-shape specification is not appropriate for VaR forecasting. However, dynamic-shape specifications perform well in the sample and often return low AICs. E. g. the 6 extensions of Bekaert et al. (2015) have the lowest AICs out of all specifications, with the lowest being -24275.2, compared to -24264.1 of the best dynamic-scale variant. The results with non-centred gamma and centred log-logistic distributions are nearly identical; however, we do not report them in this appendix to save space.

Figure 2.A.2: p-values for two VaR quality tests by distribution



Forty popular GARCH variants are compared to all opposite-sign model specifications with dynamic scales (left panel) and extensions of Bekaert et al. (2015) (right panel); all models are using de-meaned returns. The original BEGE model is shown with a black dot. The specifications that did not fail either of the exceedance tests are in the upper right quadrant. Point characters correspond to various distribution functions of baseline shocks.

2.B Particularities of numerically stable optimisation

2.B.1 Initial value selection for dynamic scale parameter series

The problem of the initial values, $\sigma^2_{\varepsilon_1^+}$ and $\sigma^2_{\varepsilon_1^-}$, required to compute the entire dynamic series $\{\sigma^2_{\varepsilon_t^+}, \sigma^2_{\varepsilon_t^-}\}_{t=2}^T$, might be solved in multiple ways. The easiest is taking plausible values, generating the series iteratively, and discarding the first point. However, we improve the initial value guess using $\sigma^2_{\varepsilon_1^+} = \sigma^2_{\varepsilon_1^-} = 0.5 \operatorname{Var} r_t$, generating the first part of the dynamic series ($\lfloor T/10 \rfloor$ points), and plugging the medians of those series as the initial values, making them dependent on the model parameters as well.

2.B.2 Copulæ with restrictions on parameter space

The dynamic series for the copula parameter might contain values outside the admissible range depending on the γ parameters and extreme values of observed returns, and values near the boundary of this range can cause numerical instability during integration (i. e. the density becomes nearly degenerate). The use of slowly decaying damping functions is therefore highly advocated. Table 2.B.1 contains the damping functions

Table 2.A.2: Bekaert et al. (2015) original model and its extensions

Copula	Dyn.	VR^{i}	p^{i}_{CC}	p^{i}_{DQ}	$p^{\text{i}}_{\text{dur}}$	AIC	VR^{o}	p^{o}_{CC}	p^{o}_{DQ}	$p^{\text{o}}_{\text{dur}}$
—	—	0.973	0.054	0.292	0.001	−24244.8	0.673	0.003	0.011	0.411
Plackett	static	1.000	0.000	0.266	0.007	−24226.7	0.691	0.005	0.019	0.441
Plackett	$\tilde{r}^2$	1.043	0.796	0.271	0.014	−24274.8	0.782	0.018	0.036	0.282
Plackett	$\tilde{r}^3$	1.043	0.796	0.408	0.007	−24275.2	0.782	0.018	0.041	0.368
cubic	static	0.883	0.012	0.175	0.001	−24244.1	0.709	0.009	0.031	0.551
cubic	$\tilde{r}^2$	0.915	0.016	0.166	0.002	−24218.2	0.618	0.000	0.002	0.358
cubic	$\tilde{r}^3$	0.984	0.050	0.099	0.000	−24274.5	0.727	0.015	0.048	0.596
AMH	static	0.995	0.031	0.122	0.002	−24222.9	0.691	0.005	0.005	0.356
AMH	$\tilde{r}^2$	0.995	0.200	0.740	0.001	−24230.3	0.709	0.009	0.028	0.551
AMH	$\tilde{r}^3$	0.968	0.309	0.781	0.008	−24228.4	0.709	0.009	0.030	0.551
Clayton	static	1.043	0.146	0.232	0.003	−24208.1	0.636	0.001	0.001	0.176
Clayton	$\tilde{r}^2$	1.000	0.000	0.748	0.001	−24236.0	0.673	0.003	0.002	0.263
Clayton	$\tilde{r}^3$	0.952	0.155	0.353	0.000	−24213.0	0.600	0.001	0.014	0.519

This table describes Bekaert et al. (2015)-like models (centred gamma distribution) with various copula specifications. All specifications in the table have centred gamma shock densities.
The superscript $^{\text{i}}$ denotes in-sample performance indicators, and $^{\text{o}}$ denotes out-of-sample indicators.
Dyn.: variable used in the dynamics of the copula parameter ($\tilde{r}^3$ for cubes of centred returns, $\tilde{r}^2$ for squares of centred returns).
VR: Violation ratio (the ratio of the number of realised exceedances to their expected number).
p_{CC}: p value of the conditional coverage test (Christoffersen, 1998).
p_{DQ}: p value of the dynamic quantile test (Engle & Manganelli, 2004).
p_{dur}: p value of the no-hit duration test (Christoffersen & Pelletier, 2004).

used in the paper; however, other functions can be used if they prevent extreme values of the copula parameter from being generated. Furthermore, to prevent degeneration of densities, the researcher may impose additional restrictions to force the parameters away from the boundary (e. g. defining the joint density to be zero if $\theta \notin (-0.99, 0.99)$).

Table 2.B.1: Damping functions for the dynamic copula parameter κ_t

Copula	Transformation
Plackett	$\begin{cases} \sqrt{\kappa_t + 0.25} + 0.5, & \kappa_t \geq 0 \\ \frac{2}{\pi} \arctan(\frac{\pi}{2}\kappa_t) + 1, & \kappa_t < 0 \end{cases}$
Cubic	$\frac{3}{\pi} \arctan(\kappa_t) + 0.5$
AMH	$\frac{2}{\pi} \arctan(\kappa_t)$
Clayton	$\begin{cases} \sqrt{\kappa_t + 0.25} + 0.5, & \kappa_t \geq 0 \\ \frac{2}{\pi} \arctan(\frac{\pi}{2}\kappa_t) + 1, & \kappa_t < 0 \end{cases}$

2.B.3 Numerically stable integration

The numerical integration used for conditional density evaluation in the proposed model must be used cautiously. Because the density $f_{\psi_t}(z)$ corresponds to de-meaned returns and the function being integrated from minus infinity to the upper limit is $f_{\varepsilon_t^+,\varepsilon_t^-}(z - v, v)$, it is natural to assume that most function mass lies in the region where 'bad' shocks take typical values similar to negative daily market returns (e. g. 99%

returns in the S&P 500 sample are higher than -0.034, and the lowest daily return is -0.104). Outside the natural range for ε_t^-, $f_{\varepsilon_t^+,\varepsilon_t^-}(z-v,v)$ takes small values even for fat-tailed distributions; thus, as a consequence, the default integration routines might converge prematurely because the estimated modulus of the integration error does not exceed the requested tolerance. The principal cause for such behaviour is a combination of two factors: (1) the function primarily concentrating in a very narrow range and (2) the standard 15-point Gauss-Kronrod rules for an unbounded interval not capturing the function mass where it lies, summing only near-zero values, and not going into the next adaptive step. The following example illustrates this phenomenon: if one attempts to numerically integrate the normal density with mean -0.2 and standard deviation 0.005 from minus infinity to 0, the most likely result will be a severe underestimation (e. g. the returned value could be approximately 10^{-5}, although it should be near 1).

Thus, the integrand should be transformed to be properly 'sampled' by the quadrature rules in the standard evaluation region. If a researcher requests integration from $-\infty$ to 0, 15 points of the quadrature are taken by default between $-10^{2.37}$ and $-10^{-2.37}$. The standard solution in such cases is scaling the argument of the integrand and the integrand itself by a measure of the function scale; no centring is required because the upper limit is 0. One such measure could be the conditional standard deviation of r_t; however, its computation requires evaluation of f_{r_t} itself, creating a vicious circle. A good feasible value for the scaling argument that we use in all integration routines is $\sqrt{\sigma^2_{\varepsilon_t^+} + \sigma^2_{\varepsilon_t^-}}$ (i. e. the true standard deviation if the shocks were independent). Another solution that prevents false convergence in the numerical integration routine is using a tighter relative tolerance as the stopping criterion. We found that requiring the relative error to be below 10^{-12} solves the false convergence issue without producing another (i. e. non-convergence due to the round-off error and impossibility of reaching the required tolerance).

2.C Vuong-like tests for likelihood-based criteria

This appendix describes a version of the likelihood ratio test developed by Vuong (1989). It is assumed that the pseudo-true value θ^* specifies the model in a family with conditional density f: $\theta^* \stackrel{\text{def}}{=} \arg\max_\theta \mathbb{E} \ln f(z_t;\theta)$. Similarly, let γ^* define the model with a different conditional density $g(z_t;\gamma^*)$. Note that both f and g may not be the density corresponding to the true DGP, and the models in the comparison may be nested, overlapping, or strictly non-nested. Under the null hypothesis that the models defined by densities f and g are equivalent (i. e. their distances from the true conditional density measured by the Kullback-Leibler information criterion, KLIC, are the same),

$$\hat{V}_{\text{KLIC}} \stackrel{\text{def}}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} [\ln f(z_t;\hat{\theta}) - \ln g(z_t;\hat{\gamma})] \xrightarrow[T\to\infty]{d} \mathcal{N}(0,\sigma_*^2),$$

where $\hat{\theta}$ and $\hat{\gamma}$ are the maximum likelihood estimators. Values of $\hat{V}_{\text{KLIC}}/\sqrt{\sigma_*^2}$ exceeding the chosen $(1-\alpha)$-quantile of the standard normal distribution imply that the two models are not observationally identical at the significance level α, and the one with conditional density f should be preferred. Calvet and Fisher (2004) suggest that σ_*^2 should be estimated with a HAC estimator due to potential serial correlation in the difference of the log-likelihood series.

This test will asymptotically choose the 'best' model (closest to the true distribution in terms of KLIC) with probability 1 if the conditions given in Calvet and Fisher (2004, Appendix A2) are met. Following the discussion in Vuong (1989, Section 5), we adjust the test statistic to penalise both models for the dimensions of the parameter vectors. For a model with density f, the Akaike information criterion is defined as

$$\mathrm{AIC} \stackrel{\mathrm{def}}{=} -2\sum_{t=1}^{T} f(z_t;\hat{\theta}) + 2\dim\hat{\theta} = -2\sum_{t=1}^{T}[f(z_t;\hat{\theta}) - \dim\hat{\theta}/T].$$

The adjusted test statistic for Akaike information criteria equivalence testing is defined as

$$\hat{V}_{\mathrm{AIC}} \stackrel{\mathrm{def}}{=} \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\big([\ln f(z_t;\hat{\theta}) - \dim\hat{\theta}/T] - [\ln g(z_t;\hat{\gamma}) - \dim\hat{\gamma}/T]\big) \xrightarrow[T\to\infty]{d} \mathcal{N}(0,\sigma_*^2),$$

The asymptotic variance of $\hat{V}_{\mathrm{AIC}}$ is the same as that of $\hat{V}_{\mathrm{KLIC}}$ because the addition of a fixed penalty acts as a shift by a constant. The penalty terms $\dim\hat{\theta}/T$ and $\dim\hat{\gamma}/T$ can be replaced with $0.5\ln T\cdot(\dim\hat{\theta})/T$ and $0.5\ln T\cdot(\dim\hat{\gamma})/T$ to compare the models by their Bayesian information criteria.

Chapter 3

Missing endogenous variables in conditional moment restriction models

This chapter is based on joint work with Antonio Cosma and Gautam Tripathi.

3.1 Introduction

Let Y_i^*, Z_i, X_i be random (column) vectors corresponding to individual i in a sample of size n, i. e. $i = 1, \ldots, n$. The vector Y_i^* consists of endogenous variables (outcomes or explanatory), all of which are simultaneously not observed for some individuals in the sample. In contrast, the vector Z_i consists of endogenous variables (outcomes or explanatory variables) observed for each individual in the sample. Similarly, X_i is a vector of exogenous variables observed for each individual in the sample. We refer to the coordinates of Y^* as being 'missing' (for some individuals). Analogously, the coordinates of (Z, X) are referred to as being 'non-missing' (for all individuals).

For each i, we also observe the dummy variable

$$D_i \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if all coordinates of } Y_i^* \text{ are observed,} \\ 0 & \text{if all coordinates of } Y_i^* \text{ are missing.} \end{cases}$$

If all coordinates of Y^* are observed for individual i, then we let

$$Y_i \stackrel{\text{def}}{=} D_i Y_i^* + (1 - D_i)\mathfrak{m} \tag{3.1.1}$$

denote the observed version of Y_i^*, where $\mathfrak{m}$ denotes missing values. The symbol $\mathfrak{m}$ can be thought of as a vector of pre-specified numbers, e. g. $\mathfrak{m} \stackrel{\text{def}}{=} (99999, \ldots, 99999)_{(\dim Y^*)\times 1}$, used to code missing values. This facilitates mathematical analysis because, e. g., then $0 \times \mathfrak{m} = 0_{(\dim Y^*)\times 1}$. Note that $D_i = \mathbb{1}(Y_i^* \neq \mathfrak{m})$, where $Y_i^* \neq \mathfrak{m}$ holds coordinate-wise to indicate that each coordinate of Y_i^* is observed. Similarly, $1 - D_i = \mathbb{1}(Y_i^* = \mathfrak{m})$, where $Y_i^* = \mathfrak{m}$ coordinate-wise to indicate that each coordinate of Y_i^* is not observed.

The econometric models we consider are characterised as a system of conditional moment equalities, namely,

$$\exists \theta^* \in \Theta \quad \text{s.t.} \quad \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] = 0 \qquad P_X\text{-a.s.}, \tag{3.1.2}$$

where g is a vector of functions known up to $\theta^* \in \Theta \subset \mathbb{R}^{\dim \theta^*}$.[45] The conditional distribution of $(Y^*, Z \mid X)$ and the marginal distribution of X, denoted by P_X, are both unknown. The objective is to use the data (D_i, Y_i, Z_i, X_i), $i = 1, \ldots, n$, to efficiently estimate θ^*.

The researcher chooses the vector-valued function g to model a system of relationships between the missing and the non-missing variables. The missing variables Y^* and the non-missing variables Z are classified to be endogenous, i. e. the variables inside g that are pairwise correlated with g, because they do not appear in the conditioning set in (3.1.2). The non-missing variables X are classified as exogenous, i. e. explanatory variables that are mean independent of g. In many applications, $X \stackrel{\text{def}}{=} (X_{\text{in}}, X_{\text{ex}})$, where X_{in} denotes the 'included' instrument variables (IV) and X_{ex} the 'excluded' IV. Included instruments refer to the exogenous variables appearing in g, whereas the excluded instruments are those exogenous variables that may not be in g due to exclusion restrictions imposed by economic theory but, based on external considerations, may appear in the conditioning set to ensure the identification of θ^*. Excluded instruments are not necessary if the included instruments suffice to identify θ^*, in which case X_{ex} is the

[45] If there is no conditioning, then (3.1.2) reduces to a system of unconditional moment equalities with some variables missing. E. g. these models are studied in X. Chen et al. (2008) and Graham (2011). For more on this, cf. Example 3.3.4.

empty vector. However, if there are no included instruments—e. g. when all explanatory variables are deemed to be potentially endogenous, in which case X_{in} is the empty vector—then excluded instruments are necessarily required to identify θ^*.

A large class of models in applied economics can be written as (3.1.2).

Example 3.1.1 (IV regression with missing outcome)**.** The canonical example of (3.1.2) is the linear IV regression model $Y^* = \alpha^* + X'_{\text{in}}\beta^* + Z'\gamma^* + U$, where only the outcome variable Y^* is missing for some observations, Z is the vector of endogenous regressors, and U is mean-independent of X, i. e. $\mathbb{E}[U \mid X] = 0$ P_X-a.s. In this example, $\theta^* \stackrel{\text{def}}{=} (\alpha^*, \beta^*, \gamma^*)$, $g(Y^*, Z, X, \theta^*) \stackrel{\text{def}}{=} Y^* - \alpha^* - X'_{\text{in}}\beta^* - Z'\gamma^*$, and Y^* is a scalar. If all the regressors are endogenous, then X_{in} is the empty vector, $X \stackrel{\text{def}}{=} X_{\text{ex}}$, and the definition of θ^* is adjusted accordingly by dropping β^*. The case where all endogenous variables in the linear regression model can be missing is handled by letting Z be the empty vector and $g(Y^*, X, \theta^*) \stackrel{\text{def}}{=} Y_1^* - \alpha^* - X'_{\text{in}}\beta^* - Y_2^{*\prime}\gamma^*$ with $Y^* \stackrel{\text{def}}{=} (Y_1^*, Y_2^*)$. Note that Y^* is now a vector. □

Example 3.1.2 (Linear regression with missing endogenous explanatory variables)**.** The linear regression model where the outcome variable is non-missing but the right-hand side endogenous variables may be missing can be handled by letting $g(Y^*, Z, X, \theta^*) \stackrel{\text{def}}{=} Z - \alpha^* - X'_{\text{in}}\beta^* - Y^{*\prime}\gamma^*$, where the outcome variable Z is now a scalar. □

Multivariate extensions of Examples 3.1.1 and 3.1.2 include systems of equations, linear or non-linear, that can be written as $g(Y^*, Z, X, \theta^*) = \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] = 0$ P_X-a.s.; cf., e. g., Newey (1993, Section 3), Powell (1994, Section 2.1), Pagan and Ullah (1999, Chapter 3), and Wooldridge (2010).

3.2 Identification

To identify, i. e. uniquely define, θ^* in the presence of missing observations without modelling how the missingness is created, we make a standard 'selection on observables' assumption that conditional on the non-missing variables (Z, X), the missing observations on Y^* are missing at random (MAR); i. e.

Assumption 3.2.1 (MAR)**.** *For all individuals, $D \perp\!\!\!\perp Y^* \mid Z, X$, where the symbol '$\perp\!\!\!\perp$' denotes stochastic independence.*

We can use MAR to evaluate $\mathbb{E}[g(Y^*, Z, X, \theta) \mid X]$, $\theta \in \Theta$, even when Y^* is missing. Indeed,

$$
\begin{aligned}
\mathbb{E}[g(Y^*, Z, X, \theta) \mid X] &\stackrel{P_X\text{-a.s.}}{=} \mathbb{E}\big[\mathbb{E}[g(Y^*, Z, X, \theta) \mid Z, X] \mid X\big] && \text{(tower property)} \\
&= \big[\mathbb{E}[g(Y^*, Z, X, \theta) \mid Z, X, D = 1] \mid X\big] && \text{(MAR)} \\
&= \mathbb{E}\big[\mathbb{E}[g(Y, Z, X, \theta) \mid Z, X, D = 1] \mid X\big] && (D = 1 \stackrel{(3.1.1)}{\Longleftrightarrow} Y^* = Y) \\
&= \mathbb{E}\left[\mathbb{E}\Big[\frac{Dg(Y, Z, X, \theta)}{\pi(Z, X)} \;\Big|\; Z, X\Big] \;\Big|\; X\right] \\
&= \mathbb{E}\Big[\frac{Dg(Y, Z, X, \theta)}{\pi(Z, X)} \;\Big|\; X\Big], && \text{(3.2.1)}
\end{aligned}
$$

where

$$
\pi(Z, X) \stackrel{\text{def}}{=} \mathbb{E}[D \mid Z, X] = \Pr(D = 1 \mid Z, X) \qquad \text{(3.2.2)}
$$

is the propensity score function. To emphasise the non-parametric nature of the propensity score function, we assume that

Assumption 3.2.2. *The functional form of* $(Z, X) \mapsto \pi(Z, X)$ *is fully unknown.*

The propensity score function, although unknown, is identified from the data as the conditional expectation of $D \mid Z, X$ because D, Z, X are non-missing. Since, under MAR,

$$\mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0 \overset{(3.2.1)}{\iff} \mathbb{E}\Big[\frac{Dg(Y, Z, X, \theta^*)}{\pi(Z, X)} \;\Big|\; X\Big] \overset{P_X\text{-a.s.}}{=} 0, \tag{3.2.3}$$

any condition that leads to the identification of θ^* in (3.1.2) will also ensure identification of θ^* in the moment condition on the right-hand side of (3.2.3), which does not contain any missing observations. To illustrate this, assume that the columns of

$$J \overset{\text{def}}{=} J(X, \theta^*)_{(\dim g)\times(\dim \theta^*)} \overset{\text{def}}{=} \partial_\theta \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] \overset{\text{def}}{=} \partial_\theta \mathbb{E}[g(Y^*, Z, X, \theta) \mid X]\big|_{\theta=\theta^*}$$

are linearly independent P_X-a.s. As shown in Appendix 3.B, this is sufficient to ensure that θ^* is locally identified.[46] Hence, as π does not depend on θ (Assumption 3.2.2),

$$J \overset{(3.2.1)}{=} \partial_\theta \mathbb{E}\Big[\frac{Dg(Y, Z, X, \theta^*)}{\pi(Z, X)} \;\Big|\; X\Big] \qquad P_X\text{-a.s.}$$

Therefore, the columns of J are linearly independent P_X-a.s. if and only if the columns of $\partial_\theta \mathbb{E}[Dg(Y, Z, X, \theta^*)/\pi(Z, X) \mid X]$ are linearly independent P_X-a.s. Since the identification of the parameters of interest in the missing data problem is not lost under MAR, for the remainder of the paper, we maintain that

Assumption 3.2.3. θ^* *is identified.*

3.3 Efficient estimation under MAR

Henceforth, we liberally use functional notation, suppressing the arguments taken by functions whenever there is no danger of confusion. In particular, we let $\pi \overset{\text{def}}{=} \pi(Z, X)$ and $g \overset{\text{def}}{=} g(Y^*, Z, X, \theta^*)$. Note that $Dg = Dg(Y^*, Z, X, \theta^*) \overset{(3.1.1)}{=} Dg(Y, Z, X, \theta^*)$. The moment condition on the right-hand side of (3.2.3) is based on the subsample of observations with $D = 1$. Adopting the terminology of Robins et al. (1994, p. 848), we henceforth refer to the subsample of observations with $D = 1$ as the 'validation sample'; i. e. the validation sample is obtained from the original sample $(D_i, Y_i^*, Z_i, X_i : 1 \leq i \leq n)$ by discarding the observations (D_i, Y_i^*, Z_i, X_i) for which $D_i = 0$. We let $L_2(Z, X)$ denote the set of real-valued functions of Z, X with finite second moments.

3.3.1 Efficiency bounds

The equivalence in (3.2.3) reveals that under MAR, θ^* can be estimated from the validation sample alone. In practice, however, it may not be a good idea to estimate θ^* using the validation sample alone because of the efficiency loss resulting from discarding the observations on (Z, X) even though they are not missing. It is, therefore,

[46] The same condition leads to the global identification of θ^* whenever g is linear in θ^*.

important to know the efficiency bound for estimating θ^* in (3.1.2) under MAR. Loosely speaking, the efficiency bound for θ^* is the smallest asymptotic variance of an estimator that best uses the information from all non-missing observations.

We motivate the efficiency bound for estimating θ^* using Hristache and Patilea (2017, Theorem 1), which extends the results in Graham (2011, Theorem 2.1) to conditional moment restrictions. Although Hristache and Patilea consider a very general model with infinite-dimensional parameters, unlike us, they do not consider efficient estimation of the parameters of interest. Instead, they focus on showing that a moment condition with missing observations and the MAR assumption are equivalent to the moment condition in the validation sample implied by the MAR and the moment condition defining the propensity score function.

Consider the system of $(\dim g + 1)$ equations

$$\mathbb{E}\Big[\frac{Dg}{\pi} \;\Big|\; X\Big] = 0 \quad P_X\text{-a.s.} \tag{3.3.1a}$$

$$\mathbb{E}\Big[\frac{D}{\pi} - 1 \;\Big|\; Z, X\Big] = 0 \quad P_{Z,X}\text{-a.s.} \tag{3.3.1b}$$

Since $Dg \overset{(3.1.1)}{=} Dg(Y, Z, X, \theta^*)$, the moment conditions in (3.3.1) do not contain any missing observations. Note that (3.3.1a) identifies θ^* using the validation sample alone, whereas (3.3.1b) is the definition of π. Remarkably, by Theorem 1 of Hristache and Patilea, the moment conditions in (3.3.1) are equivalent to (3.1.2) and MAR, i. e.

$$\text{(3.3.1)} \iff \text{(3.1.2)} \,\&\, \text{MAR}\,. \tag{3.3.2}$$

The equivalence in (3.3.2) reveals that, under MAR, the efficiency bound for θ^* in (3.1.2) is equal to the efficiency bound for estimating θ^* in (3.3.1), which is a system of conditional moment restrictions with increasing conditioning sets. Following Ai and Chen (2012, Section 2), and Hristache and Patilea (2016, Section 4.1), we convert the sequential system in (3.3.1) into a conditional-on-X moment restriction whose moment functions are orthogonal to the moment function in (3.3.1b), by considering the residual from an orthogonal projection of the moment functions in (3.3.1a) onto the '$L_2(Z, X)$-span' of the moment function in (3.3.1b). This residual, which is free from the influence of (3.3.1b) in the sense that it is orthogonal to $D/\pi - 1$, satisfies a conditional-on-X moment restriction, which is then used to estimate θ^*.

For the remainder of the paper, let

$$\mu \overset{\text{def}}{=} \mu(Z, X, \theta^*) \overset{\text{def}}{=} \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid Z, X] \tag{3.3.3}$$

and

$$\begin{aligned} \rho &\overset{\text{def}}{=} \rho(\mathcal{A}, \theta^*, \pi, \mu) \overset{\text{def}}{=} \rho(\mathcal{A}, \theta^*, \pi(Z, X), \mu(Z, X, \theta^*)) && (\mathcal{A} \overset{\text{def}}{=} (D, Y, Z, X)) \\ &\overset{\text{def}}{=} \frac{Dg(Y, Z, X, \theta^*)}{\pi(Z, X)} - \mu(Z, X, \theta^*)\Big[\frac{D}{\pi(Z, X)} - 1\Big]. \end{aligned} \tag{3.3.4}$$

The functions π and μ appearing in the definition of ρ are estimable because, under MAR, μ is just the best non-parametric imputation of $g(Y^*, Z, X, \theta^*)$ based on (Z, X), i. e.

$$\mu \overset{\text{MAR}}{=} \mathbb{E}[g(Y, Z, X, \theta^*) \mid Z, X, D = 1]. \tag{3.3.5}$$

Define the $L_2(Z,X)$-span of $D/\pi - 1$ to be $\mathfrak{A} \overset{\text{def}}{=} \{a(D/\pi - 1) : a \in L_2(Z,X)\}$. It is shown in Appendix 3.C that (3.3.1) implies that ρ is the residual from the coordinate-wise projection of Dg/π onto the linear space $\mathfrak{A}$, and that ρ satisfies the conditional-on-X moment restriction

$$\mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X] \overset{P_X\text{-a.s.}}{=} 0. \tag{3.3.6}$$

Therefore, estimation of θ^* can be based on (3.3.6).

In fact, (3.3.6) can also deliver an efficient estimator. It is shown in Appendix 3.C that

$$\text{(3.3.1)} \implies \begin{cases} \partial_{\theta^*} \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X] \overset{P_X\text{-a.s.}}{=} J & \text{(3.3.7a)} \\ \partial_{\pi} \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X] \overset{P_X\text{-a.s.}}{=} 0 & \text{(3.3.7b)} \\ \partial_{\mu} \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X] \overset{P_X\text{-a.s.}}{=} 0. & \text{(3.3.7c)} \end{cases}$$

Hence, by Ai and Chen (2003, Theorems 4.1 and 6.1), it is straightforward to verify that the efficiency bound for estimating θ^* in (3.3.6) is given by $(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$, where $\Omega_\rho \overset{\text{def}}{=} \mathbb{E}[\rho\rho' \mid X] \overset{(3.3.6)}{=} \operatorname{Var}[\rho \mid X]$. Furthermore, as demonstrated subsequently in Lemma 3.3.1, $(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}$ is also the semi-parametric efficiency bound for estimating θ^* in (3.1.2). Therefore, efficient estimation of θ^* can be based on (3.3.6). Consequently, the moment function ρ (which can be written as a weighted sum of g and μ with weights D/π and $1 - D/\pi$ respectively) may also be interpreted as the optimal linear combination of g and its best non-parametric imputation μ.

Let $\sigma_g^2 \overset{\text{def}}{=} \sigma_g^2(X) \overset{\text{def}}{=} \mathbb{E}[g'g \mid X]$, and $\|\cdot\|_\infty$ denote the supremum norm, e. g. $\|\sigma_g^2\|_\infty \overset{\text{def}}{=} \sup_{\operatorname{supp}(X)} \sigma_g^2$. The efficiency bound in Lemma 3.3.1 is obtained under the following conditions.

Assumption 3.3.1. (i) $\inf_{\operatorname{supp}(X,Z)} \pi > 0$; (ii) $\mathbb{E} \operatorname{tr} J'J < \infty$ *and* $\|\sigma_g^2\|_\infty < \infty$; (iii) *The matrix* $\mathbb{E}J'\Omega_\rho^{-1}J$ *exists and is nonsingular, and the matrix* $\mathbb{E}[J'\Omega_\rho^{-1}(1-\pi)\pi^{-1}\mu\mu'\Omega_\rho^{-1}J]$ *exists.*

(i) is necessary for θ^* to be $n^{1/2}$-estimable. In (ii), $\mathbb{E} \operatorname{tr} J'J < \infty$ implies that each element of J has finite second moment. Consequently, $\operatorname{span}(J)$, which denotes the set of all linear combinations of the column vectors of J, i. e. the column space of J, is closed in $L_2(X)^{\times \dim g}$. The condition $\|\sigma_g^2\|_\infty < \infty$ uniformly bounds the skedastic function for each coordinate of g. (i) and (ii) are used in the proof of Lemma 3.C.1. (iii), which implies that the efficiency bound in (3.3.8) is well-defined, is also necessary for θ^* to be $n^{1/2}$-estimable.

Lemma 3.3.1. *Let Assumptions* 3.2.1*,* 3.2.2*,* 3.2.3*, and* 3.3.1 *hold. Then, the efficiency bound for estimating* θ^* *in* (3.1.2) *is given by*[47]

$$\text{l.b.}(\theta^*) \overset{\text{def}}{=} (\mathbb{E}J'\Omega_\rho^{-1}J)^{-1}. \tag{3.3.8}$$

The efficiency bound for θ^* *does not decrease if* π *is known to be parametrically specified up to a finite-dimensional parameter, or even if* π *is fully known.*

Since Lemma 3.3.1 does not require $\theta \mapsto g(Y^*, Z, X, \theta)$ to be differentiable, the efficiency bound in (3.3.8) remains valid for non-smooth moment functions arising, e. g. in quantile regression models. The result in Lemma 3.3.1 that the efficiency bound

[47] The abbreviation 'l.b.' stands for 'lower bound' because the semi-parametric efficiency bound is the greatest lower bound for the asymptotic variance of any $n^{1/2}$-consistent regular estimator.

for θ^* remains the same irrespective of whether π is fully unknown, or fully known, or known up to a finite-dimensional parameter, can be interpreted as the propensity score function being ancillary to θ^* (Hahn, 1998, p. 319). This is not surprising because π does not enter the moment condition (3.1.2) through which θ^* is defined. As noted in X. Chen et al. (2008, Section 4.2, p. 822) and Graham (2011, p. 439, and the references cited therein), ancillarity of π implies that, in order to obtain an asymptotically efficient estimator of θ^*, the propensity score function should be non-parametrically estimated even if it is parametrically specified, or, indeed, even if it is fully known.[48]

To get an idea about the efficiency gains obtained when all non-missing observations are used to estimate θ^* (and not just those in the validation sample), the efficiency bound in Lemma 3.3.1 can be compared with the efficiency bound for estimating θ^* using the moment condition based on the validation sample alone, i. e. the right-hand side of (3.2.3). Let $\text{l.b.}(\theta^*)\big|_{\text{VS}}$ denote the efficiency bound for θ^* based on $\mathbb{E}[Dg(Y,Z,X,\theta^*)/\pi(Z,X) \mid X] \overset{P_X\text{-a.s.}}{=} 0$. By Ai and Chen (2003, Theorems 4.1 and 6.1),

$$\text{l.b.}(\theta^*)\big|_{\text{VS}} = (\mathbb{E}\mathcal{J}'_{\varpi_*}\Sigma^{-1}\mathcal{J}_{\varpi_*})^{-1}, \tag{3.3.9}$$

where $\Sigma \overset{\text{def}}{=} \mathbb{E}[\frac{Dgg'}{\pi^2} \mid X]$, and $\varpi_* \overset{\text{def}}{=} (\varpi_*^{(1)}, \ldots, \varpi_*^{(\dim\theta^*)}) \in L_2(Z,X)^{\times \dim\theta^*}$ is a $(\dim\theta^*) \times 1$ vector of real-valued functions chosen such that $\mathbb{E}\mathcal{J}'_{\varpi_*}\Sigma^{-1}\mathcal{J}_{\varpi_*} \leq_L \mathbb{E}\mathcal{J}'_{\varpi}\Sigma^{-1}\mathcal{J}_{\varpi}$ for all $\varpi \overset{\text{def}}{=} (\varpi^{(1)}, \ldots, \varpi^{(\dim\theta^*)}) \in L_2(Z,X)^{\times \dim\theta^*}$,[49] where

$$\mathcal{J}_{\varpi} \overset{\text{def}}{=} \left[J_1 + \mathbb{E}\Big[\frac{Dg}{\pi^2}\varpi^{(1)} \;\Big|\; X\Big] \quad \cdots \quad J_{(\dim\theta^*)} + \mathbb{E}\Big[\frac{Dg}{\pi^2}\varpi^{(\dim\theta^*)} \;\Big|\; X\Big]\right]_{(\dim g)\times(\dim\theta^*)}$$

and J_k is the kth column of J. Clearly, $\mathbb{E}\mathcal{J}'_{\varpi_*}\Sigma^{-1}\mathcal{J}_{\varpi_*} \leq_L \mathbb{E}J'\Sigma^{-1}J$ from the definition of ϖ_*. Moreover, MAR and (3.3.12) imply that

$$\Sigma = \Omega_\rho + \mathbb{E}\Big[\frac{1-\pi}{\pi}\mu\mu' \;\Big|\; X\Big]. \tag{3.3.10}$$

Consequently, $\Omega_\rho \leq_L \Sigma$ from which it follows that

$$(\mathbb{E}J'\Omega_\rho^{-1}J)^{-1} \leq_L (\mathbb{E}\mathcal{J}'_{\varpi_*}\Sigma^{-1}\mathcal{J}_{\varpi_*})^{-1} \iff \text{l.b.}(\theta^*) \leq_L \text{l.b.}(\theta^*)\big|_{\text{VS}}. \tag{3.3.11}$$

Equation (3.3.11) reveals that, in general, an asymptotically efficient estimator of θ^*—which necessarily uses all the observations and not just those in the validation sample—beats any estimator constructed using the validation sample alone.

The special case $\text{l.b.}(\theta^*) = \text{l.b.}(\theta^*)\big|_{\text{VS}}$, namely, when using the validation sample alone leads to an efficient estimator, arises when there are no non-missing endogenous variables, i. e. when Z is the empty vector (denoted by $Z = \varnothing$). In this case, the propensity score is a function of X alone so that $\mathcal{J} \overset{(3.2.3)}{=} J$, and $\mu \overset{\text{def}}{=} \mathbb{E}[g \mid Z,X] \overset{Z=\varnothing}{=} \mathbb{E}[g \mid X] \overset{(3.1.2)}{=} 0$ P_X-a.s., which implies that $\Sigma \overset{(3.3.10)}{=} \Omega_\rho$ P_X-a.s. Consequently, $\text{l.b.}(\theta^*)\big|_{Z=\varnothing} = \text{l.b.}(\theta^*)\big|_{\text{VS}}$. In other words, if all endogenous outcomes and endogenous explanatory variables in the model are missing, then estimating θ^* using the validation subsample alone is

[48] A similar issue arises in estimating models with stratified samples when the stratum shares are known (Tripathi, 2011a, Section 2).

[49] The symbol '$\leq_L$' denotes the usual (Löwner) order on the set of symmetric matrices. Namely, $M_1 \leq_L M_2$ for symmetric matrices M_1, M_2 means that the matrix $M_1 - M_2$ is negative semidefinite.

asymptotically efficient.[50] This result, which generalises the findings in Hristache and Patilea (2016, Section 4.2) and Hristache and Patilea (2017, p. 740) for missing outcomes, is worth stating separately.

Corollary 3.3.1. *The efficiency gains in estimating θ^*, measured by the coordinate-wise ratio* $\mathrm{l.b.}(\theta^*)\big|_{\mathrm{VS}}/\mathrm{l.b.}(\theta^*)$*, are due to the presence of the non-missing endogenous variables.*

Corollary 3.3.1 can be interpreted as a result about imputing the values of missing endogenous variables, which may be appealing to applied researchers. Recalling that ρ is the optimal linear combination of g and its best non-parametric imputation μ (cf. the discussion after (3.3.7)), Corollary 3.3.1 says that the efficiency gains in estimating θ^* using all of the non-missing observations—and not just those in the validation sample—arise only when $\mu \neq 0$. Since $\mu \overset{P_{Z,X}\text{-a.s.}}{=} 0 \iff \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid Z, X] \overset{P_{Z,X}\text{-a.s.}}{=} 0$, we have that

$$\begin{aligned}\mu \overset{P_{Z,X}\text{-a.s.}}{=} 0 &\implies \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0 \quad \& \quad \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid Z] \overset{P_Z\text{-a.s.}}{=} 0 \\ &\implies \mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0 \quad \& \quad Z \text{ is exogenous or empty.}\end{aligned}$$

Therefore, if the model (3.1.2) is correctly specified, then $\mu \neq 0$ only when Z is endogenous and non-empty, i. e. when the information used to impute g (the conditioning set in (3.3.3)) is strictly larger than the information used to estimate θ^* (the conditioning set in (3.1.2)); equivalently, when the conditioning set in (3.3.1b) is strictly larger than the conditioning set in (3.3.1a).

The following example illustrates that imputing missing outcomes in linear regression models can lead to efficiency gains only when non-missing endogenous regressors are present.

Example 3.3.1 (When should missing outcomes be imputed?)**.** Consider the linear regression model $Y^* = \alpha_0^* + X_{\mathrm{in}}'\beta_0^* + U$ with $\mathbb{E}[U \mid X_{\mathrm{in}}] \overset{P_{X_{\mathrm{in}}}\text{-a.s.}}{=} 0$, where the outcomes may be missing, and there are no non-missing endogenous regressors and excluded instruments. In this model, $g \overset{\mathrm{def}}{=} U$; hence, $\mu \overset{\mathrm{def}}{=} \mu(X_{\mathrm{in}}, \alpha_0^*, \beta_0^*) = \mathbb{E}[Y^* - \alpha_0^* - X_{\mathrm{in}}'\beta_0^* \mid X_{\mathrm{in}}] \overset{P_{X_{\mathrm{in}}}\text{-a.s.}}{=} 0$. Consequently, the validation sample alone can be used to construct a semi-parametrically efficient estimator of (α_0^*, β_0^*). Indeed, Lemma 3.3.1 shows that the efficiency bound for estimating (α_0^*, β_0^*) is given by $(\mathbb{E}\tilde{\pi}J'\Omega_g^{-1}J)^{-1}$, where $\tilde{\pi} \overset{\mathrm{def}}{=} \mathbb{E}[D \mid X_{\mathrm{in}}]$, $J = -[\,1 \; X_{\mathrm{in}}'\,]$, and $\Omega_g \overset{\mathrm{def}}{=} \mathbb{E}[gg' \mid X_{\mathrm{in}}]$. By (3.3.9), this coincides with the efficiency bound using the validation sample alone; moreover, based on the moment condition $\mathbb{E}[D(Y - \alpha_0^* - X_{\mathrm{in}}'\beta_0^*) \mid X_{\mathrm{in}}] \overset{P_{X_{\mathrm{in}}}\text{-a.s.}}{=} 0$, which holds only in the validation sample, the estimator proposed in (3.3.15) attains the bound. In this model, imputing the missing Y^* using X_{in} (as no other non-missing endogenous/exogenous variables are present) and employing the imputed values to estimate (α_0^*, β_0^*) does not lead to any efficiency gains. This is easily seen for the least-squares (LS) estimator, which is not semi-parametrically efficient but serves to illustrate the point. Since $\mathbb{E}[Y^* \mid X_{\mathrm{in}}, D = 1] \overset{\mathrm{MAR}}{=} \mathbb{E}[Y^* \mid X_{\mathrm{in}}] = \alpha_0^* + X_{\mathrm{in}}'\beta_0^*$, the missing Y^* can be replaced by their imputed value $\hat{Y} \overset{\mathrm{def}}{=} \hat{\alpha}_{0\mathrm{VS}} + X_{\mathrm{in}}'\hat{\beta}_{0\mathrm{VS}}$, where $(\hat{\alpha}_{0\mathrm{VS}}, \hat{\beta}_{0\mathrm{VS}}) \overset{\mathrm{def}}{=} \operatorname{argmin}_{\alpha,\beta} \sum_{i:\, D_i=1}(Y_i - \alpha - X_{\mathrm{in},i}'\beta)^2$ is the estimator of (α_0^*, β_0^*) from

[50] If there is no missingness at all, i. e. $Y^* = Y$ with probability 1, then $\rho = g$ and (3.3.8) becomes $(\mathbb{E}J'\Omega_g^{-1}J)^{-1}$, which is the well-known efficiency bound for estimating θ^* in the model $\mathbb{E}[g(Y, Z, X, \theta^*) \mid X] = 0$ P_X-a.s.

the validation sample alone. The LS estimator of (α_0^*, β_0^*) from the full sample, obtained by replacing the missing outcomes with their imputed values, is then

$$\begin{aligned}(\hat{\alpha}_{0\mathrm{LS}}, \hat{\beta}_{0\mathrm{LS}}) &\stackrel{\text{def}}{=} \operatorname*{argmin}_{\alpha,\beta} \sum_{i\colon D_i=1}^{n} (Y_i - \alpha - X_{\mathrm{in},i}'\beta)^2 + \sum_{i\colon D_i=0}^{n} (\hat{Y}_i - \alpha - X_{\mathrm{in},i}'\beta)^2 \\ &= \operatorname*{argmin}_{\alpha,\beta} \sum_{i\colon D_i=1}^{n} (Y_i - \alpha - X_{\mathrm{in},i}'\beta)^2 + \underbrace{\sum_{i\colon D_i=0}^{n} (\hat{\alpha}_{0\mathrm{VS}} + X_{\mathrm{in},i}'\hat{\beta}_{0\mathrm{VS}} - \alpha - X_{\mathrm{in},i}'\beta)^2}_{=0 \text{ if } (\alpha,\beta)=(\hat{\alpha}_{0\mathrm{VS}}, \hat{\beta}_{0\mathrm{VS}})} \\ &= (\hat{\alpha}_{0\mathrm{VS}}, \hat{\beta}_{0\mathrm{VS}}).\end{aligned}$$

Therefore, imputing the missing outcomes in linear regression models that have no non-missing endogenous regressors does not lead to efficiency gains.

Next, suppose that in the previous specification, we allow for non-missing endogenous regressors and excluded instruments, i. e. now $Y^* = \alpha^* + X_{\mathrm{in}}'\beta^* + Z'\gamma^* + \varepsilon$ with $\mathbb{E}[\varepsilon \mid X] \stackrel{P_X\text{-a.s.}}{=} 0$ and $X \stackrel{\text{def}}{=} (X_{\mathrm{in}}, X_{\mathrm{ex}})$. Here, $g \stackrel{\text{def}}{=} \varepsilon$; hence, $\mu \stackrel{\text{def}}{=} \mu(Z, X, \alpha^*, \beta^*, \gamma^*) = \mathbb{E}[\varepsilon \mid Z, X] \neq 0$. Consequently, no estimator of $(\alpha^*, \beta^*, \gamma^*)$ using the validation sample alone is semi-parametrically efficient. Indeed, the efficiency bound for estimating $(\alpha^*, \beta^*, \gamma^*)$ (cf. Example 3.3.2)—which is attained by the estimator in (3.3.15) with ρ defined in (3.3.4)—is strictly smaller than the efficiency bound for estimating $(\alpha^*, \beta^*, \gamma^*)$ from the validation sample alone because $\mu \neq 0$. In this model, imputing the missing Y^* using (Z, X) and employing the imputed values to estimate $(\alpha^*, \beta^*, \gamma^*)$ can lead to efficiency gains. This is easily seen for the two-stage least-squares (2SLS) estimator, which is not semi-parametrically efficient but illustrates the point fittingly. Since $\mathbb{E}[Y^* \mid Z, X, D=1] \stackrel{\text{MAR}}{=} \mathbb{E}[Y^* \mid Z, X] = \alpha^* + X_{\mathrm{in}}'\beta^* + Z'\gamma^* + \mu$, the missing Y^* are imputed by $\check{Y} \stackrel{\text{def}}{=} \hat{\alpha}_{\mathrm{VS}} + X_{\mathrm{in}}'\hat{\beta}_{\mathrm{VS}} + Z'\hat{\gamma}_{\mathrm{VS}} + \hat{\mu}(Z, X)$, where $(\hat{\alpha}_{\mathrm{VS}}, \hat{\beta}_{\mathrm{VS}}, \hat{\gamma}_{\mathrm{VS}}) \stackrel{\text{def}}{=} \operatorname{argmin}_{\alpha,\beta,\gamma} \sum_{i\colon D_i=1}^{n} (Y_i - \alpha - X_{\mathrm{in},i}'\beta - \hat{Z}_{\mathrm{VS},i}'\gamma)^2$ is the 2SLS estimator of $(\alpha^*, \beta^*, \gamma^*)$ in the validation sample, $\hat{Z}_{\mathrm{VS}}$ is the predicted Z from the first-stage obtained by estimating the reduced form equations for Z in the validation sample, and $\hat{\mu}(Z, X) \stackrel{\text{def}}{=} \hat{\mu}(Z, X, \hat{\alpha}_{\mathrm{VS}}, \hat{\beta}_{\mathrm{VS}}, \hat{\gamma}_{\mathrm{VS}})$ is obtained by non-parametrically regressing the second-stage residual $\hat{\varepsilon}_{\mathrm{VS}} \stackrel{\text{def}}{=} \hat{\alpha}_{\mathrm{VS}} - X_{\mathrm{in}}'\hat{\beta}_{\mathrm{VS}} - Z'\hat{\gamma}_{\mathrm{VS}}$ on (Z, X). Letting $\hat{Z}$ be the predicted Z from the first-stage reduced form equations for Z in the full sample, the 2SLS estimator of $(\alpha^*, \beta^*, \gamma^*)$ in the full sample is then

$$\begin{aligned}(\hat{\alpha}_{2\mathrm{SLS}}, \hat{\beta}_{2\mathrm{SLS}}, \hat{\gamma}_{2\mathrm{SLS}}) &\stackrel{\text{def}}{=} \operatorname*{argmin}_{\alpha,\beta,\gamma} \sum_{i\colon D_i=1}^{n} (Y_i - \alpha - X_{\mathrm{in},i}'\beta - \hat{Z}_i'\gamma)^2 + \sum_{i\colon D_i=0}^{n} (\check{Y}_i - \alpha - X_{\mathrm{in},i}'\beta - \hat{Z}_i'\gamma)^2 \\ &= \operatorname*{argmin}_{\alpha,\beta,\gamma} \sum_{i\colon D_i=1}^{n} (Y_i - \alpha - X_{\mathrm{in},i}'\beta - \hat{Z}_i'\gamma)^2 \\ &\qquad + \underbrace{\sum_{i\colon D_i=0}^{n} (\hat{\alpha}_{\mathrm{VS}} + X_{\mathrm{in},i}'\hat{\beta}_{\mathrm{VS}} + Z_i'\hat{\gamma}_{\mathrm{VS}} + \hat{\mu}(Z_i, X_i) - \alpha - X_{\mathrm{in},i}'\beta - \hat{Z}_i'\gamma)^2}_{\neq 0} \\ &\neq (\hat{\alpha}_{\mathrm{VS}}, \hat{\beta}_{\mathrm{VS}}, \hat{\gamma}_{\mathrm{VS}}).\end{aligned}$$

Therefore, imputing the missing outcomes in linear regression models where non-missing endogenous regressors are present can lead to efficiency gains. □

Before ending this section, we describe an alternative expression for Ω_ρ that is useful in many applications. It is shown in Appendix 3.C that

$$\pi\mathbb{E}[\rho\rho' \mid Z, X] + (1-\pi)\mu\mu' = \mathbb{E}[gg' \mid Z, X]. \tag{3.3.12}$$

Since $\mathrm{Var}[g \mid Z, X] = \mathbb{E}[gg' \mid Z, X] - \mu\mu'$,

$$(3.3.12) \iff \mathbb{E}[\rho\rho' \mid Z, X] = \pi^{-1}\,\mathrm{Var}[g \mid Z, X] + \mu\mu'.$$

Consequently, since $\mathbb{E}[\rho \mid X] \overset{(3.3.6)}{=} 0$ P_X-a.s.,

$$\Omega_\rho = \mathbb{E}[\pi^{-1}\,\mathrm{Var}(g \mid Z, X) \mid X] + \mathbb{E}[\mu\mu' \mid X]. \tag{3.3.13}$$

Example 3.3.2 (Example 3.1.1 contd.)**.** In the linear regression model where only the outcome variable may be missing, $g(Y^*, Z, X, \theta^*) = U = Y^* - \alpha^* - X_{\text{in}}'\beta^* - Z'\gamma^*$. Hence, by Lemma 3.3.1, the efficiency bound for estimating θ^* is given by $(\mathbb{E}J'J/\Omega_\rho)^{-1}$, where $J = -\big[\begin{smallmatrix} 1 & X_{\text{in}}' & \mathbb{E}[Z'|X] \end{smallmatrix}\big]$, $\Omega_\rho \overset{(3.3.13)}{=} \mathbb{E}[\pi^{-1}\,\mathrm{Var}(Y^* \mid Z, X) \mid X] + \mathbb{E}[\mu^2 \mid X]$, and $\mu = \mathbb{E}[U \mid Z, X] = \mathbb{E}[Y^* \mid Z, X] - \alpha^* - X_{\text{in}}'\beta^* - Z'\gamma^*$. □

Example 3.3.3 (Example 3.1.2 contd.)**.** In the linear regression model where the outcome is always observed, but the right-hand side endogenous variables may be missing, $g(Y^*, Z, X, \theta^*) = Z - \alpha^* - X_{\text{in}}'\beta^* - Y^{*\prime}\gamma^*$. Hence, by Lemma 3.3.1, the efficiency bound for estimating θ^* is given by $(\mathbb{E}J'J/\Omega_\rho)^{-1}$, where $J = -\big[\begin{smallmatrix} 1 & X_{\text{in}}' & \mathbb{E}[Y^{*\prime}|X] \end{smallmatrix}\big]$, $\Omega_\rho \overset{(3.3.13)}{=} \gamma^{*\prime}\mathbb{E}[\pi^{-1}\,\mathrm{Var}(Y^* \mid Z, X) \mid X]\gamma^* + \mathbb{E}[\mu^2 \mid X]$, and $\mu = Z - \alpha^* - X_{\text{in}}'\beta^* - \gamma^{*\prime}\mathbb{E}[Y^* \mid Z, X]$. □

Example 3.3.4 (Unconditional moment equalities)**.** If there is no conditioning in (3.1.2), then the efficiency bound in Lemma 3.3.1 reduces to the one obtained by X. Chen et al. (2008, Theorem 1) and Graham (2011, p. 439) for estimating parameters defined via unconditional moment equalities. Indeed, (3.1.2) can be formally converted into the model of X. Chen et al. (2008, Eqn. 2) by applying the following two-step procedure to (3.1.2): (i) First, let $X \overset{\text{def}}{=} \varnothing$ so that the conditioning disappears; (ii) Then, replace Z by X. This leads to the propensity score $\tilde{\pi} \overset{\text{def}}{=} \tilde{\pi}(X) \overset{\text{def}}{=} \mathbb{E}[D \mid X]$ and the unconditional moment restriction model

$$\mathbb{E}g(Y^*, X, \theta^*) = 0. \tag{3.3.14}$$

in (3.3.14), assume that $\dim g \geq \dim \theta^*$ and that θ^* is identified. Then, applying the procedure in (i) and (ii) to the efficiency bound in Lemma 3.3.1, it follows that the efficiency bound for estimating θ^* in (3.3.14) is given by $(J'\Omega_\rho^{-1}J)^{-1}$, where, now, $J = \partial_\theta \mathbb{E}g(Y^*, X, \theta^*)$, $\Omega_\rho \overset{(3.3.13)}{=} \mathbb{E}[\tilde{\pi}^{-1}\,\mathrm{Var}(g \mid X)] + \mathbb{E}\mu\mu'$ with $g = g(Y^*, X, \theta^*)$ and $\mu = \mathbb{E}[g \mid X]$. As discussed in X. Chen et al. (2008, Section 2.1), $\tilde{\pi}$ is ancillary to θ^* in (3.3.14). □

3.3.2 The smoothed empirical likelihood

If π and μ are fully known, then the smoothed empirical likelihood (SEL) estimator of θ^* (Kitamura et al., 2004) based on the conditional moment restriction (3.3.6) will be asymptotically efficient, i. e. its asymptotic variance will equal the semi-parametric efficiency bound in (3.3.8), because $J \overset{(3.3.7a)}{=} \partial_\theta \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X]$ P_X-a.s. In fact, the asymptotic variance of the SEL estimator will not change even if π and μ are replaced by

their non-parametric estimators.[51] Therefore, we estimate θ^* using the SEL approach, which entails maximising the empirical likelihood of the data subject to (3.3.6). Smoothing the empirical likelihood is required because (3.3.6) is a conditional restriction and the coordinates of X are continuously distributed.[52]

By (3.2.2) and (3.3.5), π and μ can be estimated by the kernel estimators

$$\hat{\pi}_c(Z,X) \stackrel{\text{def}}{=} \frac{\sum_{k=1}^n D_k \mathcal{K}_c(Z_k - Z, X_k - X)}{\sum_{k=1}^n \mathcal{K}_c(Z_k - Z, X_k - X)}$$
$$\hat{\mu}_d(Z,X,\theta) \stackrel{\text{def}}{=} \frac{\sum_{k=1}^n g(Y_k, Z_k, X_k, \theta)\mathcal{K}_d(Z_k - Z, X_k - X)\mathbb{1}(D_k = 1)}{\sum_{k=1}^n \mathcal{K}_d(Z_k - Z, X_k - X)\mathbb{1}(D_k = 1)},$$

where $\mathcal{K}_c(\cdot) \stackrel{\text{def}}{=} K(\cdot/c_n)$ and $\mathcal{K}_d(\cdot) \stackrel{\text{def}}{=} K(\cdot/d_n)$ are kernel functions and $c \stackrel{\text{def}}{=} (c_n)$ and $d \stackrel{\text{def}}{=} (d_n)$ the bandwidths. Hence, letting $\hat{\rho}(\mathcal{A}_j, \theta) \stackrel{\text{def}}{=} \rho\big(\mathcal{A}_j, \theta, \hat{\pi}_c(Z_j, X_j), \hat{\mu}_d(Z_j, X_j, \theta)\big)$, $j = 1, \dots, n$, the SEL estimator of θ^* is defined as

$$\hat{\theta} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta \in \Theta} \mathrm{SEL}_{\mathbb{T}}(\theta), \tag{3.3.15}$$

where

$$\mathrm{SEL}_{\mathbb{T}}(\theta) \stackrel{\text{def}}{=} -\sum_{i=1}^n \mathbb{T}_{i,n} \max_{\lambda_i \in \mathbb{R}^{\dim \rho}} \sum_{j=1}^n w_{ij} \log\big(1 + \lambda_i' \hat{\rho}(\mathcal{A}_j, \theta)\big),$$

and $w_{ij} \stackrel{\text{def}}{=} \mathcal{K}_b(X_i - X_j)/\sum_{k=1}^n \mathcal{K}_b(X_i - X_k)$ are the kernel weights, with bandwidth $b \stackrel{\text{def}}{=} (b_n)$, used to construct the local empirical likelihood $\sum_{j=1}^n w_{ij} \log\big(1 + \lambda' \hat{\rho}(\mathcal{A}_j, \theta)\big)$. The indicator $\mathbb{T}_{i,n}$ is a trimming function (defined subsequently) introduced to deal with the instability of the local empirical log-likelihood caused by the density of the conditioning variables becoming too small in the tails. Cf. Kitamura et al. (2004, Section 2) for the derivation of the SEL objective function and the intuition behind it.

3.3.3 Inference

The empirical likelihood approach provides a convenient unified environment for testing hypotheses about θ^* based on the likelihood ratio (LR) statistic $\mathrm{LR}(\theta) \stackrel{\text{def}}{=} 2[\mathrm{SEL}_{\mathbb{T}}(\hat{\theta}) - \mathrm{SEL}_{\mathbb{T}}(\theta)]$. E. g. the parametric restriction $\mathrm{H}_0 : R(\theta^*) = 0$, where R is a vector of smooth functions, is rejected for large values of $\mathrm{LR}(\hat{\theta}_R)$, where $\hat{\theta}_R \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta : R(\theta) = 0} \mathrm{SEL}_{\mathbb{T}}(\theta)$. The LR statistic is asymptotically pivotal because $\mathrm{LR}(\hat{\theta}_R)$ is distributed as a $\chi^2_{\dim R}$ random variable in large samples when H_0 is true. Although a Wald statistic based on $\hat{\theta}$ can also be constructed, it is less attractive than the LR statistic because the latter is internally studentised. As in parametric situations, the LR statistic can be inverted to obtain asymptotically valid confidence intervals. E. g. the lower level random set $\{\theta \in \Theta : \mathrm{LR}(\theta) \leq k_\tau\}$, where $\tau \in (0,1)$ and k_τ denotes the $1 - \tau$ quantile of a $\chi^2_{\dim \theta^*}$ random variable, is a LR confidence region for θ^* whose coverage probability approaches

[51] This is because the effect of estimating a parameter is captured through its Jacobian, and the Jacobians with respect to π and μ vanish by (3.3.7b) and (3.3.7c).

[52] We assume for convenience that all coordinates of X are continuously distributed. Discrete coordinates can be easily accommodated by smoothing them along with the continuous ones. If all coordinates of X are discrete, then smoothing is not necessary, and it can be shown (cf. Section 3.4 and Appendix 3.D.2) that a version of $\mathrm{SEL}_{\mathbb{T}}$ constructed with $\mathbb{T}_{i,n} \stackrel{\text{def}}{=} 1$ and $w_{ij} \stackrel{\text{def}}{=} \mathbb{1}(X_i = X_j)/\sum_{k=1}^n \mathbb{1}(X_i = X_k)$ coincides with unconditional empirical likelihood.

$1-\tau$ as $n \to \infty$. A nice property of the LR confidence regions is that they are invariant to nonsingular transformations of the moment conditions. Moreover, being subsets of Θ by construction, they also automatically respect natural range restrictions on the parameters.

3.4 Simulation study

In this section, we compare the small sample behaviour of $\hat{\theta}$ with the estimator constructed using only the validation sample.

3.4.1 The meaning of 'identification using the validation sample' in applied research

Recall from (3.2.3) that, under MAR, (3.1.2) is equivalent to the moment condition

$$\mathbb{E}\Big[\frac{Dg(Y,Z,X,\theta^*)}{\pi(Z,X)} \;\Big|\; X\Big] \overset{P_X\text{-a.s.}}{=} 0, \tag{3.4.1}$$

which only uses the validation sample. However, when applied researchers talk about identifying (3.1.2) using the validation sample alone, they usually have in mind the model

$$\mathbb{E}[Dg(Y,Z,X,\theta^*) \mid X] \overset{P_X\text{-a.s.}}{=} 0, \tag{3.4.2}$$

i. e. the original moment condition (3.1.2) holds in the validation sample without being scaled by the reciprocal of the propensity score. Since (3.4.1) reduces to (3.4.2) when $\pi(Z,X)$ does not depend on Z, in this section we assume that

Assumption 3.4.1. $\mathbb{E}[D \mid Z,X] = \mathbb{E}[D \mid X] =: \tilde{\pi}(X)$.

Assumption 3.4.1, which is equivalent to the condition that $D \perp\!\!\!\perp Z \mid X$ for all individuals, is also maintained in Hristache and Patilea (2017, p. 736), cf. their discussion of the partially linear single-index model. Under MAR and Assumption 3.4.1, $\jmath_{\varpi_*} = J$ and $\Sigma = \Omega_g/\tilde{\pi}$, where $\Omega_g \overset{\text{def}}{=} \mathbb{E}[gg' \mid X]$. Therefore,

$$\text{MAR } \& \text{ Ass. 3.4.1} \implies \text{l.b.}(\theta^*)\big|_{\text{VS}} \overset{(3.3.9)}{=} (\mathbb{E}\tilde{\pi} J'\Omega_g^{-1} J)^{-1}. \tag{3.4.3}$$

Henceforth, the SEL estimator of θ^* in (3.4.2) is denoted by $\hat{\theta}_{\text{VS}}$.

3.4.2 Designs

We consider two designs with the same structural model, namely, a simplified version of the linear IV regression in Example 3.1.1 given by $Y^* \overset{\text{def}}{=} \alpha^* + \gamma^* Z + U\sigma(X)$, where the outcome Y^* is missing for some individuals, the single regressor Z is endogenous, and X is the sole excluded IV for Z, i. e. X satisfies $\mathbb{E}[U \mid X] = 0$ P_X-a.s. The difference between the designs is in how Z and X are modelled. In Design 1, Z and X are both continuously distributed, and the reduced form equation for Z is given by $Z \overset{\text{def}}{=} \zeta_0 + \zeta_1 X + V$. In contrast, in Design 2, Z and X are both dummy variables, and the reduced form equation

for Z is given by $Z \overset{\text{def}}{=} \mathbb{1}(\zeta_0 + \zeta_1 X + V > 0)$.[53] In both designs, $\left[\begin{smallmatrix} U \\ V \end{smallmatrix}\right] \,\big|\, X \sim \mathcal{N}(\left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right], \left[\begin{smallmatrix} \sigma_U^2 & \sigma_{UV} \\ \sigma_{UV} & \sigma_V^2 \end{smallmatrix}\right])$ with $\sigma_U^2 = 1$, $\sigma_V^2 = 2$, $\sigma_{UV} = 1$, and $\alpha^* = \gamma^* = \zeta_1 = 1$. The reduced form intercept ζ_0 differs across the designs ($\zeta_0 = 1$ in Design 1, and $\zeta_0 = 0$ in Design 2) to ensure that $\mathbb{E}Z$ is close to $\mathbb{E}X$.

Throughout this section, Φ denotes the cumulative distribution function, and ϕ the probability density function, of a $\mathcal{N}(0,1)$ random variable. The results reported in this section are based on 5000 Monte Carlo replications, and $n = 500, 1000, 2000, 4000$, corresponding to 'relatively small', 'small', 'medium', and 'large' sample sizes. We call $n = 500$ to be a relatively small sample because it includes the missing observations. Indeed, if $n = 500$ then a validation sample of approximately 200 or fewer observations in some draws can be reasonably considered to be relatively small in the semi-parametric context.

Design 1

In this design, $X \sim \text{Unif}[0,1]$. The skedastic function $\sigma^2(x) \overset{\text{def}}{=} |x - r|^\nu + 1/15$, with $r = -1/3$ and $\nu = 2$, due to Cragg (1983), is popular with researchers to model conditional heteroskedasticity (and is used by us as well). The parameter ν determines the degree of heteroskedasticity (conditional homoskedasticity follows if $\nu = 0$). The regressor Z can be classified as being strongly endogenous in the heteroskedastic case because $\text{corr}\big(Z, U\sigma(X)\big) \approx 0.66$ when $\nu = 2$. This poses a serious problem because the bias of the slope coefficient, relative to its true value, when Y is regressed on Z in the validation sample, is $\approx 42.1\%$ (averaged across the simulations). There is no issue with weak IV because $\text{corr}(Z, X) \approx 0.20$ and in those Monte Carlo replications where the first-stage F-statistics are < 10 new data were re-generated until the first-stage F-statistics became ≥ 10 ($\approx 57\%$ of all replications for $n = 500$ and 0.4% for $n = 2000$). The non-missingness indicator D is drawn from a Bernoulli distribution with success probability $\tilde{\pi}(X) \overset{\text{def}}{=} l + (u - l)\Phi\big((X - r_{\tilde{\pi}})/\sigma_{\tilde{\pi}}\big)$, where $l = 0.25$, $u = 0.95$, $r_{\tilde{\pi}} = 0.1$, and $\sigma_{\tilde{\pi}} = 0.5$, are chosen to make $\tilde{\pi}$ be bounded away from 0 and 1. As explained below, the parameter $r_{\tilde{\pi}}$, which controls the horizontal shift of the propensity score function, turns out to be more important than the degree of heteroskedasticity in determining the maximum efficiency gain, as measured by the ratio $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*) \overset{(3.3.11)}{\geq} 1$, that this simulation design can deliver.

Figure 3.A.1 plots $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*)$ as a function of the propensity score shift and the heteroskedasticity parameter.[54] It can be seen from Figure 3.A.1 that the shift of the propensity score function determines how many values of Y^* are lost in the sample. The percentage of non-missing observations as a function of $r_{\tilde{\pi}}$ is shown in black. If $r_{\tilde{\pi}}$ is too large or too small, then $\tilde{\pi}$ becomes almost constant on the support of X, which resembles missing completely at random instead of MAR. In the simulations $r_{\tilde{\pi}} = 0.1$, which yields a retention rate of $\approx 42\%$ (i. e. in $\approx 58\%$ of observations the outcome Y^* is missing). The degree of heteroskedasticity does not appear to have a major impact on the efficiency gains, which is not surprising because $\text{l.b.}(\gamma^*)$ and $\text{l.b.}(\gamma^*)\big|_{\text{VS}}$ are both robust to the form of the skedastic function. Indeed, comparing $\nu = 0$ (conditional

[53] Designs where all regressors and instruments are discrete are not uncommon in microeconometric applications. As in Robins et al. (1994, Section 2.5), efficiency gains in these designs are much more apparent because no smoothing is required to implement the efficient estimator.

[54] In both designs, the ratio $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*)$ was obtained by numerical integration on the simplified expressions given in Appendix 3.D.

homoskedasticity) with $\nu = 2$, Figure 3.A.1 reveals that the maximum efficiency gains are roughly the same ($\approx 42\%$). Therefore, for both Design 1 and Design 2 (described next), we generate data and report simulation results only for heteroskedastic errors since that is the empirically relevant case.

Design 2

In this design, $p \stackrel{\text{def}}{=} \Pr(X = 1) = 0.6$. Consequently,

$$(Z, X) = \begin{cases} (0,0) & \text{w.p. } \Phi(-\zeta_0/\sigma_V)(1-p) = 0.200 \\ (0,1) & \text{w.p. } \Phi\big(-(\zeta_0+\zeta_1)/\sigma_V\big)p = 0.144 \\ (1,0) & \text{w.p. } \Phi(\zeta_0/\sigma_V)(1-p) = 0.200 \\ (1,1) & \text{w.p. } \Phi\big((\zeta_0+\zeta_1)/\sigma_V\big)p = 0.456. \end{cases}$$

The non-missingness indicator D is drawn from a Bernoulli distribution with success probability $\tilde{\pi}(X) \stackrel{\text{def}}{=} 0.9X + 0.25(1-X)$, which yields an average retention rate of $\mathbb{E}D = 64\%$. With this choice of $\tilde{\pi}$, observations corresponding to $X = 0$ are more likely to be missing than observations corresponding to $X = 1$. The skedastic function $\sigma^2(X) \stackrel{\text{def}}{=} X + 16(1-X)$ creates higher dispersion, hence, more uncertainty, when there is less data, which strengthens the case for using the efficient estimator. The maximum efficiency gain $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*)$ that Design 2 can deliver is $\approx 31\%$. Compared to Design 1, endogeneity of Z is even more of a problem in Design 2 because the relative bias of the slope coefficient when Y is regressed on Z in the validation sample is $\approx 179\%$ (averaged across the simulations). In this design, X is not a weak instrument because $\text{corr}(Z, X) \approx 0.27$ and the average first-stage F statistic is ≈ 16.3 when $n = 500$.

3.4.3 Implementation

Code for the simulation experiment is written in `R`[55], and the SEL estimator $\hat{\theta} \stackrel{\text{def}}{=} (\hat{\alpha}, \hat{\gamma})$ is implemented by maximising the SEL objective function with $\mathbb{T}_{i,n} \stackrel{\text{def}}{=} 1$. Similarly, $\hat{\theta}_{\text{VS}} \stackrel{\text{def}}{=} (\hat{\alpha}_{\text{VS}}, \hat{\gamma}_{\text{VS}})$ is implemented by maximizing the SEL objective function with $\mathbb{T}_{i,n} \stackrel{\text{def}}{=} 1$ and $\hat{\rho} \stackrel{\text{def}}{=} Dg$, where $g \stackrel{\text{def}}{=} Y^* - \alpha^* - \gamma^* Z$. The LR confidence region for the slope coefficient is obtained by treating the intercept as a nuisance parameter. Specifically, let $\hat{\alpha}(\gamma) \stackrel{\text{def}}{=} \text{argmax}_{\alpha \in \mathbb{R}} \text{SEL}_{\mathbb{T}}(\alpha, \gamma)$, and denote by

$$\text{LR}^{\text{P}}(\gamma) \stackrel{\text{def}}{=} 2\big[\text{SEL}_{\mathbb{T}}(\hat{\alpha}, \hat{\gamma}) - \text{SEL}_{\mathbb{T}}\big(\hat{\alpha}(\gamma), \gamma\big)\big] \tag{3.4.4}$$

the profile LR statistic obtained by concentrating out the intercept. Then, the lower level set $\{\gamma \in \mathbb{R} : \text{LR}^{\text{P}}(\gamma) \leq k_\tau\}$ is a $(1-\tau)100\%$ confidence region for γ^*. Whether this confidence region is an interval or not depends on the shape of $\gamma \mapsto \text{LR}^{\text{P}}(\gamma)$. If $\gamma \mapsto \text{LR}^{\text{P}}(\gamma)$ is quasi-convex, which appears to be the case in our simulation study because in both designs $\gamma \mapsto \text{SEL}_{\mathbb{T}}\big(\hat{\alpha}(\gamma), \gamma\big)$ seems close to being concave for the sample

[55] The complete code is available from GitHub as an `R` package with estimation and simulation routines at https://github.com/Fifis/smoothemplik.

sizes we consider (Figures 3.4.4 and 3.4.6), then the confidence region is an interval.[56] The endpoints of the LR confidence interval are obtained by numerically finding the roots of the equation $\mathrm{LR}^{\mathrm{P}}(\gamma) = k_\tau$ using Brent's method (the initialising points for the root-finding algorithm are chosen to be the endpoints of the Wald confidence interval). The same approach, mutatis mutandis, is used to obtain the LR confidence interval based on $\hat{\theta}_{\mathrm{VS}}$.

Design 1

Here, w_{ij} is constructed using Gaussian kernels, and $\hat{\tilde{\pi}}_c, \hat{\mu}_d$ are Nadaraya-Watson estimators with bandwidths c, d and Gaussian kernels. Before constructing $\hat{\tilde{\pi}}_{c_n}$ and $\hat{\mu}_{d_n}$, an injective transformation is applied to map distinct elements of $(Z_1, \ldots, Z_n)$ and $(X_1, \ldots, X_n)$ into the interval $(0, 1)$ such that the transformed observations become more equispaced and do not fall into the boundary region. This procedure, motivated by the discussion in Hall (1990, Section 3), is helpful in dealing with the bandwidth and edge-effect issues; e. g. equispacing the observations is a simple device for improving the performance of kernel estimators of conditional expectation functions because the bandwidth does not have to be adaptive if the observations on the conditioning variables are relatively equispaced. It is described in detail in Appendix 3.A as it may be useful to other applied researchers.

To the best of our knowledge, how to choose an optimal data-driven bandwidth when smoothing the empirical likelihood remains an open problem. Consequently, we acted as the oracle to choose the optimal b_n by repeating the simulation experiment on a grid of b_n and picking the bandwidth that minimised the average (across the simulation replications) RMSE of the estimator of γ^*. In particular, since b_n is the only bandwidth required to smooth the empirical likelihood for implementing $\hat{\gamma}_{\mathrm{VS}}$, for every sample size, we estimated $\hat{\gamma}_{\mathrm{VS}}$ on a coarse grid of bandwidths, and the oracle SEL bandwidth b_n^* was chosen to minimise the RMSE of $\hat{\gamma}_{\mathrm{VS}}$. The bandwidth b_n^* was also used to implement the efficient estimator $\hat{\gamma}$. With this optimal b_n^*, we chose (c_n^*, d_n^*) via least-squares cross-validation on each individual simulated data set to implement $\tilde{\pi}_c$ and μ_d. The oracle bandwidth b_n^*, and the median of the cross-validated bandwidths (c_n^*, d_n^*), are reported in Table 3.4.1, which contains the summary statistics for the estimated slope coefficients $\hat{\gamma}$ and $\hat{\gamma}_{\mathrm{VS}}$ averaged across the simulations. The manner in which (b_n^*, c_n^*, d_n^*) were chosen illustrates the following points: (i) Substantial efficiency gains are possible if the bandwidths are chosen appropriately; (ii) the gains in efficiency are not too sensitive to the choice of bandwidth; (iii) the bandwidths for estimating the propensity score $\tilde{\pi}$ and the function μ required for non-parametric imputation can be chosen by cross-validation.[57]

[56] Quasi-convexity of $\gamma \mapsto \mathrm{LR}^{\mathrm{P}}(\gamma)$ implies that its lower level sets are convex. Since convex sets are connected, and the only connected sets in $\mathbb{R}$ are intervals, it follows that if $\gamma \mapsto \mathrm{LR}^{\mathrm{P}}(\gamma)$ is quasi-convex, then its lower level sets are intervals.

[57] In a separate set of simulations, we also acted as the oracle for choosing (c_n, d_n) along with b_n. The efficiency gains in these simulations were marginally higher, e. g. 6% instead of 1% for $n = 500$, and 49% instead of 44% for $n = 2000$. However, we only report the results for the cross-validated bandwidths (c_n, d_n).

Design 2

Discreteness of the conditioning variable exactly identifies θ^* because

$$X \in \{0,1\} \implies \mathbb{E}[Y^* - \alpha^* - \gamma^* Z \mid X] = 0 \text{ w.p.1} \iff \mathbb{E}[Y^* - \alpha^* - \gamma^* Z]\tilde{X} = 0,$$

where $\tilde{X} \stackrel{\text{def}}{=} [\,{}^{1}_{X}\,]$. Hence, as shown in Appendix 3.D.2, if $\hat{\theta}$ solves $\sum_{j=1}^{n} \tilde{X}_j \hat{\rho}(\mathcal{A}_j, \hat{\theta}) = 0$ then it also maximises the SEL objective function with weights $w_{ij} \stackrel{\text{def}}{=} \mathbb{1}(X_i = X_j)/\sum_{k=1}^{n} \mathbb{1}(X_i = X_k)$. The same argument reveals that

$$\hat{\theta}_{\text{VS}} = \Big(\sum_{j=1}^{n} D_j \tilde{X}_j \tilde{Z}_j'\Big)^{-1} \sum_{j=1}^{n} D_j \tilde{X}_j Y_j$$

is the IV estimator obtained using the validation sample.

3.4.4 Results and discussion

We now describe the main findings of our simulation study, which follow our theoretical results fairly closely.

Design 1

The distribution of the estimators appears to be centred around the true value and is close to being normal (Figure 3.4.1). Since the mean and median biases are close to zero (Table 3.4.1), the efficiency gains (whether measured by the ratio of the Monte Carlo variances or the ratio of the Monte Carlo mean squared errors) range from about 1.3% (when $n = 500$) to about 45% (when $n = 4000$). In comparison, as noted in Section 3.4.2, the maximum efficiency gain the simulation design can deliver is about 42%. Figures 3.4.2 and 3.4.3 show that the RMSE of $\hat{\gamma}$ is relatively insensitive to the bandwidths b_n and (c_n, d_n) over a large enough interval.

Table 3.4.2 contains the coverage probabilities for LR confidence intervals and their median lengths (when the intervals are bounded) in the Monte Carlo replications. This table emphasises the following key findings. Firstly, for small sample sizes, the LR confidence intervals can be unbounded in one direction (Figure 3.4.4). E. g. the last column of Table 3.4.2 shows that when $n = 500$ and nominal coverage probability is 90%, the LR confidence intervals based on $\hat{\gamma}_{\text{VS}}$ are unbounded in 6.1% of the Monte Carlo replications, and those based on $\hat{\gamma}$, only in 0.4%. For samples of size 1000 or more, the fraction of unbounded confidence intervals was less than 0.1% for all confidence levels, whilst 0.9% of the intervals (with nominal coverage $= 90\%$) based on $\hat{\gamma}_{\text{VS}}$ were unbounded. Secondly, although their coverage probabilities are close to nominal, the LR confidence intervals based on $\hat{\gamma}$ are much shorter than those based on $\hat{\gamma}_{\text{VS}}$. The difference in the lengths of the confidence intervals is clear evidence of the efficiency gains from $\hat{\gamma}$. In large samples, the ratio of their lengths is close to the square root of relative efficiency gains, as it should be, and in small samples, the gains are even larger.

Design 2

The simulation results for Design 2 are summarised in Table 3.4.3. The increase in $\text{MSE}(\hat{\gamma}_{\text{VS}})$ compared to the $\text{MSE}(\hat{\gamma})$, i. e. $[\text{MSE}(\hat{\gamma}_{\text{VS}}) - \text{MSE}(\hat{\gamma})]/\,\text{MSE}(\hat{\gamma})$, can be very large

for small sample sizes, e. g. 533% when $n = 500$. This, however, is a sample size effect reflecting how $\mathbb{E}\tilde{X}\tilde{Z}'$, required in the implementation of $\hat{\theta}_{\mathrm{VS}}$, is estimated. Indeed, in simulation results not reported here, we replaced $\hat{\theta}_{\mathrm{VS}}$ with $(\sum_{j=1}^{n} \tilde{X}_j \tilde{Z}_j')^{-1} \sum_{j=1}^{n} D_j \tilde{X}_j Y_j$, which estimates $\mathbb{E}\tilde{X}\tilde{Z}'$ using the entire sample (because Z and X are never missing), and found that this led to significant improvement in the performance of $\hat{\gamma}_{\mathrm{VS}}$; namely, its average bias (resp. standard deviation) reduced by more than $1/4$th (resp. $1/2$) when $n = 500$. The efficiency gains stabilise as the sample size increases. For $n = 4000$, they are approximately 39%, which is close to the maximum that Design 2 can deliver. The smoothed densities of $\hat{\gamma} - \gamma^*$ and $\hat{\gamma}_{\mathrm{VS}} - \gamma^*$ are in Figure 3.4.5. Both estimators appear to be Gaussian, with a larger dispersion for $\hat{\gamma}_{\mathrm{VS}}$ as expected. In small samples, the efficiency gains for Design 2 are higher than those for Design 1 because, unlike Design 1, no non-parametric smoothing is required in Design 2.

Table 3.4.4 contains the coverage probabilities for LR confidence intervals and their lengths (when the intervals are bounded), averaged across the Monte Carlo replications. As with Design 1, we find that: (i) For small sample sizes, the LR confidence intervals based on $\hat{\gamma}_{\mathrm{VS}}$ can be unbounded in one direction (Figure 3.4.6). E. g. the last column of Table 3.4.4 reveals that when $n = 500$ and nominal coverage probability is 90%, the LR confidence intervals based on $\hat{\gamma}_{\mathrm{VS}}$ are unbounded in 3.1% of the Monte Carlo replications. In contrast, the LR confidence intervals based on $\hat{\gamma}$ are bounded even when $n = 500$ (except in one simulation sample when nominal coverage probability is 99%). (ii) The LR confidence intervals based on $\hat{\gamma}$ are much shorter than those based on $\hat{\gamma}_{\mathrm{VS}}$. Moreover, the empirical coverage probabilities are very close to nominal since both estimators are empirical-likelihood-based. For small sample sizes, the high accuracy of the empirical coverage probability in Design 2 is due to the absence of any non-parametric smoothing, whereas the slightly more conservative behaviour of confidence intervals in Design 1 is likely caused by non-parametric smoothing and the fact that the estimation-optimal bandwidths used to implement the confidence intervals need not be testing-optimal.

Acknowledgements

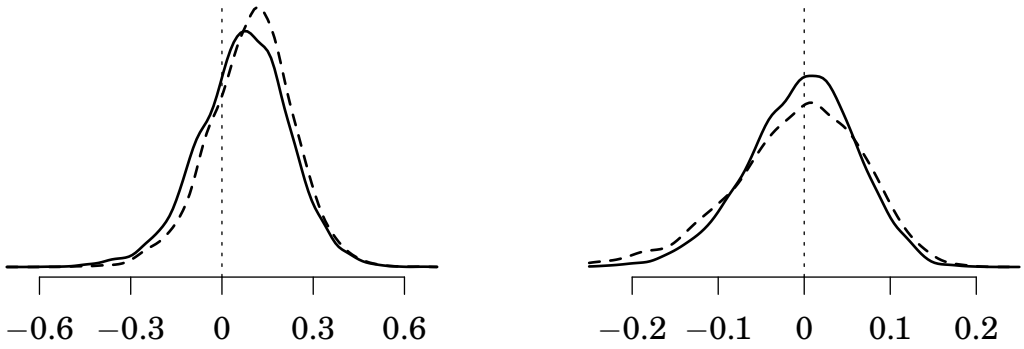
Andreï V. Kostyrka gratefully acknowledges financial support from FNR-Luxembourg through a PRIDE grant for the Migration and Labour (MINLAB) doctoral training unit. The simulation experiments reported in this paper were carried out using the HPC facilities of the University of Luxembourg (Varrette et al., 2014, https://hpc.uni.lu).

Table 3.4.1: Simulation summary for the estimated γ^* in Design 1.

n	Est.	b_n^*	c_n^*	d_n^*	Med. Bias	Mean Bias	Std. Dev.	$\frac{\text{Med. AD}(\cdot)}{\text{Med. AD}(\hat{\gamma})}$	$\frac{\text{MeanAD}(\cdot)}{\text{MeanAD}(\hat{\gamma})}$	$\frac{\text{Var}(\cdot)}{\text{Var}(\hat{\gamma})}$	$\frac{\text{MSE}(\cdot)}{\text{MSE}(\hat{\gamma})}$
500	$\hat{\gamma}$	0.150	0.144	0.321	0.0871	0.0792	0.1473	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.150	–	–	0.1040	0.0965	0.1379	1.0364	1.0138	0.8766	1.0130
1000	$\hat{\gamma}$	0.114	0.121	0.258	0.0198	0.0100	0.1269	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.114	–	–	0.0249	0.0112	0.1355	1.0646	1.0652	1.1406	1.1412
2000	$\hat{\gamma}$	0.086	0.102	0.220	−0.0024	−0.0092	0.0967	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.086	–	–	−0.0014	−0.0168	0.1155	1.1301	1.1710	1.4277	1.4449
4000	$\hat{\gamma}$	0.065	0.086	0.189	−0.0003	−0.0041	0.0642	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.065	–	–	−0.0009	−0.0085	0.0771	1.1567	1.1831	1.4415	1.4529

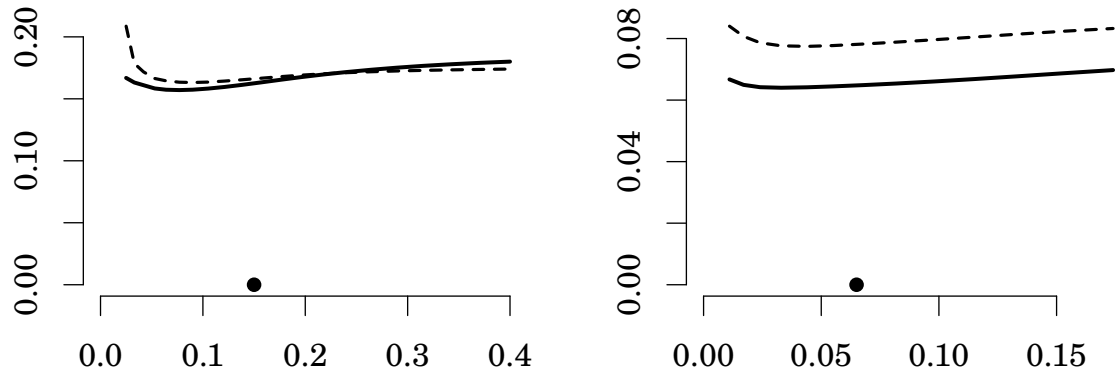
The reported c_n^*, d_n^* are the medians (across all simulations) of the bandwidths chosen via cross-validation. A '–' indicates that $\hat{\gamma}_{\text{VS}}$ does not depend on c_n^*, d_n^*. AD is short for Absolute Deviation.

Figure 3.4.1: Smoothed density of $\hat{\gamma} - \gamma^*$ (solid) and $\hat{\gamma}_{\text{VS}} - \gamma^*$ (dashed) in Design 1.



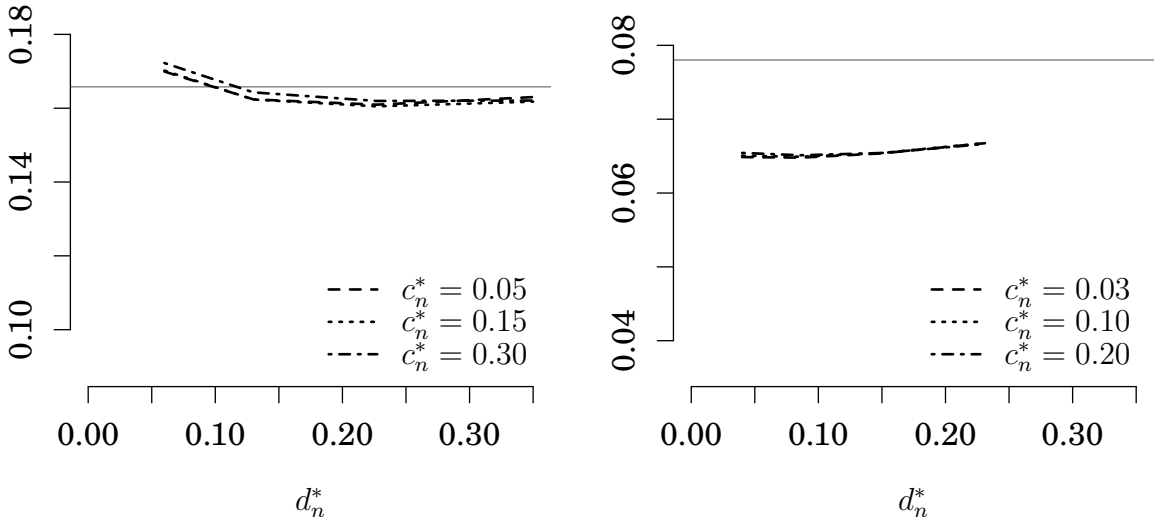
The left panel is for $n = 500$, and the right panel for $n = 4000$.

Figure 3.4.2: RMSE of $\hat{\gamma}$ (solid) and $\hat{\gamma}_{\mathrm{VS}}$ (dashed) as a function of b_n in Design 1.



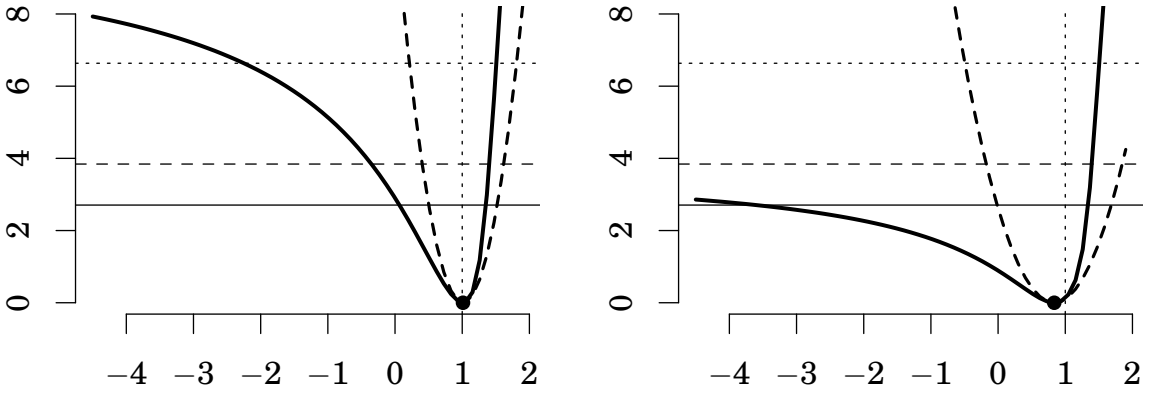
The left panel is for $n = 500$, and the right panel for $n = 4000$. The black dot is b_n^*.

Figure 3.4.3: RMSE$(\hat{\gamma})$ as a function of (c_n^*, d_n^*) in Design 1.



The left panel is for $n = 500$, and the right panel for $n = 4000$. The horizontal line is RMSE$(\hat{\gamma}_{\mathrm{VS}})$.

Figure 3.4.4: Shape of $\gamma \mapsto \mathrm{LR}^{\mathrm{p}}(\gamma)$ in Design 1.



In the left plot, the solid curve is the LR statistic defined in (3.4.4), whereas the dashed curve is the Wald statistic $\mathrm{W}(\gamma) \stackrel{\mathrm{def}}{=} |(\hat{\gamma} - \gamma)/\operatorname{se}(\hat{\gamma})|^2$. The right plot shows the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ (solid) and the corresponding Wald statistic (dashed). The vertical line is the location of the true $\gamma(= 1)$, whereas the black point shows the location of $\hat{\gamma}$ in the left plot and $\hat{\gamma}_{\mathrm{VS}}$ in the right plot. The horizontal lines are the $\{.9, .95, .99\}$-quantiles of a χ^2_1 random variable. The above plots were obtained using one simulated dataset with $n = 500$ (220 observations in the validation sample). In this dataset, the 95%—hence, the 99%—LR confidence interval based on $\hat{\gamma}_{\mathrm{VS}}$ is unbounded from the left. [Numerical evaluations reveal that the line $y = 3.81$ is a horizontal asymptote to the graph of the $\hat{\gamma}_{\mathrm{VS}}$-based LR statistic at $-\infty$. Therefore, the left branch of the graph of the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ never exceeds the .95 quantile (3.84)—hence, the .99 quantile (6.63)—of a χ^2_1 random variable.]

Table 3.4.2: LR confidence intervals for γ^* in Design 1.

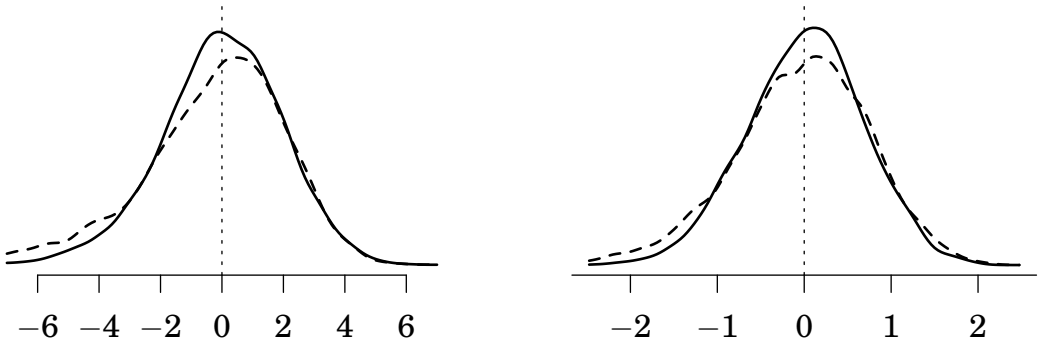
		Coverage Probability			
n	Estimator	Nominal	Empirical	Median length	% bounded
500	$\hat{\gamma}$	0.90	0.916	0.646	100
		0.95	0.968	0.828	100
		0.99	0.995	1.354	99.6
	$\hat{\gamma}_{\text{VS}}$	0.90	0.920	0.698	100
		0.95	0.970	0.943	100.0
		0.99	0.995	1.820	93.9
1000	$\hat{\gamma}$	0.90	0.933	0.489	100
		0.95	0.971	0.605	100
		0.99	0.996	0.879	100.0
	$\hat{\gamma}_{\text{VS}}$	0.90	0.942	0.584	100
		0.95	0.973	0.753	100.0
		0.99	0.996	1.246	99.1
2000	$\hat{\gamma}$	0.90	0.914	0.331	100
		0.95	0.958	0.401	100
		0.99	0.993	0.551	100
	$\hat{\gamma}_{\text{VS}}$	0.90	0.911	0.398	100
		0.95	0.957	0.491	100
		0.99	0.994	0.710	100
4000	$\hat{\gamma}$	0.90	0.913	0.219	100
		0.95	0.957	0.264	100
		0.99	0.993	0.353	100
	$\hat{\gamma}_{\text{VS}}$	0.90	0.912	0.258	100
		0.95	0.955	0.313	100
		0.99	0.994	0.429	100

A '100.0' in the last column (due to round-off rules) indicates that there are 99.95% or more bounded intervals.

Table 3.4.3: Simulation summary for the estimated γ^* in Design 2.

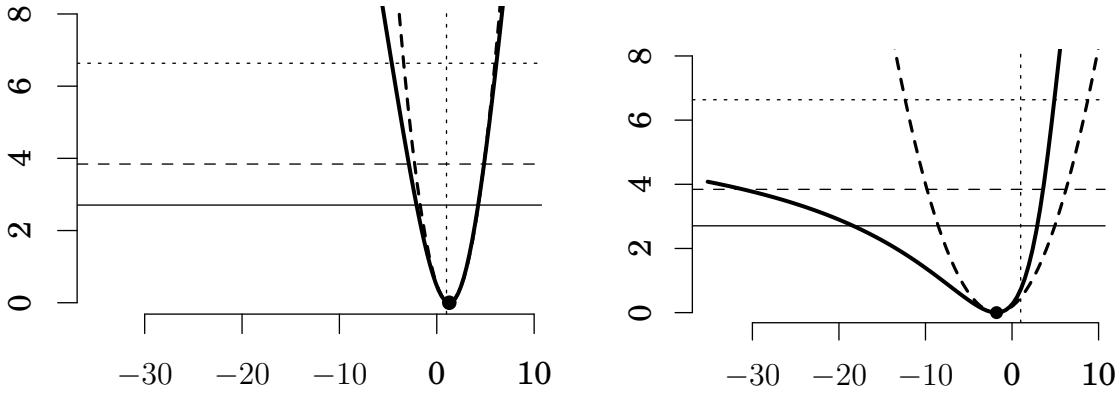
n	Estimator	Median Bias	Mean Bias	Std. Dev.	$\frac{\text{Med. AD}(\cdot)}{\text{Med. AD}(\hat{\gamma})}$	$\frac{\text{MeanAD}(\cdot)}{\text{MeanAD}(\hat{\gamma})}$	$\frac{\text{Var}(\cdot)}{\text{Var}(\hat{\gamma})}$	$\frac{\text{MSE}(\cdot)}{\text{MSE}(\hat{\gamma})}$
500	$\hat{\gamma}$	0.0418	-0.0316	2.0204	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	-0.0252	-0.6041	5.0498	1.1716	1.4167	6.2470	6.3349
1000	$\hat{\gamma}$	0.0269	-0.0266	1.3979	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.0042	-0.2288	1.8047	1.1067	1.1989	1.6668	1.6930
2000	$\hat{\gamma}$	0.0407	0.0193	0.9634	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.0150	-0.0808	1.1751	1.1572	1.1845	1.4877	1.4942
4000	$\hat{\gamma}$	0.0338	0.0136	0.6693	1.0000	1.0000	1.0000	1.0000
	$\hat{\gamma}_{\text{VS}}$	0.0224	-0.0356	0.7884	1.1519	1.1608	1.3879	1.3901

Figure 3.4.5: Smoothed density of $\hat{\gamma} - \gamma^*$ (solid) and $\hat{\gamma}_{\text{VS}} - \gamma^*$ (dashed) in Design 2.



The left panel is for $n = 500$, and the right panel for $n = 4000$.

Figure 3.4.6: Shape of $\gamma \mapsto \mathrm{LR}^{\mathrm{p}}(\gamma)$ in Design 2.



In the left plot, the solid curve is the LR statistic defined in (3.4.4), whereas the dashed curve is the Wald statistic $\mathrm{W}(\gamma) \stackrel{\mathrm{def}}{=} |(\hat{\gamma} - \gamma)/\operatorname{se}(\hat{\gamma})|^2$. The right plot shows the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ (solid) and the corresponding Wald statistic (dashed). The vertical line is the location of the true $\gamma(= 1)$, whereas the black point shows the location of $\hat{\gamma}$ in the left plot and $\hat{\gamma}_{\mathrm{VS}}$ in the right plot. The horizontal lines are the $\{.9, .95, .99\}$-quantiles of a χ^2_1 random variable. The above plots were obtained using one simulated dataset with $n = 500$ (320 observations in the validation sample). In this dataset, the 99% LR confidence interval based on $\hat{\gamma}_{\mathrm{VS}}$ is unbounded from the left. [Numerical evaluations reveal that the line $y = 6.47$ is a horizontal asymptote to the graph of the $\hat{\gamma}_{\mathrm{VS}}$-based LR statistic at $-\infty$. Therefore, the left branch of the graph of the LR statistic based on $\hat{\gamma}_{\mathrm{VS}}$ never exceeds the .99 quantile (6.63) of a χ^2_1 random variable.]

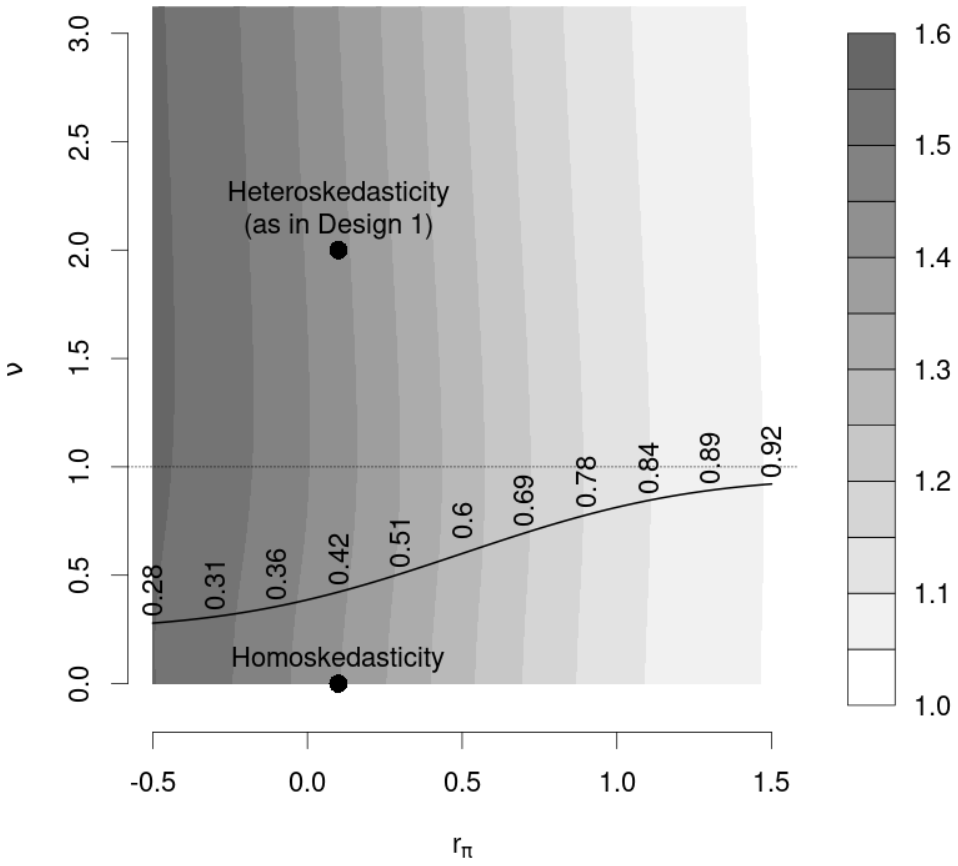
Table 3.4.4: LR confidence intervals for γ^* in Design 2.

		Coverage Probability			
n	Estimator	Nominal	Empirical	Median length	% bounded
500	$\hat{\gamma}$	.90	.905	6.66	100
		.95	.952	8.17	100
		.99	.991	11.51	100.0
	$\hat{\gamma}_{\text{VS}}$	.90	.897	8.43	96.9
		.95	.949	10.77	94.1
		.99	.990	16.67	84.2
1000	$\hat{\gamma}$	.90	.903	4.59	100
		.95	.953	5.54	100
		.99	.993	7.54	100
	$\hat{\gamma}_{\text{VS}}$	.90	.900	5.53	100.0
		.95	.952	6.83	99.8
		.99	.992	9.91	99.2
2000	$\hat{\gamma}$	.90	.898	3.19	100
		.95	.952	3.83	100
		.99	.990	5.12	100
	$\hat{\gamma}_{\text{VS}}$	.90	.897	3.73	100
		.95	.947	4.53	100
		.99	.991	6.23	100
4000	$\hat{\gamma}$	.90	.904	2.24	100
		.95	.957	2.68	100
		.99	.991	3.55	100
	$\hat{\gamma}_{\text{VS}}$	.90	.903	2.59	100
		.95	.948	3.11	100
		.99	.991	4.18	100

Appendix

3.A Additional figures and implementation details

Figure 3.A.1: Heat map of $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*)$ as a function the propensity score shift ($r_{\tilde{\pi}}$) and the degree of heteroskedasticity (ν) in Design 1.



The darker the shade, the larger the efficiency gain $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*)$. The solid line and the numbers above it show the proportion of non-missing observations.

Additional implementation details for Design 1

Motivated by the discussion in Hall (1990, Section 3), before estimating $\tilde{\pi}$ and μ we apply an injective transformation to $(Z_1, \ldots, Z_n)$ and $(X_1, \ldots, X_n)$ to map their distinct elements into $(0, 1)$ in order to simplify the problem of bandwidth choice and deal with edge effects in kernel estimators. We have found that doing so improves the

performance of our kernel estimators. We now describe this procedure in detail as it may be useful to other applied researchers as well.

Let the random variable A denote Z or X (if Z or X are vectors, the procedure is applied coordinate-wise). There are no missing observations in $\mathscr{A} \stackrel{\text{def}}{=} (A_1, \ldots, A_n)$ because Z and X are observed for each i. Let $M_n \stackrel{\text{def}}{=} \sum_{i=1}^{n} D_i$ be the size of the validation sample. Since the validation sample only contains those i for which $D_i = 1$, we have $\mathscr{A} = \mathscr{V} \cup \mathscr{N}$, where $\mathscr{V}$ is the ordered array (keeping ties preserved) of observations in the validation sample, and $\mathscr{N}$ is the ordered array (keeping ties preserved) of observations not in the validation sample.[58] Let $A^{\text{VS}}_{(1)} \leq \ldots \leq A^{\text{VS}}_{(M_n)}$ denote the ordered observations in $\mathscr{V}$, and define

$$\begin{aligned}
\mathscr{N}_1 &\stackrel{\text{def}}{=} \text{ordered array of elements of } \mathscr{N} \text{ in } (-\infty, A^{\text{VS}}_{(1)}) \\
\mathscr{N}_j &\stackrel{\text{def}}{=} \text{ordered array of elements of } \mathscr{N} \text{ in } (A^{\text{VS}}_{(j-1)}, A^{\text{VS}}_{(j)}) \qquad (j = 2, \ldots, M_n) \\
\mathscr{N}_{M_n+1} &\stackrel{\text{def}}{=} \text{ordered array of elements of } \mathscr{N} \text{ in } (A^{\text{VS}}_{(M_n)}, \infty).
\end{aligned}$$

If $(A^{\text{VS}}_{(j-1)}, A^{\text{VS}}_{(j)})$ is empty (e. g. when there are ties in $\mathscr{V}$), then $\mathscr{N}_j$ is the empty array. Let $\hat{F}_{\mathscr{V}}(a) \stackrel{\text{def}}{=} M_n^{-1} \sum_{j=1}^{M_n} \mathbb{1}(A^{\text{VS}}_{(j)} \leq a)$, $a \in \mathbb{R}$, be the empirical cumulative distribution function (cdf) of the observations in $\mathscr{V}$, and define

$$\begin{aligned}
\mathscr{T}_1 &\stackrel{\text{def}}{=} \text{set of tick marks in an equispaced grid on } \left(0, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - \frac{0.5}{M_n}\right) \\
&\qquad \text{with as many ticks as the number of distinct elements in } \mathscr{N}_1, \\
\mathscr{T}_j &\stackrel{\text{def}}{=} \text{set of tick marks in an equispaced grid on } \left(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(j-1)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(j)}) - \frac{0.5}{M_n}\right) \\
&\qquad \text{with as many ticks as the number of distinct elements in } \mathscr{N}_j,\ j = 2, \ldots, M_n, \\
\mathscr{T}_{M_n+1} &\stackrel{\text{def}}{=} \text{set of tick marks in an equispaced grid on } \left(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(M_n)}) - \frac{0.5}{M_n}, 1\right) \\
&\qquad \text{with as many ticks as the number of distinct elements in } \mathscr{N}_{M_n}.
\end{aligned}$$

Note that $\mathscr{T}_j$ is empty if $\mathscr{N}_j$ is the empty array.

Now, for $i = 1, \ldots, n$, map $A_i \to (0, 1)$ as follows:

$$\Psi_n(A_i) \stackrel{\text{def}}{=} \begin{cases} \hat{F}_{\mathscr{V}}(A_i) - \dfrac{0.5}{M_n} & \text{if } A_i \in \{A^{\text{VS}}_{(1)}, \ldots, A^{\text{VS}}_{(M_n)}\}, \\ \text{tick in } \mathscr{T}_j \text{, repeated as many times as the} \\ \text{multiplicity of } A_i \in \mathscr{N}_j \text{, such that the order} & \text{if } A_i \in \mathscr{N}_j \ (j = 1, \ldots, M_n + 1). \\ \text{in which } A_i \text{ appears in } \mathscr{N}_j \text{ is preserved} \end{cases}$$

In words, Ψ_n makes distinct elements of $(A_1, \ldots, A_n)$ equally spaced in the validation and non-validation subsamples by placing observations from the validation sample $0.5/M_n$ units below their values under $\hat{F}_{\mathscr{V}}$, whereas observations not in the validation sample are placed equally apart in $0 < \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - 0.5/M_n \leq \ldots \leq \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(M_n)}) - 0.5/M_n < 1$, taking ties into account. Spacing the observations equally in each subsample mitigates the problem of bandwidth selection in low-density regions of the conditioning variable, and ensuring that the observations stay away from the boundary improves the small-sample properties of the kernel estimators. As Ψ_n is injective by construction, the

[58] Note that $\mathscr{V}$ and $\mathscr{N}$ may have elements in common (corresponding to different i).

information set used to estimate the conditional expectations remains unchanged.

The following numerical example illustrates how Ψ_n works.

Example 3.A.1. Let $n = 11$, $M_n = 5$, $\mathscr{V} = (1, 1, 3, 4, 6)$ and $\mathscr{N} = (0, 2, 2, 5.9, 7, 8)$. In this dataset, $A^{\text{VS}}_{(1)} = 1$, $A^{\text{VS}}_{(2)} = 1$, $A^{\text{VS}}_{(3)} = 3$, $A^{\text{VS}}_{(4)} = 4$, $A^{\text{VS}}_{(5)} = 6$. Hence, $\mathscr{N}_1 = (0)$, $\mathscr{N}_2 = (\varnothing)$, $\mathscr{N}_3 = (2, 2) = (2^{\text{multiplicity}=2})$, $\mathscr{N}_4 = (\varnothing)$, $\mathscr{N}_5 = (5.9)$, and $\mathscr{N}_6 = (7, 8)$. Next, as $\hat{F}_{\mathscr{V}}(a) = [2\mathbb{1}(1 \leq a) + \mathbb{1}(3 \leq a) + \mathbb{1}(4 \leq a) + \mathbb{1}(6 \leq a)]/5$, we have

$$
\begin{aligned}
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) = \frac{2}{5} &\implies \Big(0, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - \frac{0.5}{M_n}\Big) = (0, 0.3) \overset{\mathscr{N}_1=(0)}{\Longrightarrow} \mathscr{T}_1 = \{0.15\} \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(2)}) = \frac{2}{5} &\implies \Big(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(1)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(2)}) - \frac{0.5}{M_n}\Big) = (0.3, 0.3) \overset{\mathscr{N}_2=(\varnothing)}{\Longrightarrow} \mathscr{T}_2 = \varnothing \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(3)}) = \frac{3}{5} &\implies \Big(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(2)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(3)}) - \frac{0.5}{M_n}\Big) = (0.3, 0.5) \overset{\mathscr{N}_3=(2^{\text{m-plicity}=2})}{\Longrightarrow} \mathscr{T}_3 = \{0.4\} \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(4)}) = \frac{4}{5} &\implies \Big(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(3)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(4)}) - \frac{0.5}{M_n}\Big) = (0.5, 0.7) \overset{\mathscr{N}_4=(\varnothing)}{\Longrightarrow} \mathscr{T}_4 = \varnothing \\
\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(5)}) = 1 &\implies \Big(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(4)}) - \frac{0.5}{M_n}, \hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(5)}) - \frac{0.5}{M_n}\Big) = (0.7, 0.9) \overset{\mathscr{N}_5=(5.9)}{\Longrightarrow} \mathscr{T}_5 = \{0.8\} \\
&\Big(\hat{F}_{\mathscr{V}}(A^{\text{VS}}_{(5)}) - \frac{0.5}{M_n}, 1\Big) = (0.9, 1) \overset{\mathscr{N}_6=(7,8)}{\Longrightarrow} \mathscr{T}_6 = \{28/30, 29/30\}.
\end{aligned}
$$

Consequently,

$$
\begin{aligned}
\Psi_{11}(\mathscr{V}) &= \Big(\hat{F}_{\mathscr{V}}(1) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(1) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(3) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(4) - \frac{0.5}{5}, \hat{F}_{\mathscr{V}}(6) - \frac{0.5}{5}\Big) \\
&= (0.3, 0.3, 0.5, 0.7, 0.9); \\
\Psi_{11}(\mathscr{N}) &= (\text{ticks in } \mathscr{T}_1, \ldots, \mathscr{T}_6 \text{ preserving the multiplicity and order in } \mathscr{N}_1, \ldots, \mathscr{N}_6) \\
&= (0.15, 0.4, 0.4, 0.8, 28/30, 29/30).
\end{aligned}
$$

The graph of Ψ_{11} is shown in Figure 3.A.2. □

3.B Proofs for Section 3.2

3.B.1 Local identification in conditional moment restriction models

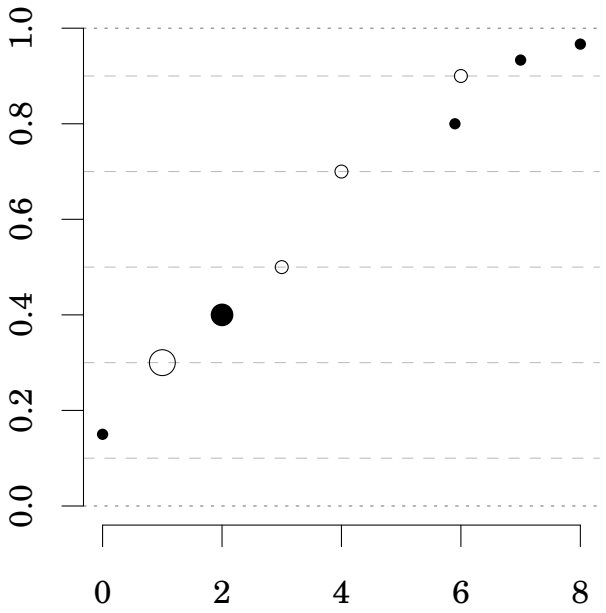
Let $\|\cdot\|$ be the Euclidean norm and $J(X, \theta)_{(\dim g)\times(\dim \theta^*)} \overset{\text{def}}{=} \partial_\theta \mathbb{E}[g(Y^*, Z, X, \theta) \mid X]$. Recall that $J \overset{\text{def}}{=} J(X, \theta^*)$.

Definition 3.B.1 (Linear independence P_X-a.s.)**.** The columns of J are said to be linearly independent P_X-a.s. if, for all $\alpha \in \mathbb{R}^{\dim \theta^*}$, $P_X(J\alpha = 0) = 1 \implies \alpha = 0$.

In this section, we extend Rothenberg (1971) to show that θ^* in (3.1.2) is locally identified if the columns of J are linearly independent P_X-a.s. We begin by defining the notion of observational equivalence for the conditional moment equality $\mathbb{E}[g \mid X] = 0$ P_X-a.s.

Definition 3.B.2 (Observational equivalence)**.** The parameters $\theta^*, \theta^\dagger \in \Theta$ are said to be observationally equivalent if $\mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] = 0 = \mathbb{E}[g(Y^*, Z, X, \theta^\dagger) \mid X]$ P_X-a.s.

Figure 3.A.2: The graph of Ψ_{11}.



The empty circles denote points in $\mathscr{V}$, and the filled circles denote points in $\mathscr{N}$. The larger circles correspond to observations with multiplicity 2.

In other words, θ^* and $\theta^\dagger$ are observationally equivalent if they satisfy the same conditional moment equality. Next, we define what it means for θ^* to be locally identified.

Definition 3.B.3 (Local identification)**.** The parameter $\theta^* \in \Theta$ is said to be locally identified if there exists an open ball centred at θ^*, say $\mathcal{N}^* \subset \Theta$, such that the punctured open ball $\mathcal{N}^* \setminus \{\theta^*\}$ does not contain any element observationally equivalent to θ^*.

We now prove the local identification result stated in Section 3.2. Although Lemma 3.B.1 below looks as if it should be well-known, we have been unable to find it in the literature.[59]

Lemma 3.B.1. *Let $\mathbb{E}[g(Y^*, Z, X, \theta^*) \mid X] = 0$ P_X-a.s. for some $\theta^* \in \Theta$. Assume that (P_X-a.s.):* (a) *$\theta \mapsto J(X, \theta)$ is well-defined on an open (in Θ) ball centred at θ^*, and* (b) *$\theta \mapsto J(X, \theta)$ is continuous at θ^*. If the columns of J are linearly independent P_X-a.s., then θ^* is locally identified.*

Note that for (a) to hold, it is necessary that Θ has a non-empty interior.

Proof of Lemma 3.B.1. Suppose to the contrary that θ^* is not locally identified. Then, by Definition 3.B.3, each punctured open ball centred at θ^* contains at least one element different from θ^* that is observationally equivalent to θ^*. This yields a sequence $(\theta_j)_{j\in\mathbb{N}} \subset \Theta$ such that (i) $\lim_{j\to\infty} \theta_j = \theta^*$, (ii) $\theta_j \neq \theta^*$ for each $j \in \mathbb{N}$, and (iii) θ_j is observationally equivalent to θ^* for each $j \in \mathbb{N}$. Letting $m(X, \theta) \stackrel{\text{def}}{=} \mathbb{E}[g(Y^*, Z, X, \theta) \mid X]$ and $q \stackrel{\text{def}}{=} \dim g$, an element-by-element mean value expansion of $m(X, \theta_j)$ about θ^*

[59] Identification of parameters defined via unconditional moment equalities is discussed in Newey and McFadden (1994, Section 2.2.3).

reveals that

$$m(X,\theta_j) \overset{\text{(a)}}{=} m(X,\theta^*) + \begin{bmatrix} d_1'\big(X,\theta^* + \lambda_1(\theta_j-\theta^*)\big) \\ \vdots \\ d_q'\big(X,\theta^* + \lambda_q(\theta_j-\theta^*)\big) \end{bmatrix} (\theta_j - \theta^*) \qquad P_X\text{-a.s.}, \tag{3.B.1}$$

where d_k' denotes the kth row of J, and each $\lambda_k \in (0,1)$. Hence, by (iii) and Definition 3.B.2,

$$\begin{bmatrix} d_1'\big(X,\theta^* + \lambda_1(\theta_j-\theta^*)\big) \\ \vdots \\ d_q'\big(X,\theta^* + \lambda_q(\theta_j-\theta^*)\big) \end{bmatrix} (\theta_j - \theta^*) = 0 \qquad P_X\text{-a.s.} \qquad (j \in \mathbb{N})$$

By (ii), $r_j \overset{\text{def}}{=} (\theta_j - \theta^*)/\|\theta_j - \theta^*\|$ is well-defined for each j. Hence, we can write the previous displayed equation as

$$\begin{bmatrix} d_1'(X,\theta^* + \lambda_1 r_j\|\theta_j - \theta^*\|) \\ \vdots \\ d_q'(X,\theta^* + \lambda_q r_j\|\theta_j - \theta^*\|) \end{bmatrix} r_j = 0 \qquad P_X\text{-a.s.} \qquad (j \in \mathbb{N})$$

Now, (r_j) is a bounded sequence in $\mathbb{R}^{\dim\theta^*}$ because $\|r_j\| = 1$ for each j. Hence, by the Bolzano-Weierstrass theorem, there exists a subsequence $(s_j) \subset (r_j)$, and $r^* \in \mathbb{R}^{\dim\theta^*}$ with $\|r^*\| = 1$, such that $\lim_{j\to\infty} s_j = r^*$. In particular, since (s_j) is a subsequence of (r_j), we have that

$$\begin{bmatrix} d_1'(X,\theta^* + \lambda_1 s_j\|\theta_j - \theta^*\|) \\ \vdots \\ d_q'(X,\theta^* + \lambda_q s_j\|\theta_j - \theta^*\|) \end{bmatrix} s_j = 0 \qquad P_X\text{-a.s.} \qquad (j \in \mathbb{N})$$

Since each row of $J(X,\theta)$ is continuous at θ^* (P_X-a.s.) if and only if $J(X,\theta)$ is continuous at θ^* (P_X-a.s.), letting $j \to \infty$ in the previous displayed equation, (b) and (i) imply that

$$\begin{bmatrix} d_1'(X,\theta^*) \\ \vdots \\ d_q'(X,\theta^*) \end{bmatrix} r^* = 0 \qquad P_X\text{-a.s.} \iff Jr^* = 0 \qquad P_X\text{-a.s.}$$

But, as $r^* \neq 0$, this contradicts the assumption that the columns of J are linearly independent P_X-a.s. The desired result follows. □

As mentioned in Footnote 46, Lemma 3.B.1 implies the global identification of θ^* whenever g is linear in θ^*.[60] This is easily verified for linear regression models.

Example 3.B.1. In Example 3.1.1, $g \overset{\text{def}}{=} Y^* - \alpha^* - X_{\text{in}}'\beta^* - Z'\gamma^*$. Hence, $\mathbb{E}[g \mid X] = \mathbb{E}[Y^* \mid X] - \alpha^* - X_{\text{in}}'\beta^* - \mathbb{E}[Z' \mid X]\gamma^*$. Consequently, it is straightforward to verify that $\theta^* \overset{\text{def}}{=} (\alpha^*, \beta^*, \gamma^*)$ is globally identified if and only if the columns of $\begin{bmatrix} 1 & X_{\text{in}}' & \mathbb{E}[Z' \mid X] \end{bmatrix}$ are linearly independent P_X-a.s. The connection with Lemma 3.B.1 is apparent because, in this example, $J = -\begin{bmatrix} 1 & X_{\text{in}}' & \mathbb{E}[Z' \mid X] \end{bmatrix}$.

[60] Because the mean value expansion in (3.B.1) is exact when g is linear in θ^*.

3.C Proofs for Section 3.3

The following notation is used throughout this section. For a generic random vector W, the set of real-valued functions of W with finite second moments is denoted by $L_2(W)$, and $L_{2,0}(W) \stackrel{\text{def}}{=} \{\psi \in L_2(W) : \mathbb{E}\psi(W) = 0\}$ is the subset of functions of W whose expectation is zero. If $S \subset L_2(W)$, then $S^\perp$ is the orthogonal complement of S in $L_2(W)$, and $\mathscr{P}_S(a)$ denotes the orthogonal projection of $a \in L_2(W)$ onto S using the inner product $\langle a_1, a_2\rangle \stackrel{\text{def}}{=} \mathbb{E}[a_1 a_2]$. The norm induced by the inner product is $\|a\|_2 \stackrel{\text{def}}{=} \sqrt{\mathbb{E}a^2}$.

Proof that ρ is the residual from the coordinate-wise projection of Dg/π onto $\mathfrak{A}$. Let (3.3.1) hold, which, by the equivalence in (3.3.2), implies that MAR holds. Let $\alpha \in \mathbb{R}^{\dim g}$ be such that $\|\alpha\| = 1$. Then, $\alpha'\rho = D\alpha' g/\pi - \alpha'\mu[D/\pi - 1]$. Now,

$$\begin{aligned}
\mathbb{E}(\alpha'\mu)^2 &\leq \mathbb{E}\|\mu\|^2 && \text{(Cauchy-Schwarz)}\\
&= \mathbb{E}\big\|\mathbb{E}[g \mid Z, X]\big\|^2 && \text{(defn. } \mu\text{)}\\
&= \mathbb{E}\big[(\mathbb{E}[g^{(1)} \mid Z, X])^2 + \ldots + (\mathbb{E}[g^{(\dim g)} \mid Z, X])^2\big]\\
&\leq \mathbb{E}\big[\mathbb{E}[(g^{(1)})^2 \mid Z, X] + \ldots + \mathbb{E}[(g^{(\dim g)})^2 \mid Z, X]\big] && \text{(cond. Jensen)}\\
&= \mathbb{E}[g^{(1)}]^2 + \ldots + \mathbb{E}[g^{(\dim g)}]^2 && \text{(iterated expectations)}\\
&= \mathbb{E}\|g\|^2\\
&< \infty. && \text{(Ass. 3.3.1(ii))}
\end{aligned}$$

Consequently, $\alpha'\mu \in L_2(Z, X) \implies \alpha'\mu[D/\pi - 1] \in \mathfrak{A}$. It remains to show that $\alpha'\rho$ is orthogonal to $\mathfrak{A}$. Begin by observing that

$$\begin{aligned}
\alpha'\rho &\stackrel{(3.3.4)}{=} \frac{D\alpha' g(Y, Z, X, \theta^*)}{\pi(Z, X)} - \alpha'\mu(Z, X, \theta^*)\Big[\frac{D}{\pi(Z, X)} - 1\Big]\\
&= \frac{D\alpha' g}{\pi} - \alpha'\mu\Big[\frac{D}{\pi} - 1\Big]. && (Dg(Y^*, Z, X, \theta^*) \stackrel{(3.1.1)}{=} Dg(Y, Z, X, \theta^*))
\end{aligned}$$

Let $\mathfrak{a} \in \mathfrak{A}$ so that $\mathfrak{a} = a(D/\pi - 1)$ for some $a \in L_2(Z, X)$. Then,

$$\mathbb{E}[\alpha'\rho\mathfrak{a}] = \mathbb{E}\Big[\alpha'\rho a\Big(\frac{D}{\pi} - 1\Big)\Big] = \alpha'\mathbb{E}\Big[\frac{Dg}{\pi}a\Big(\frac{D}{\pi} - 1\Big)\Big] - \alpha'\mathbb{E}\Big[\mu a\Big(\frac{D}{\pi} - 1\Big)^2\Big].$$

Thus, we are done if we can show that $\mathbb{E}\Big[\frac{Dg}{\pi}a\Big(\frac{D}{\pi} - 1\Big)\Big] = \mathbb{E}\Big[a\mu\Big(\frac{D}{\pi} - 1\Big)^2\Big]$. Now,

$$\begin{aligned}
\mathbb{E}\Big[\frac{Dg}{\pi}a\Big(\frac{D}{\pi} - 1\Big) \;\Big|\; Y^*, Z, X\Big] &= ag\mathbb{E}\Big[\frac{D}{\pi}\Big(\frac{D}{\pi} - 1\Big) \;\Big|\; Y^*, Z, X\Big]\\
&= \frac{ag}{\pi}\mathbb{E}\Big[\frac{D}{\pi} - D \;\Big|\; Y^*, Z, X\Big] && (D^2 = D)\\
&\stackrel{\text{MAR}}{=} \frac{ag}{\pi}\mathbb{E}\Big[\frac{D}{\pi} - D \;\Big|\; Z, X\Big] && ((3.3.1) \stackrel{(3.3.2)}{\implies} \text{MAR})\\
&= ag\Big(\frac{1}{\pi} - 1\Big). && (\pi \stackrel{\text{def}}{=} \mathbb{E}[D \mid Z, X])
\end{aligned}$$

Hence, conditioning on (Z, X), we obtain that

$$\mathbb{E}\Big[\frac{Dg}{\pi}a\Big(\frac{D}{\pi} - 1\Big) \;\Big|\; Z, X\Big] = \mathbb{E}\Big[ag\Big(\frac{1}{\pi} - 1\Big) \;\Big|\; Z, X\Big] = a\mu\Big(\frac{1}{\pi} - 1\Big). \qquad (\mu \stackrel{\text{def}}{=} \mathbb{E}[g \mid Z, X])$$

Consequently,

$$\mathbb{E}\Big[\frac{Dg}{\pi}a\Big(\frac{D}{\pi}-1\Big)\Big]=\mathbb{E}\Big[a\mu\Big(\frac{1}{\pi}-1\Big)\Big].$$

Moreover,

$$\mathbb{E}\Big[a\mu\Big(\frac{D}{\pi}-1\Big)^2\;\Big|\;Z,X\Big]\overset{(D^2=D)}{=}a\mu\mathbb{E}\Big[\frac{D}{\pi^2}+1-2\frac{D}{\pi}\;\Big|\;Z,X\Big]\overset{(3.3.1b)}{=}a\mu\Big(\frac{1}{\pi}-1\Big).$$

Therefore,

$$\mathbb{E}\Big[a\mu\Big(\frac{D}{\pi}-1\Big)^2\Big]=\mathbb{E}\Big[a\mu\Big(\frac{1}{\pi}-1\Big)\Big].$$

The desired result follows. □

Proof of (3.3.6). Let (3.3.1) hold, which, by the equivalence in (3.3.2), implies that MAR holds. Observe that $\mathbb{E}[\rho\mid X]\overset{(3.3.4)}{=}\mathbb{E}\Big[\frac{Dg}{\pi}\;\Big|\;X\Big]-\mathbb{E}\Big[\mu\Big(\frac{D}{\pi}-1\Big)\;\Big|\;X\Big]$. But $\mathbb{E}\Big[\frac{Dg}{\pi}\;\Big|\;X\Big]\overset{(3.3.1a)}{=}0$ P_X-a.s., and

$$\mathbb{E}\Big[\mu\Big(\frac{D}{\pi}-1\Big)\;\Big|\;Z,X\Big]\overset{P_{Z,X}\text{-a.s.}}{=}\mu\mathbb{E}\Big[\frac{D}{\pi}-1\;\Big|\;Z,X\Big]\overset{(3.3.1b)}{=}0\implies\mathbb{E}\Big[\mu\Big(\frac{D}{\pi}-1\Big)\;\Big|\;X\Big]\overset{P_X\text{-a.s.}}{=}0.$$

The desired result follows. □

Proof of (3.3.7). Let (3.3.1) hold, which, by the equivalence in (3.3.2), implies that MAR holds. Assume that derivatives with respect to (θ^*,π,μ) can be exchanged with conditional (on Z,X) expectations. Then, since $\pi\overset{\text{def}}{=}\mathbb{E}[D\mid Z,X]$ does not depend on θ^* (Assumption 3.2.2),

$$\begin{aligned}
&\partial_{\theta^*}\mathbb{E}[\rho(\mathcal{A},\theta^*,\pi,\mu)\mid Z,X]\\
&\quad\overset{(3.3.4)}{=}\partial_{\theta^*}\mathbb{E}\Big[\frac{Dg(Y,Z,X,\theta^*)}{\pi(Z,X)}\;\Big|\;Z,X\Big]-\mathbb{E}\Big[\frac{D}{\pi(Z,X)}-1\;\Big|\;Z,X\Big]\partial_{\theta^*}\mu(Z,X,\theta^*)\\
&\quad\overset{(3.3.1b)}{=}\frac{1}{\pi(Z,X)}\partial_{\theta^*}\mathbb{E}[Dg(Y,Z,X,\theta^*)\mid Z,X]\\
&\quad=\frac{1}{\pi(Z,X)}\partial_{\theta^*}\mathbb{E}[Dg(Y^*,Z,X,\theta^*)\mid Z,X]\quad(Dg(Y^*,Z,X,\theta^*)\overset{(3.1.1)}{=}Dg(Y,Z,X,\theta^*))\\
&\quad\overset{\text{MAR}}{=}\frac{1}{\pi(Z,X)}\mathbb{E}[D\mid Z,X]\partial_{\theta^*}\mathbb{E}[g(Y^*,Z,X,\theta^*)\mid Z,X]\qquad((3.3.1)\overset{(3.3.2)}{\implies}\text{MAR})\\
&\quad=\partial_{\theta^*}\mathbb{E}[g(Y^*,Z,X,\theta^*)\mid Z,X].
\end{aligned}$$

Consequently, conditioning on X, and recalling that the tower property of conditional expectations holds almost surely and $J\overset{\text{def}}{=}\partial_\theta\mathbb{E}[g(Y^*,Z,X,\theta^*)\mid X]$, we have that

$$\partial_{\theta^*}\mathbb{E}[\rho(\mathcal{A},\theta^*,\pi,\mu)\mid X]\overset{P_{Z,X}\text{-a.s.}}{=}\partial_{\theta^*}\mathbb{E}[g(Y^*,Z,X,\theta^*)\mid X]=J;$$

i. e. (3.3.7a) holds.

Next,

$$
\begin{aligned}
&\partial_\pi \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid Z, X] \\
&\quad \overset{(3.3.4)}{=} -\mathbb{E}\Big[\frac{Dg(Y, Z, X, \theta^*)}{\pi^2(Z, X)} \;\Big|\; Z, X\Big] + \mu(Z, X, \theta^*)\mathbb{E}\Big[\frac{D}{\pi^2(Z, X)} \;\Big|\; Z, X\Big] \\
&\quad \overset{(3.3.1b)}{=} -\mathbb{E}\Big[\frac{Dg(Y, Z, X, \theta^*)}{\pi^2(Z, X)} \;\Big|\; Z, X\Big] + \frac{\mu(Z, X, \theta^*)}{\pi(Z, X)} \\
&\quad = -\frac{1}{\pi^2(Z, X)}\mathbb{E}\Big[Dg(Y^*, Z, X, \theta^*) \;\Big|\; Z, X\Big] + \frac{\mu(Z, X, \theta^*)}{\pi(Z, X)} \\
&\qquad\qquad\qquad\qquad\qquad\qquad (Dg(Y^*, Z, X, \theta^*) \overset{(3.1.1)}{=} Dg(Y, Z, X, \theta^*)) \\
&\quad \overset{\text{MAR}}{=} -\frac{1}{\pi(Z, X)}\mathbb{E}[g(Y^*, Z, X, \theta^*) \mid Z, X] + \frac{\mu(Z, X, \theta^*)}{\pi(Z, X)} \qquad ((3.3.1) \overset{(3.3.2)}{\Longrightarrow} \text{MAR}) \\
&\quad \overset{\text{defn. } \mu}{=} 0.
\end{aligned}
$$

Consequently, conditioning on X, and recalling that the tower property of conditional expectations holds almost surely, we have that

$$\partial_\pi \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid X] \overset{P_{Z,X}\text{-a.s.}}{=} 0;$$

i. e. (3.3.7b) holds.

Finally,

$$\partial_\mu \mathbb{E}[\rho(\mathcal{A}, \theta^*, \pi, \mu) \mid Z, X] \overset{(3.3.4)}{=} -\mathbb{E}\Big[\frac{D}{\pi(Z, X)} - 1 \;\Big|\; Z, X\Big] \overset{(3.3.1b)}{=} 0.$$

Consequently, conditioning on X, and recalling that the tower property of conditional expectations holds almost surely, we have that (3.3.7c) also holds. □

Proof of (3.3.12). Observe that

$$
\begin{aligned}
\mathbb{E}[\rho\rho' \mid Z, X] &\overset{(3.3.4)}{=} \mathbb{E}\bigg[\Big(\frac{Dg}{\pi} - \mu\Big[\frac{D}{\pi} - 1\Big]\Big)\Big(\frac{Dg}{\pi} - \mu\Big[\frac{D}{\pi} - 1\Big]\Big)' \;\bigg|\; Z, X\bigg] \\
&= \frac{1}{\pi^2}\mathbb{E}[Dgg' \mid Z, X] - \mathbb{E}\bigg[\frac{Dg}{\pi}\mu'\Big[\frac{D}{\pi} - 1\Big] \;\bigg|\; Z, X\bigg] \\
&\quad - \mathbb{E}\bigg[\mu\Big[\frac{D}{\pi} - 1\Big]\frac{Dg'}{\pi} \;\bigg|\; Z, X\bigg] + \mu\mu'\mathbb{E}\Big[\Big(\frac{D}{\pi} - 1\Big)^2 \;\Big|\; Z, X\Big]. \qquad \text{(3.C.1)}
\end{aligned}
$$

Now,

$$\mathbb{E}[Dgg' \mid Z, X] \overset{\text{MAR}}{=} \mathbb{E}[D \mid Z, X]\mathbb{E}[gg' \mid Z, X] = \pi\mathbb{E}[gg' \mid Z, X],$$

which implies that

$$\frac{1}{\pi^2}\mathbb{E}[Dgg' \mid Z, X] = \frac{1}{\pi}\mathbb{E}[gg' \mid Z, X].$$

Next,

$$\begin{aligned}
\mathbb{E}\Big[\frac{Dg}{\pi}\mu'\Big[\frac{D}{\pi}-1\Big]\;\Big|\;Z,X\Big] &= \mathbb{E}\Big[\frac{g}{\pi}\mu'\Big[\frac{D^2}{\pi}-D\Big]\;\Big|\;Z,X\Big] \\
&= \mathbb{E}\Big[\frac{Dg}{\pi}\mu'\Big[\frac{1}{\pi}-1\Big]\;\Big|\;Z,X\Big] && (D^2=D) \\
&= \mathbb{E}[Dg\mid Z,X]\mu'\frac{1}{\pi}\Big[\frac{1}{\pi}-1\Big] \\
&\overset{\text{MAR}}{=} \mathbb{E}[D\mid Z,X]\mathbb{E}[g\mid Z,X]\mu'\frac{1}{\pi}\Big[\frac{1}{\pi}-1\Big] \\
&= \mu\mu'\Big[\frac{1}{\pi}-1\Big]. && (\pi \overset{\text{def}}{=} \mathbb{E}[D\mid Z,X], \mu \overset{\text{def}}{=} \mathbb{E}[g\mid Z,X])
\end{aligned}$$

Moreover, since $\mathbb{E}\Big[\frac{D}{\pi}-1\mid Z,X\Big]=0$,

$$\mathbb{E}\Big[\Big(\frac{D}{\pi}-1\Big)^2\;\Big|\;Z,X\Big] = \operatorname{Var}\Big[\frac{D}{\pi}-1\;\Big|\;Z,X\Big] = \frac{1}{\pi^2}\operatorname{Var}[D\mid Z,X] = \frac{\pi(1-\pi)}{\pi^2} = \frac{1}{\pi}-1. \tag{3.C.2}$$

Hence,

$$\mathbb{E}[\rho\rho'\mid Z,X] \overset{(3.C.1)}{=} \frac{1}{\pi}\mathbb{E}[gg'\mid Z,X]-2\mu\mu'\Big[\frac{1}{\pi}-1\Big]+\mu\mu'\Big[\frac{1}{\pi}-1\Big] = \frac{1}{\pi}\mathbb{E}[gg'\mid Z,X]-\mu\mu'\Big[\frac{1}{\pi}-1\Big],$$

which implies that

$$\pi\mathbb{E}[\rho\rho'\mid Z,X]+(1-\pi)\mu\mu' = \mathbb{E}[gg'\mid Z,X]. \tag{3.C.3}$$

The desired result follows. □

Remark 3.C.1. Applying the $\operatorname{tr}\circ\operatorname{diag}$ operator to both sides of (3.C.3), we get that

$$\pi\mathbb{E}[\rho'\rho\mid Z,X]+(1-\pi)\mu'\mu = \mathbb{E}[g'g\mid Z,X].$$

Hence,

$$\mathbb{E}[\rho'\rho\mid Z,X] = \frac{1}{\pi}\mathbb{E}[g'g\mid Z,X]-\frac{(1-\pi)}{\pi}\mu'\mu \le \frac{1}{\inf\pi}\mathbb{E}[g'g\mid Z,X].$$

Consequently,

$$\mathbb{E}[\rho'\rho\mid X] \le \frac{1}{\inf\pi}\mathbb{E}[g'g\mid X] \le \frac{\|\sigma_g^2\|_\infty}{\inf\pi} \overset{\text{Ass. 3.3.1(i,ii)}}{<} \infty. \tag{3.C.4}$$

This bound is used in the proof of Lemma 3.C.1. □

Proof of Lemma 3.3.1. We use the approach of Severini and Tripathi (2001, 2013, Section 12) to derive the efficiency bound for θ^*. Let $I_0 \overset{\text{def}}{=} [0,t_0]$ for some $t_0>0$. With respect to an appropriate dominating measure, define the probability density function $v^2 \overset{\text{def}}{=} \operatorname{pdf}_{Y^*,Z|X}$. Let $t\mapsto v_t$ be a real-valued function defined on I_0 such that $v_t\big|_{t=0} = v$, and, suppressing the dominating measure, $\int v_t^2(y,z\mid x)=1$ for all $(t,x)\in I_0\times\operatorname{supp}(X)$. The score corresponding to $\dot v$, the tangent to v_t at $t=0$, is $S_{\dot v}\overset{\text{def}}{=} 2\dot v/v \in L_2(Y^*,Z,X)\cap L_2(X)^\perp$.

Let $t\mapsto\theta_t^*$ denote a $\mathbb{R}^{\dim\theta^*}$-valued function defined on I_0 such that $\theta_t^*\big|_{t=0}=\theta^*$ and

$\int_{\mathrm{supp}(Y^*)\times\mathrm{supp}(Z)} g(y,z,x,\theta_t^*)v_t^2(y,z \mid x) = 0$ for all $(t,x) \in I_0 \times \mathrm{supp}(X)$. Then, differentiating with respect to t and evaluating at $t = 0$, we have

$$J\dot{\theta}^* + \mathbb{E}[gS_{\dot{v}} \mid X] = 0 \qquad P_X\text{-a.s.}, \tag{3.C.5}$$

where the vector $\dot{\theta}^*$ is the tangent to θ_t^* at $t = 0$. Note that (3.C.5) restricts the tangent vector $S_{\dot{v}}$ to be such that

$$\mathbb{E}[gS_{\dot{v}} \mid X] \in \mathrm{span}(J) \qquad P_X\text{-a.s.} \tag{3.C.6}$$

Let $A \stackrel{\mathrm{def}}{=} A(X)$ be a $r \times (\dim g)$ matrix with $r \geq \dim g$ such that $B \stackrel{\mathrm{def}}{=} \mathbb{E}[AJ]$ has column rank $\dim \theta^*$,[61] and let B^+ denote the generalised inverse of B. Then,

$$\text{(3.C.5)} \implies \dot{\theta}^* = -B^+\mathbb{E}[A\mathbb{E}(gS_{\dot{v}} \mid X)]. \tag{3.C.7}$$

Since observations on Y^* can be missing, θ^* has to be identified as a feature of $q^2 \stackrel{\mathrm{def}}{=} \mathrm{pdf}_{D,Y^*,Z,X}$, the joint density of D, Y^*, Z, X. In particular, suppose that we want the efficiency bound for estimating the functional $\eta(\log q^2) \stackrel{\mathrm{def}}{=} c'\theta^*$, where $c \in \mathbb{R}^{\dim\theta^*}$ is arbitrary. Now,

$$\begin{aligned}
q^2 &\stackrel{\mathrm{def}}{=} \mathrm{pdf}_{D,Y^*,Z,X} \\
&= \mathrm{pdf}_{D,Y^*|Z,X}\,\mathrm{pdf}_{Z,X} \\
&\stackrel{\mathrm{MAR}}{=} \mathrm{pdf}_{D|Z,X}\,\mathrm{pdf}_{Y^*|Z,X}\,\mathrm{pdf}_{Z,X} \\
&= \mathrm{pdf}_{D|Z,X}\,\mathrm{pdf}_{Y^*,Z,X} \\
&= \mathrm{pdf}_{D|Z,X}\,\mathrm{pdf}_{Y^*,Z|X}\,\mathrm{pdf}_{X} \\
&= p^2v^2f^2,
\end{aligned}$$

where $p^2 \stackrel{\mathrm{def}}{=} \mathrm{pdf}_{D|Z,X}$ and $f^2 \stackrel{\mathrm{def}}{=} \mathrm{pdf}_X$. Hence, $\eta(\log q^2) \stackrel{\mathrm{MAR}}{=} \eta(\log p^2 + \log v^2 + \log f^2)$.

Let $t \mapsto p_t$ and $t \mapsto f_t$ be real-valued functions defined on I_0 such that $p_t\big|_{t=0} = p$, $f_t\big|_{t=0} = f$, and (suppressing the dominating measures), $\int p_t^2(d \mid z,x) = 1$ for all $(t,z,x) \in I_0 \times \mathrm{supp}(Z) \times \mathrm{supp}(X)$ and $\int f_t^2(x) = 1$ for all $(t,x) \in I_0 \times \mathrm{supp}(X)$ for all $t \in I_0$. Therefore, since $\log q_t^2 \stackrel{\mathrm{MAR}}{=} \log p_t^2 + \log v_t^2 + \log f_t^2$ and θ_t^* are related via the requirement that $\eta(\log q_t^2) = c'\theta_t^*$ for all $t \in I_0$, it follows that $\nabla\eta(S_{\dot{q}}) = c'\dot{\theta}^*$, where $\nabla\eta$ is the pathwise derivative of η and $S_{\dot{q}} \stackrel{\mathrm{def}}{=} S_{\dot{p}} + S_{\dot{v}} + S_{\dot{f}}$, with $S_{\dot{p}} \stackrel{\mathrm{def}}{=} 2\dot{p}/p \in L_2(D,Z,X) \cap L_2(Z,X)^\perp$ and $S_{\dot{f}} \stackrel{\mathrm{def}}{=} 2\dot{f}/f \in L_{2,0}(X)$. Hence, by (3.C.7),

$$\nabla\eta(S_{\dot{q}}) = -c'B^+\mathbb{E}[A\mathbb{E}(gS_{\dot{v}} \mid X)]. \tag{3.C.8}$$

To show that $\nabla\eta$ is a linear functional of $S_{\dot{q}}$, we also have to write the right-hand side of (3.C.8) in terms of $S_{\dot{q}}$. To do so, we now obtain an expression for $\mathbb{E}[gS_{\dot{v}} \mid X]$ in

[61] For instance, let $A(X) \stackrel{\mathrm{def}}{=} J'w(X)$, where $w(X)$ is a $(\dim g) \times (\dim g)$ matrix that is positive definite P_X-a.s. Since the columns of J are linearly independent P_X-a.s., so are the columns of the $(\dim\theta^*) \times (\dim\theta^*)$ matrix $J'w(X)J$. Hence, $\mathbb{E}[A(X)J]$ has column rank $\dim\theta^*$.

terms of $S_{\dot{q}}$. Let $\mu \stackrel{\text{def}}{=} \mathbb{E}[g \mid Z, X]$ and[62]

$$\rho \stackrel{\text{def}}{=} \rho(D, Y^*, Z, X) \stackrel{\text{def}}{=} \frac{Dg}{\pi} - \mu\Big[\frac{D}{\pi} - 1\Big]. \tag{3.C.9}$$

Then, as shown after the proof of this lemma,

$$\begin{aligned}
&\mathbb{E}[\rho S_{\dot{p}} \mid Z, X] = 0, \qquad \forall S_{\dot{p}} \in L_2(D, Z, X) \cap L_2(Z, X)^{\perp}, \\
&\mathbb{E}[\rho S_{\dot{v}} \mid Z, X] = \mathbb{E}[g S_{\dot{v}} \mid Z, X], \qquad \forall S_{\dot{v}} \in L_2(Y^*, Z, X) \cap L_2(X)^{\perp}, \\
&\mathbb{E}[\rho S_{\dot{f}} \mid Z, X] = \mu S_{\dot{f}}, \qquad \forall S_{\dot{f}} \in L_{2,0}(X).
\end{aligned} \tag{3.C.10}$$

Hence, since $S_{\dot{q}} \stackrel{\text{def}}{=} S_{\dot{p}} + S_{\dot{v}} + S_{\dot{f}}$, we have that

$$\mathbb{E}[\rho S_{\dot{q}} \mid Z, X] \stackrel{(3.C.10)}{=} \mathbb{E}[g S_{\dot{v}} \mid Z, X] + \mu S_{\dot{f}}.$$

Consequently, as $S_{\dot{f}} \in L_{2,0}(X)$ and $\mathbb{E}[\mu \mid X] = \mathbb{E}[g \mid X] \stackrel{P_X\text{-a.s.}}{=} 0$,

$$\mathbb{E}[\rho S_{\dot{q}} \mid X] = \mathbb{E}[g S_{\dot{v}} \mid X] + S_{\dot{f}} \mathbb{E}[\mu \mid X] = \mathbb{E}[g S_{\dot{v}} \mid X] \stackrel{(3.C.6)}{\in} \operatorname{span}(J) \quad P_X\text{-a.s.} \tag{3.C.11}$$

With the restrictions on $S_{\dot{q}}$ collected, it follows that

$$\nabla\eta(S_{\dot{q}}) \stackrel{(3.C.8)}{=} -c' B^+ \mathbb{E}[A \mathbb{E}(g S_{\dot{v}} \mid X)] \stackrel{(3.C.11)}{=} -c' B^+ \mathbb{E}[A \mathbb{E}(\rho S_{\dot{q}} \mid X)] \tag{3.C.12}$$

is a linear functional defined on the tangent space

$$\begin{aligned}
\dot{\mathcal{M}} \stackrel{\text{def}}{=} \{S_{\dot{q}} \in L_2(D, Y^*, Z, X) : {} & S_{\dot{q}} = S_{\dot{p}} + S_{\dot{v}} + S_{\dot{f}}, \text{ where } S_{\dot{p}} \in L_2(D, Z, X) \cap L_2(Z, X)^{\perp}, \\
& S_{\dot{v}} \in L_2(Y^*, Z, X) \cap L_2(X)^{\perp}, S_{\dot{f}} \in L_{2,0}(X), \\
& \text{and } \mathbb{E}[\rho S_{\dot{q}} \mid X] \in \operatorname{span}(J) \ P_X\text{-a.s.}\}
\end{aligned} \tag{3.C.13}$$

The tangent space is closed in the norm induced by the inner product $\langle \cdot, \cdot \rangle$ (Lemma 3.C.1).

Note that

$$\begin{aligned}
\nabla\eta(S_{\dot{q}}) \stackrel{(3.C.12)}{=} -c' B^+ \mathbb{E}[A \mathbb{E}(\rho S_{\dot{q}} \mid X)] &= -c' B^+ \mathbb{E}[A \rho S_{\dot{q}}] \qquad (S_{\dot{q}} \in \dot{\mathcal{M}}) \\
&= \langle -c' B^+ A \rho, S_{\dot{q}} \rangle \\
&= \langle -c' B^+ A \rho, \mathscr{P}_{\dot{\mathcal{M}}}(S_{\dot{q}}) \rangle \\
&= \langle -\mathscr{P}_{\dot{\mathcal{M}}}(c' B^+ A \rho), S_{\dot{q}} \rangle,
\end{aligned} \tag{3.C.14}$$

where the last equality is due to the fact that projection operators are self-adjoint. Since $-\mathscr{P}_{\dot{\mathcal{M}}}(c' B^+ A \rho) \in \dot{\mathcal{M}}$, it follows by (3.C.14) and the Riesz-Fréchet theorem (Luenberger, 1969, Theorem 2, p. 109) that if $\mathbb{E}[\mathscr{P}_{\dot{\mathcal{M}}}(c' B^+ A \rho)]^2 < \infty$, then $\nabla\eta$ is a bounded linear functional on the tangent space with representer $-\mathscr{P}_{\dot{\mathcal{M}}}(c' B^+ A \rho)$. This implies that η is a pathwise differentiable functional, and the efficiency bound for estimating η is given by the squared operator norm of $\nabla\eta$, namely, $\mathbb{E}[\mathscr{P}_{\dot{\mathcal{M}}}(c' B^+ A \rho)]^2$.

To obtain $\mathscr{P}_{\dot{\mathcal{M}}}(c' B^+ A \rho)$, we proceed as follows. Let

$$\dot{\mathcal{S}} \stackrel{\text{def}}{=} \{\dot{S} \in L_{2,0}(D, Y^*, Z, X) : \mathbb{E}[\rho \dot{S} \mid X] \in \operatorname{span}(J) \ P_X\text{-a.s.}\}.$$

[62] Since $Dg \stackrel{(3.1.1)}{=} Dg(Y, Z, X, \theta^*)$, the definitions of ρ in (3.3.4) and (3.C.9) are equivalent.

Then, $\dot{\mathcal{S}}$ is closed in the norm topology (same proof as for Lemma 3.C.1), and $\dot{\mathcal{M}} \subset \dot{\mathcal{S}}$.[63] Letting $V \stackrel{\text{def}}{=} \mathbb{E}J'\Omega_\rho^{-1}J$, where $\Omega_\rho \stackrel{\text{def}}{=} \mathbb{E}[\rho\rho' \mid X]$, we have that

$$\mathscr{P}_{\dot{\mathcal{S}}}(c'B^+A\rho) \stackrel{\text{Lemma 3.C.2}}{=} \rho'\Omega_\rho^{-1}JV^{-1}c.$$

But, as shown towards the end of this proof, $\rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathcal{M}}$. Therefore,

$$\mathscr{P}_{\dot{\mathcal{M}}}(c'B^+A\rho) = \rho'\Omega_\rho^{-1}JV^{-1}c. \qquad (\dot{\mathcal{M}} \subset \dot{\mathcal{S}})$$

Consequently, the efficiency bound for estimating η is given by[64]

$$\begin{aligned}
\mathbb{E}[\mathscr{P}_{\dot{\mathcal{M}}}(-c'B^+A\rho)]^2 &= c'V^{-1}\mathbb{E}[J\Omega_\rho^{-1}\rho\rho'\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}\mathbb{E}[J\Omega_\rho^{-1}\mathbb{E}(\rho\rho' \mid X)\Omega_\rho^{-1}J]V^{-1}c \\
&= c'V^{-1}c \qquad (3.C.15) \\
&< \infty. \qquad (\text{Ass. 3.3.1(iii)})
\end{aligned}$$

The desired result follows because c is arbitrary.

It remains to verify that $\rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathcal{M}}$. Observe that

$$\dot{m} \stackrel{\text{def}}{=} \rho'\Omega_\rho^{-1}JV^{-1}c \stackrel{(3.C.9)}{=} \Big(\frac{Dg}{\pi} - \mu\Big[\frac{D}{\pi} - 1\Big]\Big)'\Omega_\rho^{-1}JV^{-1}c =: \dot{m}_1 + \dot{m}_2 + \dot{m}_3,$$

where $\dot{m}_1 \stackrel{\text{def}}{=} -\mu'\Omega_\rho^{-1}JV^{-1}(D/\pi - 1)c$, $\dot{m}_2 \stackrel{\text{def}}{=} Dg'\Omega_\rho^{-1}JV^{-1}c/\pi$, and $\dot{m}_3 \stackrel{\text{def}}{=} 0$. Now,

$$\begin{aligned}
\mathbb{E}\dot{m}_1^2 &= c'V^{-1}\mathbb{E}\Big[J'\Omega_\rho^{-1}\Big(\frac{D}{\pi} - 1\Big)^2\mu\mu'\Omega_\rho^{-1}J\Big]V^{-1}c \\
&= c'V^{-1}\mathbb{E}\Big[J'\Omega_\rho^{-1}\mathbb{E}\Big[\Big(\frac{D}{\pi} - 1\Big)^2 \;\Big|\; Z, X\Big]\mu\mu'\Omega_\rho^{-1}J\Big]V^{-1}c \\
&\stackrel{(3.C.2)}{=} c'V^{-1}\mathbb{E}\Big[J'\Omega_\rho^{-1}\frac{1-\pi}{\pi}\mu\mu'\Omega_\rho^{-1}J\Big]V^{-1}c \\
&< \infty \qquad (\text{Ass. 3.3.1(iii)})
\end{aligned}$$

and

$$\mathbb{E}[\dot{m}_1 \mid Z, X] = -\mu'\Omega_\rho^{-1}JV^{-1}c\mathbb{E}\Big[\frac{D}{\pi} - 1 \;\Big|\; Z, X\Big] = 0 \implies \dot{m}_1 \in L_2(Z, X)^\perp.$$

---

[63] Let $\dot{m} \in \dot{\mathcal{M}}$. By (3.C.13), $\mathbb{E}[\rho\dot{m} \mid X] \in \text{span}(J)$ P_X-a.s., and $\dot{m} = S_{\dot{p}} + S_{\dot{v}} + S_{\dot{f}}$, where $S_{\dot{p}} \in L_2(D, Z, X) \cap L_2(Z, X)^\perp$, $S_{\dot{v}} \in L_2(Y^*, Z, X) \cap L_2(X)^\perp$, $S_{\dot{f}} \in L_{2,0}(X)$, which imply that $\mathbb{E}\dot{m} = 0$. Hence, $\dot{m} \in \dot{\mathcal{S}}$.

[64] Since $\mathbb{E}[\rho S_{\dot{p}} \mid Z, X] = 0$ in (3.C.10) holds irrespective of whether π is fully known or known up to a finite-dimensional parameter, and the argument leading to (3.C.15) does not depend on the form of $S_{\dot{p}}$, it follows that the efficiency bound for θ^* does not decrease if π is fully known, or known up to a finite-dimensional parameter.

Thus, $\dot{m}_1 \in L_2(D, Z, X) \cap L_2(X)^\perp$. In addition,

$$\begin{aligned}
\mathbb{E}\dot{m}_2^2 &= c'V^{-1}\mathbb{E}\Big[\frac{DJ'\Omega_\rho^{-1}gg'\Omega_\rho^{-1}J}{\pi^2}\Big]V^{-1}c \\
&= c'V^{-1}\mathbb{E}\Big[\frac{J'\Omega_\rho^{-1}}{\pi^2}\mathbb{E}[Dgg' \mid Z, X]\Omega_\rho^{-1}J\Big]V^{-1}c \\
&\overset{\text{MAR}}{=} c'V^{-1}\mathbb{E}\Big[\frac{J'\Omega_\rho^{-1}}{\pi^2}\mathbb{E}[D \mid Z, X]\mathbb{E}[gg' \mid Z, X]\Omega_\rho^{-1}J\Big]V^{-1}c \\
&= c'V^{-1}\mathbb{E}\Big[\frac{J'\Omega_\rho^{-1}\mathbb{E}[gg' \mid Z, X]\Omega_\rho^{-1}J}{\pi}\Big]V^{-1}c \\
&\overset{(3.3.12)}{=} c'V^{-1}\mathbb{E}\Big[J'\Omega_\rho^{-1}\Big(\mathbb{E}[\rho\rho' \mid Z, X] + \frac{1-\pi}{\pi}\mu\mu'\Big)\Omega_\rho^{-1}J\Big]V^{-1}c \\
&= c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\rho\rho'\Omega_\rho^{-1}J]V^{-1}c + c'V^{-1}\mathbb{E}\Big[J'\Omega_\rho^{-1}\frac{1-\pi}{\pi}\mu\mu'\Omega_\rho^{-1}J\Big]V^{-1}c \\
&= c'V^{-1}c + c'V^{-1}\mathbb{E}\Big[J'\Omega_\rho^{-1}\frac{1-\pi}{\pi}\mu\mu'\Omega_\rho^{-1}J\Big]V^{-1}c \qquad (\Omega_\rho \overset{\text{def}}{=} \mathbb{E}[\rho\rho' \mid X]) \\
&< \infty. \qquad \text{(Ass. 3.3.1(iii))}
\end{aligned}$$

Moreover, since $\pi(Z, X) \overset{\text{def}}{=} \mathbb{E}[D \mid Z, X]$ and $\mu \overset{\text{def}}{=} \mathbb{E}[g \mid Z, X]$,

$$\begin{aligned}
\mathbb{E}[\dot{m}_2 \mid Z, X] &= \mathbb{E}\Big[\frac{Dg'\Omega_\rho^{-1}JV^{-1}c}{\pi} \;\Big|\; Z, X\Big] \\
&= \frac{1}{\pi}\mathbb{E}[D \mid Z, X]\mathbb{E}[g' \mid Z, X]\Omega_\rho^{-1}JV^{-1}c \\
&= \mu'\Omega_\rho^{-1}JV^{-1}c.
\end{aligned}$$

Hence, as $\mathbb{E}[\mu \mid X] = \mathbb{E}[g \mid X] \overset{P_X\text{-a.s.}}{=} 0$,

$$\mathbb{E}[\dot{m}_2 \mid X] = \mathbb{E}[\mu' \mid X]\Omega_\rho^{-1}JV^{-1}c \overset{P_X\text{-a.s.}}{=} 0 \implies \dot{m}_2 \in L_2(X)^\perp.$$

Therefore, $\dot{m}_2 \in L_2(Y^*, Z, X) \cap L_2(X)^\perp$. Furthermore,

$$\mathbb{E}[\rho\dot{m} \mid Z, X] = \mathbb{E}[\rho\rho'\Omega_\rho^{-1}JV^{-1}c \mid Z, X] = \mathbb{E}[\rho\rho' \mid Z, X]\Omega_\rho^{-1}JV^{-1}c,$$

which implies that (recall $\Omega_\rho \overset{\text{def}}{=} \mathbb{E}[\rho\rho' \mid X]$)

$$\mathbb{E}[\rho\dot{m} \mid X] = JV^{-1}c \in \operatorname{span}(J).$$

Hence, $\rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathcal{M}}$. □

Lemma 3.C.1. *Under Assumptions* 3.2.1 *and* 3.3.1(i, ii)*, $\dot{\mathcal{M}}$ is closed.*

Proof of Lemma 3.C.1. The tangent vectors $S_{\dot{p}} \in L_2(D, Z, X) \cap L_2(Z, X)^\perp =: \dot{\mathcal{M}}_1$, $S_{\dot{v}} \in L_2(Y^*, Z, X) \cap L_2(X)^\perp =: \dot{\mathcal{M}}_2$, and $S_{\dot{f}} \in L_{2,0}(X) =: \dot{\mathcal{M}}_3$ are pairwise orthogonal. Indeed,

$$\mathbb{E}[S_{\dot{p}}S_{\dot{v}} \mid Z, X] \overset{\text{MAR}}{=} \mathbb{E}[S_{\dot{p}} \mid Z, X]\mathbb{E}[S_{\dot{v}} \mid Z, X] \overset{S_{\dot{p}} \in L_2(Z,X)^\perp}{=} 0 \implies S_{\dot{p}} \perp S_{\dot{v}}.$$

Similarly,

$$\mathbb{E}[S_{\dot{p}}S_{\dot{f}} \mid Z, X] = S_{\dot{f}}\mathbb{E}[S_{\dot{p}} \mid Z, X] \overset{S_{\dot{p}} \in L_2(Z,X)^\perp}{=} 0 \implies S_{\dot{p}} \perp S_{\dot{f}}.$$

and

$$\mathbb{E}[S_{\dot{v}}S_{\dot{f}} \mid X] = S_{\dot{f}}\mathbb{E}[S_{\dot{v}} \mid X] \overset{S_{\dot{v}}\in L_2(X)^\perp}{=} 0 \implies S_{\dot{v}} \perp S_{\dot{f}}.$$

Pairwise orthogonality of $\dot{\mathcal{M}}_1, \dot{\mathcal{M}}_2, \dot{\mathcal{M}}_3$ is used to show that $\mathrm{cl}(\dot{\mathcal{M}}) \subset \dot{\mathcal{M}}$.

Let $\dot{m} \in \mathrm{cl}(\dot{\mathcal{M}})$. Then, there exists a sequence $(\dot{m}_j)_{j\in\mathbb{N}} \subset \dot{\mathcal{M}}$ such that $\lim_{j\to\infty} \|\dot{m}_j - \dot{m}\|_2 = 0$. It remains to prove that $\dot{m} \in \dot{\mathcal{M}}$. Since $\dot{m}_j \in \dot{\mathcal{M}}$ for each j, by (3.C.13) we have that $\dot{m}_j = \dot{m}_{j1}+\dot{m}_{j2}+\dot{m}_{j3}$, where $(\dot{m}_{j1}, \dot{m}_{j2}, \dot{m}_{j3}) \in \dot{\mathcal{M}}_1\times\dot{\mathcal{M}}_2\times\dot{\mathcal{M}}_3$ and $\mathbb{E}[\rho\dot{m}_j \mid X] \in \mathrm{span}(J)$ P_X-a.s. Note that $\dot{\mathcal{M}}_1, \dot{\mathcal{M}}_2, \dot{\mathcal{M}}_3$ are $\|\cdot\|_2$-closed because L_2-spaces, and the orthogonal complements of their linear subspaces, are closed. Since $\dot{\mathcal{M}}_1, \dot{\mathcal{M}}_2, \dot{\mathcal{M}}_3$ are also pairwise orthogonal under MAR, $\dot{\mathscr{P}}_{\dot{\mathcal{M}}_1+\dot{\mathcal{M}}_2+\dot{\mathcal{M}}_3} = \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1} + \dot{\mathscr{P}}_{\dot{\mathcal{M}}_2} + \dot{\mathscr{P}}_{\dot{\mathcal{M}}_3}$. Consequently,

$$\begin{aligned}\dot{m}_j - \dot{m} &= \dot{m}_{j1} + \dot{m}_{j2} + \dot{m}_{j3} - \dot{m} \\ &= \dot{m}_{j1} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1}(\dot{m}) + \dot{m}_{j2} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_2}(\dot{m}) + \dot{m}_{j3} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_3}(\dot{m}) - [\dot{m} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1+\dot{\mathcal{M}}_2+\dot{\mathcal{M}}_3}(\dot{m})].\end{aligned}$$

Since $(\dot{m}_{j1}, \dot{m}_{j2}, \dot{m}_{j3}) \in \dot{\mathcal{M}}_1 \times \dot{\mathcal{M}}_2 \times \dot{\mathcal{M}}_3$, pairwise orthogonality of $\dot{\mathcal{M}}_1, \dot{\mathcal{M}}_2, \dot{\mathcal{M}}_3$ and the fact that $\dot{m} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1+\dot{\mathcal{M}}_2+\dot{\mathcal{M}}_3}(\dot{m}) \perp \dot{\mathcal{M}}_1 + \dot{\mathcal{M}}_2 + \dot{\mathcal{M}}_3$ imply that

$$\begin{aligned}\|\dot{m}_j - \dot{m}\|_2^2 = \|\dot{m}_{j1} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1}(\dot{m})\|_2^2 + \|\dot{m}_{j2} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_2}(\dot{m})\|_2^2 + \|\dot{m}_{j3} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_3}(\dot{m})\|_2^2 \\ + \|\dot{m} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1+\dot{\mathcal{M}}_2+\dot{\mathcal{M}}_3}(\dot{m})\|_2^2.\end{aligned}$$

Therefore, since $\lim_{j\to\infty} \|\dot{m}_j - \dot{m}\|_2 = 0$, we have $\lim_{j\to\infty} \|\dot{m}_{j1} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1}(\dot{m})\|_2 = 0$, $\lim_{j\to\infty} \|\dot{m}_{j2} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_2}(\dot{m})\|_2 = 0$, $\lim_{j\to\infty} \|\dot{m}_{j3} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_3}(\dot{m})\|_2 = 0$, and $\|\dot{m} - \dot{\mathscr{P}}_{\dot{\mathcal{M}}_1+\dot{\mathcal{M}}_2+\dot{\mathcal{M}}_3}(\dot{m})\|_2 = 0$. The last condition reveals that $\dot{m} \in \dot{\mathcal{M}}_1 + \dot{\mathcal{M}}_2 + \dot{\mathcal{M}}_3$, which means that $\dot{m} = \dot{m}_1 + \dot{m}_2 + \dot{m}_3$ for some $(\dot{m}_1, \dot{m}_2, \dot{m}_3) \in \dot{\mathcal{M}}_1 \times \dot{\mathcal{M}}_2 \times \dot{\mathcal{M}}_3$. Hence, $\dot{m} \in \dot{\mathcal{M}}$ follows if we can establish that $\mathbb{E}[\rho\dot{m} \mid X] \in \mathrm{span}(J)$ P_X-a.s.

We show that $\mathbb{E}[\rho\dot{m} \mid X] \in \mathrm{span}(J)$ P_X-a.s. by demonstrating that $\lim_{j\to\infty} \|\mathbb{E}[\rho\dot{m}_j \mid X] - \mathbb{E}[\rho\dot{m} \mid X]\|_2 = 0$ (recall that $\|b\|_2 \overset{\text{def}}{=} \sqrt{\mathbb{E}b'b}$ if b is a random vector). Indeed, $\lim_{j\to\infty} \|\mathbb{E}[\rho\dot{m}_j \mid X] - \mathbb{E}[\rho\dot{m} \mid X]\|_2 = 0$ implies that, P_X-a.s., $(\mathbb{E}[\rho\dot{m}_j \mid X])_{j\in\mathbb{N}}$ is a convergent sequence in $\mathrm{span}(J)$ because $\mathbb{E}[\rho\dot{m}_j \mid X] \in \mathrm{span}(J)$ P_X-a.s. for each j. Therefore, P_X-a.s., its limit $\mathbb{E}[\rho\dot{m} \mid X] \in \mathrm{span}(J)$ because $\mathrm{span}(J)$ is $\|\cdot\|_2$-closed (cf. the discussion after Assumption 3.3.1).

To show that $\lim_{j\to\infty} \|\mathbb{E}[\rho\dot{m}_j \mid X] - \mathbb{E}[\rho\dot{m} \mid X]\|_2 = 0$, we proceed as follows. Let $\rho^{(k)}$ denote the k^{th} coordinate of ρ. Then,

$$\begin{aligned}\|\mathbb{E}[\rho(\dot{m}_j - \dot{m}) \mid X]\|_2^2 &= \mathbb{E}\sum_{k=1}^{\dim g}\big(\mathbb{E}[\rho^{(k)}(\dot{m}_j - \dot{m}) \mid X]\big)^2 && (\|b\|_2 \overset{\text{def}}{=} \sqrt{\mathbb{E}b'b}, \dim\rho = \dim g) \\ &\le \mathbb{E}\sum_{k=1}^{\dim g}\mathbb{E}[(\rho^{(k)})^2 \mid X]\,\mathbb{E}[(\dot{m}_j - \dot{m})^2 \mid X] && \text{(cond. Cauchy-Schwarz)} \\ &= \mathbb{E}\big(\mathbb{E}[\rho'\rho \mid X]\,\mathbb{E}[(\dot{m}_j - \dot{m})^2 \mid X]\big) \\ &\overset{\text{(3.C.4)}}{\le} \frac{\|\sigma_g^2\|_\infty}{\inf\pi}\mathbb{E}(\dot{m}_j - \dot{m})^2 && \text{(Ass. 3.3.1(i,ii))} \\ &\xrightarrow[j\to\infty]{} 0. && (\lim_{j\to\infty}\|\dot{m}_j - \dot{m}\|_2 = 0)\end{aligned}$$

The desired result follows. □

Proof of (3.C.10). Observe that

$$\begin{aligned}
\mathbb{E}[\rho S_{\dot p} \mid Z, X] &= \mathbb{E}\Big[\frac{Dg}{\pi} S_{\dot p} \;\Big|\; Z, X\Big] - \mathbb{E}\Big[\mu\Big(\frac{D}{\pi} - 1\Big) S_{\dot p} \;\Big|\; Z, X\Big] \\
&\qquad\qquad (S_{\dot p} \in L_2(D, Z, X) \cap L_2(Z, X)^\perp) \\
&= \frac{1}{\pi}\mathbb{E}[DgS_{\dot p} \;\big|\; Z, X] - \mu\mathbb{E}\Big[\Big(\frac{D}{\pi} - 1\Big) S_{\dot p} \mid Z, X\Big] \\
&\overset{\text{MAR}}{=} \frac{1}{\pi}\mathbb{E}[DS_{\dot p} \mid Z, X]\mathbb{E}[g \mid Z, X] - \mu\Big(\frac{1}{\pi}\mathbb{E}[DS_{\dot p} \mid Z, X] - \mathbb{E}[S_{\dot p} \mid Z, X]\Big) \\
&= \frac{\mu}{\pi}\mathbb{E}[DS_{\dot p} \mid Z, X] - \frac{\mu}{\pi}\mathbb{E}[DS_{\dot p} \mid Z, X] \qquad (\mu \overset{\text{def}}{=} \mathbb{E}[g \mid Z, X], S_{\dot p} \in L_2(Z, X)^\perp) \\
&= 0.
\end{aligned}$$

Next,

$$\begin{aligned}
\mathbb{E}[\rho S_{\dot v} \mid Z, X] &= \mathbb{E}\Big[\frac{Dg}{\pi} S_{\dot v} \;\Big|\; Z, X\Big] - \mathbb{E}\Big[\mu\Big(\frac{D}{\pi} - 1\Big) S_{\dot v} \;\Big|\; Z, X\Big] \;\; (S_{\dot v} \in L_2(Y^*, Z, X) \cap L_2(X)^\perp) \\
&= \frac{1}{\pi}\mathbb{E}[DgS_{\dot v} \mid Z, X] - \mu\mathbb{E}\Big[\Big(\frac{D}{\pi} - 1\Big) S_{\dot v} \;\Big|\; Z, X\Big] \\
&\overset{\text{MAR}}{=} \frac{1}{\pi}\mathbb{E}[D \mid Z, X]\mathbb{E}[gS_{\dot v} \mid Z, X] - \mu\mathbb{E}\Big[\frac{D}{\pi} - 1 \;\Big|\; Z, X\Big]\mathbb{E}[S_{\dot v} \mid Z, X] \\
&= \mathbb{E}[gS_{\dot v} \mid Z, X]. \qquad (\mathbb{E}[\tfrac{D}{\pi} - 1 \mid Z, X] = 0)
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}[\rho S_{\dot f} \mid Z, X] &= \mathbb{E}\Big[\frac{Dg}{\pi} S_{\dot f} \;\Big|\; Z, X\Big] - \mathbb{E}\Big[\mu\Big(\frac{D}{\pi} - 1\Big) S_{\dot f} \;\Big|\; Z, X\Big] \qquad (S_{\dot f} \in L_{2,0}(X)) \\
&= \frac{S_{\dot f}}{\pi}\mathbb{E}[Dg \mid Z, X] - \mu S_{\dot f}\mathbb{E}\Big[\frac{D}{\pi} - 1 \;\Big|\; Z, X\Big] \\
&\overset{\text{MAR}}{=} \frac{S_{\dot f}}{\pi}\mathbb{E}[D \mid Z, X]\mathbb{E}[g \mid Z, X] \qquad (\mathbb{E}[\tfrac{D}{\pi} - 1 \mid Z, X] = 0) \\
&= S_{\dot f}\mu. \qquad (\mu \overset{\text{def}}{=} \mathbb{E}[g \mid Z, X])
\end{aligned}$$

The desired result follows. □

Lemma 3.C.2. $\mathscr{P}_{\dot{\mathcal{S}}}(c'B^+A\rho) = \rho'\Omega_\rho^{-1}JV^{-1}c$.

Proof of Lemma 3.C.2. Observe that $\dot m \overset{\text{def}}{=} \dot m(D, Y^*, Z, X) \overset{\text{def}}{=} \rho'\Omega_\rho^{-1}JV^{-1}c \in \dot{\mathcal{S}}$ because

$$\begin{aligned}
\mathbb{E}\dot m^2 = c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\rho\rho'\Omega_\rho^{-1}J]V^{-1}c = c'V^{-1}c &\overset{\text{Ass. 3.3.1(iii)}}{<} \infty, \\
\mathbb{E}[\dot m \mid X] = \mathbb{E}[\rho' \mid X]\Omega_\rho^{-1}JV^{-1}c &\overset{(3.3.6)}{=} 0 \implies \mathbb{E}\dot m = 0, \\
\mathbb{E}[\rho\dot m \mid X] = \mathbb{E}[\rho\rho' \mid X]\Omega_\rho^{-1}JV^{-1}c &= JV^{-1}c \in \operatorname{span}(J).
\end{aligned}$$

Next, let $\dot{S} \in \dot{\mathbb{S}}$. Then, since $\mathbb{E}[\rho\dot{S} \mid X] \in \operatorname{span}(J)$, which implies that $\mathbb{E}[\rho\dot{S} \mid X] = J\alpha$ for some $\alpha \in \mathbb{R}^{\dim \theta^*}$, we have that

$$\begin{aligned}
\langle c'B^+A\rho - \dot{m}, \dot{S}\rangle &= c'B^+\mathbb{E}[A\rho\dot{S}] - \mathbb{E}[\rho'\Omega_\rho^{-1}JV^{-1}c\dot{S}] && \text{(defn. } \dot{m}\text{)} \\
&= c'B^+\mathbb{E}[A\rho\dot{S}] - c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\rho\dot{S}] \\
&= c'B^+\mathbb{E}[A\mathbb{E}(\rho\dot{S} \mid X)] - c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}\mathbb{E}(\rho\dot{S} \mid X)] \\
&= c'B^+\mathbb{E}[AJ]\alpha - c'V^{-1}\mathbb{E}[J'\Omega_\rho^{-1}J]\alpha \\
&= c'B^+\mathbb{E}[AJ]\alpha - c'\alpha && (V \stackrel{\text{def}}{=} \mathbb{E}[J\Omega_\rho^{-1}J]) \\
&= c'B^+B\alpha - c'\alpha && (B \stackrel{\text{def}}{=} \mathbb{E}[AJ]) \\
&= c'\alpha - c'\alpha && (B \text{ full column rank}) \\
&= 0.
\end{aligned}$$

The desired result follows because $\dot{S}$ was arbitrary. □

3.D Efficiency gains in the simulation designs

Let $\tilde{Z} \stackrel{\text{def}}{=} (1, Z)_{2\times 1}$. The efficiency bound for estimating $\theta^* = (\alpha^*, \gamma^*)_{2\times 1}$ in the structural model $Y^* = \tilde{Z}'\theta^* + \sigma(X)U$ is given by $\text{l.b.}(\theta^*) = (\mathbb{E}J'J/\Omega_\rho)^{-1}$, where, cf. Example 3.3.2,

$$\begin{aligned}
J &= -\begin{bmatrix} 1 & \mathbb{E}[Z \mid X] \end{bmatrix} \\
\Omega_\rho &= \tilde{\pi}^{-1}\mathbb{E}[\operatorname{Var}(Y^* \mid Z, X) \mid X] + \mathbb{E}[\mu^2 \mid X] \\
\mu &= \mathbb{E}[Y^* \mid Z, X] - \tilde{Z}'\theta^*.
\end{aligned}$$

Furthermore, from (3.4.3), the efficiency bound for estimating θ^* using the validation sample alone is given by $\text{l.b.}(\theta^*)\big|_{\text{VS}} = (\mathbb{E}\tilde{\pi}J'J/\Omega_g)^{-1}$, where

$$\Omega_g = \mathbb{E}[(Y^* - \tilde{Z}'\theta^*)^2 \mid X] = \mathbb{E}[\sigma^2(X)U^2 \mid X] = \sigma^2(X)\sigma_U^2,$$

because $U \perp\!\!\!\perp X$. Hence, $\text{l.b.}(\theta^*)\big|_{\text{VS}} = \sigma_U^2\big(\mathbb{E}\tilde{\pi}J'J/\sigma^2(X)\big)^{-1}$. We now obtain the efficiency gain $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*)$ by simplifying the expressions for J and Ω_ρ in the two designs. For notational convenience, let $\tilde{X} \stackrel{\text{def}}{=} (1, X)_{2\times 1}$ and $\zeta \stackrel{\text{def}}{=} (\zeta_0, \zeta_1)_{2\times 1}$, so that $\zeta_0 + \zeta_1 X = \tilde{X}'\zeta$.

3.D.1 Design 1

In this design, $Z = \tilde{X}'\zeta + V$. Hence,

$$\begin{aligned}
\mathbb{E}[U \mid Z, X] &= \mathbb{E}[U \mid X, \tilde{X}'\zeta + V] && \text{(defn. of } Z\text{)} \\
&= \mathbb{E}[U \mid X, V] && ((X, V) \mapsto (X, \tilde{X}'\zeta + V) \text{ is injective}) \\
&= \mathbb{E}[U \mid V]. && ((U, V) \perp\!\!\!\perp X)
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[Y^* \mid Z, X] &= \tilde{Z}'\theta^* + \sigma(X)\mathbb{E}[U \mid Z, X] \\ &= \tilde{Z}'\theta^* + \sigma(X)\mathbb{E}[U \mid V] \\ &= \tilde{Z}'\theta^* + \sigma(X)\frac{\sigma_{UV}}{\sigma_V^2}V. && (U, V \text{ jointly normal})\end{aligned}$$

Consequently,

$$\mu = \sigma(X)\frac{\sigma_{UV}}{\sigma_V^2}V \implies \mathbb{E}[\mu^2 \mid X] = \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}.$$

Next, as $\mathrm{Var}[U \mid X] = \mathbb{E}[\mathrm{Var}\big[U \mid Z, X\big] \mid X\big] + \mathrm{Var}\big[\mathbb{E}[U \mid Z, X] \mid X\big]$ by variance decomposition,

$$\begin{aligned}\mathbb{E}\big[\mathrm{Var}[U \mid Z, X] \mid X\big] &= \mathrm{Var}[U \mid X] - \mathrm{Var}\big[\mathbb{E}[U \mid Z, X] \mid X\big] \\ &= \mathrm{Var}[U \mid X] - \mathrm{Var}\big[\mathbb{E}[U \mid V] \mid X\big] \\ &= \mathrm{Var}[U \mid X] - \mathrm{Var}\Big[\frac{\sigma_{UV}}{\sigma_V^2}V \;\Big|\; X\Big] \\ &= \mathrm{Var}[U \mid X] - \frac{\sigma_{UV}^2}{\sigma_V^4}\mathrm{Var}[V \mid X] \\ &= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}. && ((U, V) \perp\!\!\!\perp X)\end{aligned}$$

Consequently,

$$\mathbb{E}\big[\mathrm{Var}[Y^* \mid X, Z] \mid X\big] = \sigma^2(X)\mathbb{E}\big[\mathrm{Var}[U \mid Z, X] \mid X\big] = \sigma^2(X)\Big[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}\Big].$$

Combining these results, we get that

$$\Omega_\rho = \frac{1}{\tilde{\pi}(X)}\sigma^2(X)\Big[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}\Big] + \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}.$$

Therefore, the efficiency bound for θ^* in design 1 is

$$\mathrm{l.b.}(\theta^*) = \Big(\mathbb{E}\frac{J'J}{\Omega_\rho}\Big)^{-1} = \left(\mathbb{E}\frac{\begin{bmatrix} 1 & \tilde{X}'\zeta \\ \tilde{X}'\zeta & (\tilde{X}'\zeta)^2 \end{bmatrix}}{\frac{\sigma^2(X)}{\tilde{\pi}(X)}[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2}] + \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}}\right)^{-1}.$$

Furthermore, the efficiency bound for estimating θ^* using only the validation sample is

$$\mathrm{l.b.}(\theta^*)\big|_{\text{VS}} = \sigma_U^2\Big(\mathbb{E}\frac{\tilde{\pi}(X)}{\sigma^2(X)}J'J\Big)^{-1} = \sigma_U^2\Big(\mathbb{E}\frac{\tilde{\pi}(X)}{\sigma^2(X)}\begin{bmatrix} 1 & \tilde{X}'\zeta \\ \tilde{X}'\zeta & (\tilde{X}'\zeta)^2 \end{bmatrix}\Big)^{-1}.$$

Hence, the efficiency gain $\mathrm{l.b.}(\gamma^*)\big|_{\text{VS}}/\mathrm{l.b.}(\gamma^*)$ can be obtained from the expressions for the 2×2 matrices $\mathrm{l.b.}(\theta^*)\big|_{\text{VS}}$ and $\mathrm{l.b.}(\theta^*)$ by extracting their $(2, 2)$ elements.

3.D.2 Design 2

In this design, $Z = \mathbb{1}(\tilde{X}'\zeta + V > 0)$. Hence,

$$J = -[1 \quad \mathbb{E}[Z \mid X]] = -[1 \quad \Pr(Z = 1 \mid X)] = -[1 \quad \Phi(\tfrac{\tilde{X}'\zeta}{\sigma_V})].$$

For $d \in \mathbb{R}$, joint normality of (U, V) implies that

$$\mathbb{E}[U\mathbb{1}(V \leq d)] = \mathbb{E}\big[\mathbb{E}[U \mid V]\mathbb{1}(V \leq d)\big] = \frac{\sigma_{UV}}{\sigma_V^2}\mathbb{E}[V\mathbb{1}(V \leq d)] = -\frac{\sigma_{UV}}{\sigma_V}\phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big).$$

Hence, $\mathbb{E}[U\mathbb{1}(V > d)] = \dfrac{\sigma_{UV}}{\sigma_V}\phi\Big(\dfrac{\tilde{X}'\zeta}{\sigma_V}\Big)$ because $\mathbb{E}U = 0$. Consequently,

$$\begin{aligned}
\mathbb{E}[U \mid Z, X] &= (1 - Z)\mathbb{E}[U \mid Z = 0, X] + Z\mathbb{E}[U \mid Z = 1, X] && (Z \in \{0, 1\}) \\
&= (1 - Z)\frac{\mathbb{E}[U\mathbb{1}(Z = 0) \mid X]}{\Pr(Z = 0 \mid X)} + Z\frac{\mathbb{E}[U\mathbb{1}(Z = 1) \mid X]}{\Pr(Z = 1 \mid X)} \\
&= (1 - Z)\frac{\mathbb{E}[U\mathbb{1}(V \leq -\tilde{X}'\zeta) \mid X]}{\Pr(V \leq -\tilde{X}'\zeta \mid X)} + Z\frac{\mathbb{E}[U\mathbb{1}(V > -\tilde{X}'\zeta) \mid X]}{\Pr(V > -\tilde{X}'\zeta \mid X)} \\
&= -(1 - Z)\frac{\sigma_{UV}}{\sigma_V}\frac{\phi(\frac{\tilde{X}'\zeta}{\sigma_V})}{\Phi(\frac{-\tilde{X}'\zeta}{\sigma_V})} + Z\frac{\sigma_{UV}}{\sigma_V}\frac{\phi(\frac{\tilde{X}'\zeta}{\sigma_V})}{\Phi(\frac{\tilde{X}'\zeta}{\sigma_V})} && ((U, V) \text{ normal and indep. of } X) \\
&= \frac{\sigma_{UV}}{\sigma_V}\Big[Z - \Phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Big]G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big),
\end{aligned}$$

where $G(t) \stackrel{\text{def}}{=} \phi(t)/[\Phi(t)\Phi(-t)]$, $t \in \mathbb{R}$, is the probit weight function (Schumann & Tripathi, 2018). Therefore,

$$\begin{aligned}
\mathbb{E}[Y^* \mid Z, X] &= \tilde{Z}'\theta^* + \sigma(X)\mathbb{E}[U \mid Z, X] \\
&= \tilde{Z}'\theta^* + \sigma(X)\frac{\sigma_{UV}}{\sigma_V}\Big[Z - \Phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Big]G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big).
\end{aligned}$$

Consequently,

$$\mu = \sigma(X)\frac{\sigma_{UV}}{\sigma_V}\Big[Z - \Phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Big]G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big).$$

Therefore, since $\mathbb{E}[\mu \mid X] = 0$,

$$\begin{aligned}
\mathbb{E}[\mu^2 \mid X] &= \operatorname{Var}[\mu \mid X] \\
&= \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}G^2\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\operatorname{Var}[Z \mid X] \\
&= \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}G^2\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Phi\Big(-\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \\
&= \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2}G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big).
\end{aligned}$$

Next, as $\mathrm{Var}[U \mid X] = \mathbb{E}\big[\mathrm{Var}[U \mid Z, X] \mid X\big] + \mathrm{Var}\big[\mathbb{E}[U \mid Z, X] \mid X\big]$ by variance decomposition,

$$\begin{aligned}
\mathbb{E}[\mathrm{Var}\big[U \mid Z, X\big] \mid X\big] &= \mathrm{Var}[U \mid X] - \mathrm{Var}\big[\mathbb{E}[U \mid Z, X] \mid X\big] \\
&= \sigma_U^2 - \mathrm{Var}\left[\frac{\sigma_{UV}}{\sigma_V}\Big[Z - \Phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Big] G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \;\Big|\; X\right] \qquad (U \perp\!\!\!\perp X) \\
&= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G^2\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\, \mathrm{Var}[Z \mid X] \\
&= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G^2\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \Phi\Big(-\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \Phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \\
&= \sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big).
\end{aligned}$$

Consequently,

$$\begin{aligned}
\mathbb{E}\big[\mathrm{Var}[Y^* \mid X, Z] \mid X\big] &= \sigma^2(X) \mathbb{E}\big[\mathrm{Var}[U \mid Z, X] \mid X\big] \\
&= \sigma^2(X)\Big[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Big].
\end{aligned}$$

Combining these results, we get that

$$\Omega_\rho = \frac{1}{\tilde{\pi}(X)} \sigma^2(X)\Big[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big)\Big] + \sigma^2(X) \frac{\sigma_{UV}^2}{\sigma_V^2} G\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big) \phi\Big(\frac{\tilde{X}'\zeta}{\sigma_V}\Big).$$

Therefore, the efficiency bound for θ^* in design 2 is

$$\begin{aligned}
\text{l.b.}(\theta^*) &= \Big(\mathbb{E}\frac{J'J}{\Omega_\rho}\Big)^{-1} \\
&= \left(\mathbb{E}\frac{\begin{bmatrix} 1 & \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) \\ \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) & \Phi^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \end{bmatrix}}{\frac{\sigma^2(X)}{\tilde{\pi}(X)}[\sigma_U^2 - \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V})\phi(\frac{\tilde{X}'\zeta}{\sigma_V})] + \sigma^2(X)\frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V})\phi(\frac{\tilde{X}'\zeta}{\sigma_V})}\right)^{-1} \\
&= \left(\mathbb{E}\frac{\begin{bmatrix} 1 & \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) \\ \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) & \Phi^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \end{bmatrix}}{\frac{(1-X)\sigma^2(0)+X\sigma^2(1)}{(1-X)\tilde{\pi}(0)+X\tilde{\pi}(1)}[\sigma_U^2 - \mathfrak{p}] + [(1-X)\sigma^2(0) + X\sigma^2(1)]\mathfrak{p}}\right)^{-1},
\end{aligned}$$

where $\mathfrak{p} \stackrel{\text{def}}{=} \frac{\sigma_{UV}^2}{\sigma_V^2} G(\frac{\tilde{X}'\zeta}{\sigma_V})\phi(\frac{\tilde{X}'\zeta}{\sigma_V})$ and the last equality follows because $X \in \{0, 1\}$.

Furthermore, the efficiency bound for estimating θ^* using only the validation sample is

$$\text{l.b.}(\theta^*)\big|_{\text{VS}} = \sigma_U^2\Big(\mathbb{E}\frac{\tilde{\pi}(X)}{\sigma^2(X)} J'J\Big)^{-1} = \sigma_U^2\left(\mathbb{E}\frac{(1-X)\tilde{\pi}(0) + X\tilde{\pi}(1)}{(1-X)\sigma^2(0) + X\sigma^2(1)} \begin{bmatrix} 1 & \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) \\ \Phi(\frac{\tilde{X}'\zeta}{\sigma_V}) & \Phi^2(\frac{\tilde{X}'\zeta}{\sigma_V}) \end{bmatrix}\right)^{-1}.$$

Hence, the efficiency gain $\text{l.b.}(\gamma^*)\big|_{\text{VS}}/\text{l.b.}(\gamma^*)$ can be obtained from the expressions for the 2×2 matrices $\text{l.b.}(\theta^*)\big|_{\text{VS}}$ and $\text{l.b.}(\theta^*)$ by extracting their $(2, 2)$ elements.

SEL in Design 2

Here, $\hat{\rho}_j(\theta) \overset{\text{def}}{=} \hat{\rho}(\mathcal{A}_j, \theta)$ is scalar and $\hat{\theta}$ maximises the version of $\mathrm{SEL}_{\mathbb{T}}$ with $\mathbb{T}_{i,n} \overset{\text{def}}{=} 1$ and the weights given in Footnote 52, namely,

$$\mathrm{SEL}(\theta) \overset{\text{def}}{=} -\sum_{i=1}^{n} \max_{\lambda_i \in \mathbb{R}} \sum_{j=1}^{n} w_{ij} \log(1 + \lambda_i \hat{\rho}_j(\theta)),$$

where the redefined weights

$$w_{ij} \overset{\text{def}}{=} \frac{\mathbb{1}(X_i = X_j)}{\sum_{k=1}^{n} \mathbb{1}(X_i = X_k)} = \mathbb{1}(X_j = X_i)\Big[\frac{\mathbb{1}(X_i = 0)}{n(1-\bar{X})} + \frac{\mathbb{1}(X_i = 1)}{n\bar{X}}\Big] \tag{3.D.1}$$

and $\bar{X} \overset{\text{def}}{=} \sum_{j=1}^{n} X_j/n$. The maximisers of the inner optimization problems in $\mathrm{SEL}(\theta)$, denoted by $\hat{\lambda}_i$, $i = 1, \ldots, n$, satisfy the FOC

$$\begin{aligned}
0 &= \sum_{j=1}^{n} \frac{w_{ij}\hat{\rho}_j(\theta)}{1 + \hat{\lambda}_i \hat{\rho}_j(\theta)} && (i = 1, \ldots, n) \\
&\overset{(3.D.1)}{=} \frac{\mathbb{1}(X_i = 0)}{n(1-\bar{X})} \sum_{j=1}^{n} \frac{\mathbb{1}(X_j = X_i)\hat{\rho}_j(\theta)}{1 + \hat{\lambda}_i \hat{\rho}_j(\theta)} + \frac{\mathbb{1}(X_i = 1)}{n\bar{X}} \sum_{j=1}^{n} \frac{\mathbb{1}(X_j = X_i)\hat{\rho}_j(\theta)}{\mathbb{1} + \hat{\lambda}_i \hat{\rho}_j(\theta)} \\
&= \begin{cases} \displaystyle\sum_{j=1}^{n} \frac{\mathbb{1}(X_j = 0)\hat{\rho}_j(\theta)}{1 + \hat{l}_0 \hat{\rho}_j(\theta)} & \text{if } X_i = 0 \\ \displaystyle\sum_{j=1}^{n} \frac{\mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)}{1 + \hat{l}_1 \hat{\rho}_j(\theta)} & \text{if } X_i = 1, \end{cases} && \text{(3.D.2)}
\end{aligned}$$

where the real numbers $\hat{l}_0, \hat{l}_1$ solve (3.D.2). Consequently, $\hat{\lambda}_i = \hat{l}_0 \mathbb{1}(X_i = 0) + \hat{l}_1 \mathbb{1}(X_i = 1)$, $i = 1, \ldots, n$, and we have

$$\begin{aligned}
\mathrm{SEL}(\theta) &= -\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \log\big(1 + \hat{\lambda}_i \hat{\rho}_j(\theta)\big) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^{n} \mathbb{1}(X_i = X_k)} \log\big(1 + \hat{l}_0 \mathbb{1}(X_i = 0)\hat{\rho}_j(\theta) + \hat{l}_1 \mathbb{1}(X_i = 1)\hat{\rho}_j(\theta)\big) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^{n} \mathbb{1}(X_i = X_k)} \log\big(1 + \hat{l}_0 \mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + \hat{l}_1 \mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)\big) \\
&= -\sum_{j=1}^{n} \log\big(1 + \hat{l}_0 \mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + \hat{l}_1 \mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)\big) \sum_{i=1}^{n} \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^{n} \mathbb{1}(X_i = X_k)} \\
&= -\sum_{j=1}^{n} \log\big(1 + \hat{l}_0 \mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + \hat{l}_1 \mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)\big) && \text{(3.D.3)} \\
&\overset{(3.D.2)}{=} -\max_{l_0, l_1 \in \mathbb{R}} \sum_{j=1}^{n} \log\big(1 + l_0 \mathbb{1}(X_j = 0)\hat{\rho}_j(\theta) + l_1 \mathbb{1}(X_j = 1)\hat{\rho}_j(\theta)\big),
\end{aligned}$$

where (3.D.3) follows from the fact that

$$\begin{aligned}
\sum_{i=1}^{n} \frac{\mathbb{1}(X_j = X_i)}{\sum_{k=1}^{n} \mathbb{1}(X_i = X_k)} &= \sum_{i=1}^{n} \frac{\mathbb{1}(X_j = X_i)}{\mathbb{1}(X_i = 0)n(1-\bar{X}) + \mathbb{1}(X_i = 1)n\bar{X}} \\
&= \sum_{i=1}^{n} \frac{\mathbb{1}(X_j = X_i)[\mathbb{1}(X_j = 0) + \mathbb{1}(X_j = 1)]}{\mathbb{1}(X_i = 0)n(1-\bar{X}) + \mathbb{1}(X_i = 1)n\bar{X}} \\
&= \sum_{i=1}^{n} \frac{\mathbb{1}(X_i = 0)\mathbb{1}(X_j = 0) + \mathbb{1}(X_i = 1)\mathbb{1}(X_j = 1)}{\mathbb{1}(X_i = 0)n(1-\bar{X}) + \mathbb{1}(X_i = 1)n\bar{X}} \\
&= \frac{\mathbb{1}(X_j = 0)}{n(1-\bar{X})} \sum_{i=1}^{n} \mathbb{1}(X_i = 0) + \frac{\mathbb{1}(X_j = 1)}{n\bar{X}} \sum_{i=1}^{n} \mathbb{1}(X_i = 1) \\
&= \mathbb{1}(X_j = 0) + \mathbb{1}(X_j = 1) \\
&= 1.
\end{aligned}$$

Therefore, $\mathrm{SEL}(\cdot)$ coincides with unconditional empirical likelihood because

$$\begin{aligned}
\mathrm{SEL}(\theta) &= -\max_{l_0, l_1 \in \mathbb{R}} \sum_{j=1}^{n} \log\big(1 + \mathbb{1}(X_j = 0) l_0 \hat{\rho}_j(\theta) + \mathbb{1}(X_j = 1) l_1 \hat{\rho}_j(\theta)\big) \\
&= -\max_{l_0, l_1 \in \mathbb{R}} \sum_{j=1}^{n} \log\big(1 + l_0 \hat{\rho}_j(\theta) + \mathbb{1}(X_j = 1)(l_1 - l_0) \hat{\rho}_j(\theta)\big) \\
&= -\max_{l \in \mathbb{R}^2} \sum_{j=1}^{n} \log\big(1 + l' \tilde{X}_j \hat{\rho}_j(\theta)\big). \qquad (X \in \{0,1\} \implies \mathbb{1}(X = 1) = X)
\end{aligned}$$

Consequently, if $\hat{\theta}$ solves $\sum_{j=1}^{n} \tilde{X}_j \hat{\rho}_j(\hat{\theta}) = 0$, then it also maximises $\mathrm{SEL}(\cdot)$.

Bibliography

Adrian, T., Crump, R. K. & Vogt, E. (2019). Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds. *Journal of Finance*, *74*(4), 1931–1973.

Ai, C. & Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, *71*, 1795–1843.

Ai, C. & Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, *170*, 442–457.

Ait-Sahalia, Y. (2004). Disentangling diffusion from jumps. *Journal of Financial Economics*, *74*(3), 487–528.

Alberg, D., Shalit, H. & Yosef, R. (2008). Estimating stock market volatility using asymmetric GARCH models. *Applied Financial Economics*, *18*(15), 1201–1208.

Alexander, C. & Lazar, E. (2009). Modelling regime-specific stock price volatility. *Oxford Bulletin of Economics and Statistics*, *71*(6), 761–797.

Anatolyev, S. & Petukhov, A. (2016). Uncovering the skewness news impact curve. *Journal of Financial Econometrics*, *14*(4), 746–771.

Anatolyev, S. & Tarasyuk, I. (2015). Missing mean does no harm to volatility! *Economics Letters*, *134*, 62–64.

Andersen, T. G., Bollerslev, T. & Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *Review of Economics and Statistics*, *89*(4), 701–720.

Ang, A., Chen, J. & Xing, Y. (2006). Downside risk. *Review of Financial Studies*, *19*(4), 1191–1239.

Ang, A., Hodrick, R. J., Xing, Y. & Zhang, X. (2006). The cross-section of volatility and expected returns. *Journal of Finance*, *61*(1), 259–299.

Ang, A., Hodrick, R. J., Xing, Y. & Zhang, X. (2009). High idiosyncratic volatility and low returns: International and further US evidence. *Journal of Financial Economics*, *91*(1), 1–23.

Ardia, D., Boudt, K., Carl, P., Mullen, K. M. & Peterson, B. G. (2011). Differential evolution with DEoptim. *R Journal*, *3*(1), 27–34.

Atilgan, Y., Bali, T. G., Demirtas, K. O. & Gunaydin, A. D. (2019). Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns. *Journal of Financial Economics*.

Babenko, I., Boguth, O. & Tserlukevich, Y. (2016). Idiosyncratic cash flows and systematic risk. *Journal of Finance*, *71*(1), 425–456.

Bajgrowicz, P., Scaillet, O. & Treccani, A. (2015). Jumps in high-frequency data: Spurious detections, dynamics, and news. *Management Science*, *62*(8), 2198–2217.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2008). Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, *76*(6), 1481–1536.

Barndorff-Nielsen, O. E., Kinnebrouk, S. & Shephard, N. (2010). Measuring downside risk: Realised semivariance. *Volatility and time series econometrics: Essays in honor of Robert*

F. Engle ((Edited by T. Bollerslev, J. Russell and M. Watson), pp. 117–136). Oxford University Press.

Bartram, S. M., Brown, G. & Stulz, R. M. (2012). Why are US stocks more volatile? *Journal of Finance*, *67*(4), 1329–1370.

Bekaert, G. & Engstrom, E. (2017). Asset return dynamics under habits and bad environment–good environment fundamentals. *Journal of Political Economy*, *125*(3), 713–760.

Bekaert, G., Engstrom, E. & Ermolov, A. (2015). Bad environments, good environments: A non-gaussian asymmetric volatility model. *Journal of Econometrics*, *186*(1), 258–275.

Bekaert, G. & Wu, G. (2000). Asymmetric volatility and risk in equity markets. *Review of Financial Studies*, *13*(1), 1–42.

Bhattacharya, D. (2005). Asymptotic inference from multi-stage samples. *Journal of Econometrics*, *126*, 145–171.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics*, *137*, 674–707.

Bickel, P. J., Klassen, C. A. J., Ritov, Y. & Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, MD, USA.

Bickel, P. J. & Ritov, Y. (1991). Large sample theory of estimation in biased sampling regression models. *Annals of Statistics*, *19*, 797–816.

Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economical Statistics Section*, 177–181.

Bollerslev, T., Li, J., Patton, A. J. & Quaedvlieg, R. (2020). Realized semicovariances. *Econometrica*, *88*(4), 1515–1551.

Bollerslev, T., Li, S. Z. & Todorov, V. (2016). Roughing up beta: Continuous versus discontinuous betas and the cross section of expected stock returns. *Journal of Financial Economics*, *120*(3), 464–490.

Bollerslev, T., Li, S. Z. & Zhao, B. (2019). Good volatility, bad volatility, and the cross section of stock returns. *Journal of Financial and Quantitative Analysis*, 1–31.

Bollerslev, T., Litvinova, J. & Tauchen, G. (2006). Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics*, *4*(3), 353–384.

Bollerslev, T. & Wooldridge, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Reviews*, *11*(2), 143–172.

Butler, J. S. (2000). Efficiency results of MLE and GMM estimation with sampling weights. *Journal of Econometrics*, *96*, 25–37.

Calvet, L. E. & Fisher, A. J. (2004). How to forecast long-run volatility: Regime switching and the estimation of multifractal processes. *Journal of Financial Econometrics*, *2*(1), 49–83.

Campbell, J. Y., Hilscher, J. & Szilagyi, J. (2008). In search of distress risk. *Journal of Finance*, *63*(6), 2899–2939.

Caporin, M. & McAleer, M. (2006). Dynamic asymmetric GARCH. *Journal of Financial Econometrics*, *4*(3), 385–412.

Carr, P. & Wu, L. (2007). Stochastic skew in currency options. *Journal of Financial Economics*, *86*(1), 213–247.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, *34*, 305–334.

Chen, J., Hong, H. & Stein, J. C. (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics*, *61*(3), 345–381.

Chen, X., Hong, H. & Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Annals of Statistics*, *36*, 343–366.

Christensen, K., Oomen, R. C. A. & Podolskij, M. (2014). Fact or friction: Jumps at ultra high frequency. *Journal of Financial Economics*, *114*(3), 576–599.

Christie, A. A. (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics*, *10*(4), 407–432.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, *39*(4), 841–862.
Christoffersen, P. F. & Pelletier, D. (2004). Backtesting value-at-risk: A duration-based approach. *Journal of Financial Econometrics*, *2*(1), 84–108.
Clark, T. E. & McCracken, M. W. (2015). Nested forecast model comparisons: A new approach to testing equal accuracy. *Journal of Econometrics*, *186*(1), 160–177.
Cosslett, S. R. (1981a). Efficient estimation of discrete choice models. In C. F. Manski & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications* (pp. 51–111). MIT Press, Cambridge, MA, USA.
Cosslett, S. R. (1981b). Maximum likelihood estimation for choice-based samples. *Econometrica*, *49*, 1289–1316.
Cosslett, S. R. (1991). Efficient estimation from endogenously stratified samples with prior information on marginal probabilities [Manuscript available at https://www.asc.ohio-state.edu/cosslett.1/papers/cbsample1.pdf].
Cosslett, S. R. (1993). Estimation from endogenously stratified samples. In G. Maddala, C. Rao & H. Vinod (Eds.), *Handbook of statistics, vol. 11* (pp. 1–43). Elsevier, The Netherlands.
Cragg, J. G. (1983). More efficient estimation in the presence of heteroscedasticity of unknown form. *Econometrica*, *49*, 751–764.
Crouhy, M. & Rockinger, M. (1997). Volatility clustering, asymmetry and hysteresis in stock returns: International evidence. *Financial Engineering and the Japanese Markets*, *4*(1), 1–35.
Deaton, A. (1997). *The analysis of household surveys*. Johns Hopkins University Press, Baltimore, MD, USA.
DeMets, D. & Halperin, M. (1977). Estimation of a simple regression coefficient in samples arising from a subsampling procedure. *Biometrics*, *33*, 47–56.
Diebold, F. X., Gunther, T. A. & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 863–883.
Diebold, F. X. & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.
Ding, Z., Granger, C. W. J. & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, *1*(1), 83–106.
El Babsiri, M. & Zakoian, J.-M. (2001). Contemporaneous asymmetry in GARCH processes. *Journal of Econometrics*, *101*(2), 257–294.
El-Barmi, H. & Rothmann, M. (1998). Nonparametric estimation in selection biased models in the presence of estimating equations. *Nonparametric Statistics*, *9*, 381–399.
Engle, R. F., Lilien, D. M. & Robins, R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica*, 391–407.
Engle, R. F. & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, *22*(4), 367–381.
Engle, R. F. & Mistry, A. (2014). Priced risk and asymmetric volatility in the cross section of skewness. *Journal of Econometrics*, *182*(1), 135–144.
Engle, R. F. & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, *48*(5), 1749–1778.
Epstein, L. G. & Schneider, M. (2008). Ambiguity, information quality, and asset pricing. *Journal of Finance*, *63*(1), 197–228.
Fan, J., Qi, L. & Xiu, D. (2014). Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics*, *32*(2), 178–191.
Fernández, C. & Steel, M. F. J. (1998). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, *93*(441), 359–371.
Feunou, B. & Okou, C. (2019). Good volatility, bad volatility, and option pricing. *Journal of Financial and Quantitative Analysis*, *54*(2), 695–727.
Ghalanos, A. (2020). Introduction to the rugarch package. (version 1.3-8).

Ghysels, E., Guérin, P. & Marcellino, M. (2014). Regime switches in the risk-return trade-off. *Journal of Empirical Finance*, *28*, 118–138.

Gill, R. D., Vardi, Y. & Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics*, *16*, 1069–1112.

Glosten, L. R., Jagannathan, R. & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, *48*, 1779–1801.

González-Rivera, G. & Sun, Y. (2015). Generalized autocontours: Evaluation of multivariate density models. *International Journal of Forecasting*, *31*(3), 799–814.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, *79*, 437–452.

Guégan, D. & Diebolt, J. (1994). Probabilistic properties of the β-ARCH model. *Statistica Sinica*, 71–87.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*, 315–331.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of Multivariate Analysis*, *32*, 177–203.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 705–730.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 413–430.

Harvey, C. R. & Siddique, A. (1999). Autoregressive conditional skewness. *Journal of Financial and Quantitative Analysis*, *34*(4), 465–487.

Hausman, J. A. & Wise, D. A. (1981). Stratification on endogenous variables and estimation: The Gary income maintenance experiment. In C. F. Manski & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications* (pp. 365–391). MIT Press, Cambridge, MA, USA.

Heber, G., Lunde, A., Shephard, N. & Sheppard, K. (2009). Oxford-Man Institute's realized library, version 0.3.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Heckman, J. J., Lochner, L. J. & Todd, P. E. (2003). *Fifty years of Mincer earnings regressions* (tech. rep.). National Bureau of Economic Research.

Hentschel, L. (1995). All in the family: Nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics*, *39*, 71–104.

Hirose, Y. (2007). M-estimators in semi-parametric multi-sample models [Manuscript available at https://sms.wgtn.ac.nz/foswiki/pub/Main/ResearchReportSeries/mscs08-05.pdf].

Hirose, Y. & Lee, A. J. (2008). Semi-parametric efficiency bounds for regression models under generalised case-control sampling: The profile likelihood approach. *Annals of the Institute of Statistical Mathematics*, *62*, 1023–1052.

Holt, D., Smith, T. M. F. & Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of The Royal Statistical Society, Series A*, *143*, 474–487.

Hong, H., Stein, J. C. & Yu, J. (2007). Simple forecasts and paradigm shifts. *Journal of Finance*, *62*(3), 1207–1242.

Hou, K. & Loh, R. K. (2016). Have we solved the idiosyncratic volatility puzzle? *Journal of Financial Economics*, *121*(1), 167–194.

Hristache, M. & Patilea, V. (2016). Semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables. *Econometric Theory*, *32*, 917–946.

Hristache, M. & Patilea, V. (2017). Conditional moment models with data missing at random. *Biometrica*, *104*, 735–742.
Imbens, G. W. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, *60*, 1187–1214.
Imbens, G. W. & Lancaster, T. (1996). Efficient estimation and stratified sampling. *Journal of Econometrics*, *74*, 289–318.
Jewell, N. P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, *72*, 11–21.
Jorion, P. (1988). On jump processes in the foreign exchange and stock markets. *Review of Financial Studies*, *1*(4), 427–445.
Kalbfleisch, J. D. & Lawless, J. F. (1988). Estimation of reliability in field-performance studies (with discussion). *Technometrics*, *30*, 365–388.
Kapadia, N. & Zekhnini, M. (2019). Do idiosyncratic jumps matter? *Journal of Financial Economics*, *131*(3), 666–692.
Kheifets, I. L. (2015). Specification tests for nonlinear dynamic models. *Econometrics Journal*, *18*(1), 67–94.
Kiliç, M. & Shaliastovich, I. (2018). Good and bad variance premia and expected returns. *Management Science*, *65*, 2522–2544.
Kitamura, Y., Tripathi, G. & Ahn, H. (2004). Empirical likelihood based inference in conditional moment restriction models. *Econometrica*, *72*, 1667–1714.
Kumar, A. (2009). Who gambles in the stock market? *Journal of Finance*, *64*(4), 1889–1933.
Luenberger, D. G. (1969). *Optimization by vector space methods*. John Wiley & Sons.
Maheu, J. M. & McCurdy, T. H. (2004). News arrival, jump dynamics, and volatility components for individual stock returns. *Journal of Finance*, *59*(2), 755–793.
Maheu, J. M., McCurdy, T. H. & Zhao, X. (2013). Do jumps contribute to the dynamics of the equity premium? *Journal of Financial Economics*, *110*(2), 457–477.
Manski, C. F. (1989). Anatomy of the selection problem. *Journal of Human Resources*, *24*, 343–360.
Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press, Cambridge, MA, USA.
Manski, C. F. & Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, *45*, 1977–1988.
Manski, C. F. & McFadden, D. (1981). Alternative estimators and sample design for discrete choice analysis. In C. F. Manski & D. McFadden (Eds.), *Structural analysis of discrete data with econometric applications* (pp. 2–50). MIT Press, Cambridge, MA, USA.
Markowitz, H. M. (1959). *Portfolio selection: Efficient diversification of investments*. Wiley New York.
Merton, R. C. (1973). An intertemporal capital asset pricing model. *Econometrica*, 867–887.
Nam, K., Pyun, C. S. & Arize, A. C. (2002). Asymmetric mean-reversion and contrarian profits: ANST-GARCH approach. *Journal of Empirical Finance*, *9*(5), 563–588.
Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, *59*, 347–370.
Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In G. S. Maddala, C. R. Rao & H. D. Vinod (Eds.), *Handbook of statistics, vol. 11* (pp. 2111–2245). Elsevier, The Netherlands.
Newey, W. K. & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics, vol. IV* (pp. 2111–2245). Elsevier, The Netherlands.
Newey, W. K. & Steigerwald, D. G. (1997). Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models. *Econometrica*, 587–599.
Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, *75*, 237–249.

Owen, A. B. (2001). *Empirical likelihood*. Chapman & Hall/CRC, New York, NY, USA.
Owen, A. B. (2013). Self-concordance for empirical likelihood. *Canadian Journal of Statistics*, *41*, 387–397.
Owen, A. B. (2017). *A weighted self-concordant optimization for empirical likelihood* [Manuscript available at https://statweb.stanford.edu/~owen/empirical/countnotes.pdf].
Pagan, A. & Ullah, A. (1999). *Nonparametric econometrics*. Cambridge University Press, Cambridge, UK.
Palandri, A. (2015). Do negative and positive equity returns share the same volatility dynamics? *Journal of Banking & Finance*, *58*, 486–505.
Park, Y.-H. (2016). The effects of asymmetric volatility and jumps on the pricing of VIX derivatives. *Journal of Econometrics*, *192*(1), 313–328.
Patil, G. P. & Rao, C. R. (1977). The weighted distributions: A survey of their applications. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 383–405). North Holland, Amsterdam, The Netherlands.
Patil, G. P. & Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, *34*, 179–189.
Patton, A. J. (2006). Modeling asymmetric exchange rate dependence. *International Economic Review*, *47*(2), 527–556.
Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, *160*(1), 246–256.
Patton, A. J. & Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, *97*(3), 683–697.
Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford University Press.
Pelagatti, M. M. (2009). Modelling good and bad volatility. *Studies in Nonlinear Dynamics & Econometrics*, *13*(1).
Powell, J. L. (1994). Estimation of semiparametric models. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics, vol. iv* (pp. 2443–2521). Elsevier, The Netherlands.
Qin, J. (1993). Empirical likelihood in biased sample problems. *Annals of Statistics*, *21*, 1182–1196.
Quesenberry, C. P. & Jewell, N. P. (1986). Regression analysis based on stratified samples. *Biometrika*, *73*, 605–614.
Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*, 846–866.
Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica*, *55*, 875–891.
Rockinger, M. & Jondeau, E. (2002). Entropy densities with an application to autoregressive conditional skewness and kurtosis. *Journal of Econometrics*, *106*(1), 119–142.
Romano, J. P. & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, *73*(4), 1237–1282.
Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, *39*, 577–591.
Schumann, M. & Tripathi, G. (2018). Convexity of probit weights. *Statistics and Probability Letters*, *143*, 81–85.
Scott, A. J. & Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *Journal of The Royal Statistical Society, Series B*, *48*, 170–182.
Severini, T. A. & Tripathi, G. (2001). A simplified approach to computing efficiency bounds in semiparametric models. *Journal of Econometrics*, *102*, 23–66.
Severini, T. A. & Tripathi, G. (2013). Semiparametric efficiency bounds for microeconometric models: A survey. *Foundations and Trends in Econometrics*, *6*, 163–397.
Shiller, R. (2015). Online data: U.S. stock markets 1871–present and CAPE ratio [Accessed: 2018-07-11].

Sklar, A. (1959). *Fonctions de répartition à n dimensions et leurs marges* (tech. rep.). Université Paris 8.
Stambaugh, R. F., Yu, J. & Yuan, Y. (2015). Arbitrage asymmetry and the idiosyncratic volatility puzzle. *Journal of Finance*, *70*(5), 1903–1948.
Strebulaev, I. A., Whited, T. M. et al. (2012). Dynamic models and structural estimation in corporate finance. *Foundations and Trends in Finance*, *6*(1–2), 1–163.
Tauchen, G. & Zhou, H. (2011). Realized jumps on financial markets and predicting credit spreads. *Journal of Econometrics*, *160*(1), 102–118.
Todorov, V. & Tauchen, G. (2011). Volatility jumps. *Journal of Business & Economic Statistics*, *29*(3), 356–371.
Tripathi, G. (1999). A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters*, *63*, 1–3.
Tripathi, G. (2011a). Generalized method of moments (GMM) based inference with stratified samples when the aggregate shares are known. *Journal of Econometrics*, *165*, 258–265.
Tripathi, G. (2011b). Moment based inference with stratified data. *Econometric Theory*, *27*, 47–73.
Tripathi, G. & Kitamura, Y. (2003). Testing conditional moment restrictions. *Annals of Statistics*, *31*, 2059–2095.
Tversky, A. & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quarterly Journal of Economics*, *106*(4), 1039–1061.
Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, *10*, 616–620.
Vardi, Y. (1985). Empirical distributions in selection biased models. *Annals of Statistics*, *13*, 178–203.
Varrette, S., Bouvry, P., Cartiaux, H. & Georgatos, F. (2014). Management of an academic HPC cluster: The UL experience. *Proceedings of the 2014 International Conference on High Performance Computing and Simulation (HPCS 2014)*, 959–967.
Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 307–333.
White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 1–25.
Wooldridge, J. M. (1994). Chapter 45: Estimation and inference for dependent processes. *Handbook of econometrics* (pp. 2639–2738). Elsevier.
Wooldridge, J. M. (1999). Asymptotic properties of weighted M-estimators for variable probability samples. *Econometrica*, *67*, 1385–1406.
Wooldridge, J. M. (2001). Asymptotic properties of weighted M-estimators for standard stratified samples. *Econometric Theory*, *17*, 451–470.
Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd). MIT Press, Cambridge, MA, USA.
Zakoian, J.-M. (1994). Threshold heteroskedastic models. *Journal of Economic Dynamics and Control*, *18*(5), 931–955.