PML vs minimum χ^2: the comeback*

Dante Amengual
CEMFI, Casado del Alisal 5, E-28014 Madrid, Spain
<amengual@cemfi.es>

Gabriele Fiorentini
Università di Firenze and RCEA, Viale Morgagni 59, I-50134 Firenze, Italy
<gabriele.fiorentini@unifi.it>

Enrique Sentana
CEMFI, Casado del Alisal 5, E-28014 Madrid, Spain
<sentana@cemfi.es>

October 2022
Revised: March 2023

Abstract

Arellano (1989a) showed that valid equality restrictions on covariance matrices could result in efficiency losses for Gaussian PMLEs in simultaneous equations models. We revisit his two-equation example using finite normal mixtures PMLEs instead, which are also consistent for mean and variance parameters regardless of the true distribution of the shocks. Because such mixtures provide good approximations to many distributions, we relate the asymptotic variance of our estimators to the relevant semiparametric efficiency bound. Our Monte Carlo results indicate that they systematically dominate MD, and that the version that imposes the valid covariance restriction is more efficient than the unrestricted one.

Keywords: Covariance restrictions, Distributional misspecification, Efficiency bound, Finite normal mixtures, Partial adaptivity, Sieves.

JEL: C30, C36.

*We would like to thank Manuel Arellano, as well as audiences at CEMFI, NYU, the UNIA Workshop on Empirical Microeconomics and Applied Econometrics, and the Bologna-Waseda Time Series Workshop for useful comments. The suggestions of Stéphane Bonhomme (the editor) and a referee have also improved the paper. Of course, the usual caveat applies. The first and third authors gratefully acknowledge financial support from the Spanish Ministry of Science and Innovation through grant PID2021-128963NB-I00, while the second one is thankful to MIUR through the PRIN project "High-dimensional time series for structural macroeconomic analysis in times of pandemic".

1 Introduction

Maximum likelihood and minimum chi-square methods have been competing for the estimator throne for a long time. At the turn of the 19th century, Legendre (1805) and Gauss (1809) put forward least squares estimation as a Gaussian-based alternative to Laplace's (1774) least absolute deviation method, which relied on his eponymous distribution. Almost a century later, Pearson proposed not only the method of moments (see Pearson (1894)), but also the chi-square criterion in the context of matching theoretical and empirical frequencies (see Pearson (1900)). In turn, the development of maximum likelihood estimation (MLE) by Fisher (1922, 1925) was one of the most important achievements in 20th century statistics. Under standard regularity conditions, MLE asymptotically achieves the Cramér-Rao lower bound (see Cramér (1946) and Rao (1945)), which makes it at least as good as any minimum χ^2 estimator. In addition, it achieves second-order efficiency after a bias correction (see Rao (1961)). Moreover, the imposition of valid equality restrictions on the parameters systematically leads to efficiency gains (see Rothenberg (1973)).

However, not everybody was convinced (see Neyman and Scott (1948) on the incidental parameter problem, as well as the inconsistent MLE examples in Basu (1955), Kraft and Le Cam (1956) and Bahadur (1958)), and minimum χ^2 methods remained popular. In fact, Berkson (1980) argued that ML was often just a special case of minimum χ^2, and not necessarily the best one. Soon afterwards, White (1982), building on earlier work by Huber (1967), and Gouriéroux, Monfort and Trognon (1984) studied the properties of Pseudo MLEs, characterising their consistency and general inefficiency. Arellano (1989a) put another nail on the ML coffin by showing that valid equality restrictions could result in efficiency losses for Gaussian PMLEs. Arguably, the wooden stake to the heart was driven by Newey and Steigerwald (1997), who described the inconsistency of non-Gaussian PMLE procedures under distributional misspecification. Since then, graduate students with non-Bayesian teachers learn the normal distribution only, and Gaussian PMLE is just an example of Hansen's (1982) GMM. In this paper, though, we argue that non-Gaussian PMLE, like a B-movie vampire, deserves a second life (or death).

We do so by revisiting the two-equation textbook example in Arellano (1989a),[1] except that instead of basing PMLE on the Gaussian distribution, as he did, we use discrete mixtures of normals. The reason is twofold. First, Fiorentini and Sentana (2023) show that, under standard regularity conditions, such estimators are consistent for the conditional mean and variance

[1] Surprisingly, Arellano (1989a), which should be mentioned in all graduate econometric textbooks, has received very few citations: Pollock (1988), Islam (1993), Monés and Ventura (1996), Calzolari, Fiorentini and Sentana (2004), and Sentana (2005), plus a handful of self-citations, and two more which really meant to cite Arellano (1989b).

parameters regardless of the true distributions of the shocks to the model and the number of mixture components, thereby nesting the results for Gaussian PMLE in Gouriéroux, Monfort and Trognon (1984) while simultaneously avoiding the concerns raised by Newey and Steigerwald (1997). Second, finite normal mixtures with a sufficiently large number of components can provide good approximations to many distributions (see Nguyen et al (2020)), so it is reasonable to conjecture that PMLEs based on them may get close to achieving the semiparametric (SP) efficiency bound, and therefore exploit the potential adaptivity of some of the parameters when it exists, at least asymptotically.[2]

The rest of the paper is organised as follows. Section 2 introduces the example in Arellano (1989a) and summarises his main results. Then, in section 3 we extend those results to the entire parameter vector, derive the relevant semiparametric efficiency bounds, and use them to benchmark the different estimators, including the PMLEs based on finite Gaussian mixtures. Next, section 4 contains the results of our extensive Monte Carlo experiments while section 5 concludes. Proofs and auxiliary results are relegated to the appendices.

2 The example

Consider the following textbook example:

$$y_1 = \gamma + \alpha y_2 + \beta z_1 + u_1, \tag{1}$$

$$y_2 = \mu_0 + \mu_1 z_1 + \mu_2 z_2 + u_2, \tag{2}$$

with

$$\left(\begin{array}{c} u_1 \\ u_2 \end{array} \right) \Bigg| z_1, z_2 \sim D\left[\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array} \right) \right].$$

As is well known, the unrestricted Gaussian PMLE of α and β coincides with the IV estimator that uses a constant, z_1 and z_2 as instruments in the first equation. In turn, the restricted Gaussian PMLE that imposes $\sigma_{12} = 0$ coincides with the OLS estimator of the first equation.

When the joint conditional distribution of u_1 and u_2 is Gaussian, OLS is at least as efficient as IV, which justifies the Durbin-Wu-Hausman test.[3] But Arellano's (1989a) seemingly counterintuitive result says that when the true conditional distribution is not Gaussian, IV may be more efficient than OLS for α and β even though $\sigma_{12} = 0$. Specifically, he showed that IV will

[2] See Fiorentini and Sentana (2022) for a related discussion in the context of structural VARs.

[3] Wu (1973) compared OLS with IV in linear single equation models to assess regressor exogeneity unaware that Durbin (1954) had already suggested this. Hausman (1978) provided a procedure with far wider applicability.

beat OLS if and only if

$$\mu_{22} \geq 1 + \rho^{-2}_{y_2 z_2 . z_1}, \tag{3}$$

where

$$\mu_{22} = E\left(\frac{u_1^2}{\sigma_1^2} \frac{u_2^2}{\sigma_2^2} \middle| z_1, z_2 \right)$$

is the co-kurtosis coefficient between the two structural shocks and $\rho_{y_2 z_2 . z_1}$ is the correlation coefficient between y_2 and z_2 after partialling out the effect of z_1. Intuitively, μ_{22} affects the correct sandwich version of the asymptotic covariance matrix of the OLS estimators of the slope parameters.

Appendix A contains detailed expressions for the asymptotic variances of the OLS and IV estimators of α and β. We have used those expressions to create Figure 1, which displays in $(\rho_{y_2 z_2 . z_1}, \mu_{22})$ space (minus one plus) the ratio of the asymptotic variances of the OLS and IV estimators of α for positive values of $\rho_{y_2 z_2 . z_1}$.[4] We do so for the special case in which the R^2 of equation (2) coincides with $\rho^2_{y_2 z_2 . z_1}$, which allows this parameter to vary freely from 0 to 1.[5] As expected, OLS is more/less efficient than IV to the left/right of the boundary line (3).

This figure also shows the locus of $(\rho_{y_2 z_2 . z_1}, \mu_{22})$ combinations for which the IV estimator of α reaches its maximum asymptotic efficiency relative to the corresponding OLS estimator in this set-up, which is given by the curve

$$\rho^2_{y_2 z_2 . z_1} = \frac{\mu_{22}}{2\left(\mu_{22} - 1\right)}.$$

Further increases in $\rho_{y_2 z_2 . z_1}$ for a given μ_{22} result in decreases in relative efficiency, with OLS and IV becoming indistinguishable as $\rho_{y_2 z_2 . z_1} \rightarrow 1$, in which case z_2 becomes a perfect instrument for y_2.

In this context, Arellano's (1989a) proposed solution is to replace Gaussian PMLE by Minimum Distance (MD) estimators, a special case of minimum chi-square methods popularised in econometrics by Malinvaud (1970). The rationale is as follows. Let $\boldsymbol{\theta} = (\gamma, \alpha, \beta, \mu_0, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)'$ denote the vector of structural parameters. Given that the reduced form of model (2) is

$$\begin{aligned}
\left(\begin{array}{c} y_1 \\ y_2 \end{array} \right) \Bigg| z_1, z_2 &\sim D\left[\boldsymbol{\mu}(z_1, z_2; \boldsymbol{\theta}), \boldsymbol{\Omega}(z_1, z_2; \boldsymbol{\theta})\right] && (4) \\
\boldsymbol{\mu}(z_1, z_2; \boldsymbol{\theta}) &= \left[\begin{array}{c} (\gamma + \alpha\mu_0) + (\beta + \alpha\mu_1) z_1 + \alpha\mu_2 z_2 \\ \mu_0 + \mu_1 z_1 + \mu_2 z_2 \end{array} \right] && (5) \\
\boldsymbol{\Omega}(z_1, z_2; \boldsymbol{\theta}) &= \left(\begin{array}{cc} \sigma_1^2 + \alpha^2\sigma_2^2 + 2\sigma_{12}\alpha & \alpha\sigma_2^2 + \sigma_{12} \\ \alpha\sigma_2^2 + \sigma_{12} & \sigma_2^2 \end{array} \right), && (6)
\end{aligned}$$

[4] The plot would be the mirror image of Figure 1 for negative values.

[5] As we shall see in Proposition 1 below, though, this special case is such that, asymptotically, the diffference between the IV and OLS estimators affects α exclusively.

which is exactly identified, the unrestricted MD estimator coincides with IV, which is Indirect Least Squares. Then, Arellano (1989a) shows that imposing the restriction $\sigma_{12} = 0$ leads to an overidentified optimal MD procedure (weakly) more efficient than both IV and OLS for α and β.

This optimal MD estimator requires an asymptotic covariance of the reduced form parameter estimators which recognises that the third- and fourth-order multivariate cumulants of u_1 and u_2 are not usually 0 when they are jointly non-normally distributed.

Appendix A also contains detailed expressions for the asymptotic variances of the optimal MD estimators of α and β. We have used those expressions to create Figure 2, which depicts in $(\rho_{y_2 z_2 . z_1}, \mu_{22})$ space (minus one plus) the ratio of the asymptotic variance of the restricted optimal MD of α to the asymptotic variance of either the OLS estimator (to the left of (3)) or the IV one (to its right) in the same set up as Figure 1. As can be seen, the efficiency gains are relatively small over the displayed range, and they vanish when either the partial correlation goes to 0 or 1 or the co-kurtosis term goes to 0.[6]

The predictable reaction of a fervent ML believer to Figures 1 and 2 would be to argue that condition (3) requires the combination of a very good instrument (a high $\rho^2_{y_2 z_2 . z_1}$) with a substantial amount of non-normality (a large μ_{22}), in which case the Gaussian assumption would be very inappropriate. For example, a joint Student t distribution for u_1 and u_2 cannot satisfy this condition when the number of degrees of freedom is six or more, and the requirement becomes increasingly difficult for poor instruments.

A naïve ML solution would be to assume that u_1 and u_2 follow a bivariate Student t distribution to estimate the model parameters, which should dominate MD. In this respect, we have used the expressions in Appendix A to create Figures 3a and 3b, which display in $(\rho_{y_2 z_2 . z_1}, \mu_{22})$ space (minus one plus) the ratio of the asymptotic variances of the t-based MLE of α and β that impose $\sigma_{12} = 0$ to the asymptotic variances of the corresponding restricted optimal MD. As can be seen, these figures confirm that ML does indeed dominate MD in this case.

The problem with this naïve approach is that if the assumed joint distribution is incorrect, the resulting PMLEs may be inconsistent, as forcefully argued by Newey and Steigerwald (1997).

However, this does not mean that all parameters will be inconsistently estimated. Specifically, Proposition 3 in Fiorentini and Sentana (2019) implies that the unrestricted t-based PMLEs of α and β are always consistent irrespective of the true distribution. Similarly, their Proposition 1 implies that the restricted t-based PMLEs of α and β will remain consistent when the conditional

[6] Again, Proposition 1 below implies that the differences in asymptotic variances between the MD, IV and OLS estimators affect α exclusively in the special case in which the (squared) partial correlation of y_2 and z_2 given z_1 coincides with the R^2 in the regression of y_2 on z_1 and z_2

distribution of $\sigma_1^{-1}u_1$ and $\sigma_2^{-1}u_2$ is elliptical even though it does not coincide with the distribution assumed for estimation purposes. Besides, it may be possible to obtain two-step consistent estimators in closed-form along the lines of Fiorentini and Sentana (2019).

More importantly, Fiorentini and Sentana (2023) show that all parameters will always be consistently estimated if one assumes for estimation purposes that u_1 and u_2 follow a finite mixture of bivariate normals regardless of the true distribution of those innovations and the number of components of the mixture, as long as the shape parameters are simultaneous estimated with the mean and variance parameters.[7] Thus, the consistency of the Gaussian PMLE is just a special case.

The ability of finite Gaussian mixtures to approximate many other distributions mentioned in the introduction suggests that we may be able relate these finite mixture PMLEs to SP estimators which simply exploit the independence of the shocks and the conditioning variables without making any parametric assumptions. For that reason, in the next section we take SP estimators as our benchmark to study:

1. the efficiency of the OLS, IV, MD and correct ML estimators relative to SP ones,

2. the relative efficiency of restricted and unrestricted versions of these SP estimators, and

3. the relative efficiency of finite mixture-based PMLEs relative to SP estimators

in the context of model (2).

3 Theoretical analysis

3.1 Minimum distance revisited

Although the main focus of the analysis in Arellano (1989a) was α and β, it is of some interest to study the asymptotic efficiency of the optimal MD estimators of the remaining structural model parameters relative to their OLS and IV counterparts. Given that the number of different bivariate cumulants of orders three and four is 4 and 5, respectively, we focus on the special case in which the joint distribution of the (standardised) structural shocks conditional on the instruments is spherical, or $s(\mathbf{0}, \mathbf{I}_2, \boldsymbol{\eta})$ for short, where $\boldsymbol{\eta}$ is the possibly infinite vector of shape parameters. More formally,

Assumption 1

$$\left.\frac{u_1}{\sigma_1}, \frac{u_2}{\sigma_2}\right| z_1, z_2; \boldsymbol{\theta}, \boldsymbol{\eta} \sim i.i.d.\ s(\mathbf{0}, \mathbf{I}_2, \boldsymbol{\eta}) \tag{7}$$

[7] On the other hand, if the shape parameters of the mixture are fixed, then Theorem 7 in Gouriéroux, Monfort and Trognon (1984) guarantees the inconsistency of the resulting estimators except in the Gaussian limiting case.

To simplify the expressions further, we are going to follow Appendix B in Fiorentini and Sentana (2019) and re-parametrise the unrestricted covariance matrix of the structural residuals as

$$\left(\begin{array}{cc} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array}\right) = \sigma^2 \left(\begin{array}{cc} 1 & 0 \\ \psi_{12} & 1 \end{array}\right) \left(\begin{array}{cc} e^{\omega} & 0 \\ 0 & e^{-\omega} \end{array}\right) \left(\begin{array}{cc} 1 & \psi_{12} \\ 0 & 1 \end{array}\right), \tag{8}$$

where ψ_{12} is the coefficient in the least squares projection of u_2 on u_1, and σ^2 and ω the geometric mean of their variances and the natural log of the ratio of the standard deviations of these shocks, respectively, under the maintained assumption that they are uncorrelated.[8] Let $\boldsymbol{\theta}^{\dagger} = (\gamma, \alpha, \beta, \mu_0, \mu_1, \mu_2, \omega, \sigma^2)'$ denote the vector of structural parameters implied by (8) under the restriction $\psi_{12} = 0$. Using the expressions for the Jacobian linking $\boldsymbol{\theta}^{\dagger}$ and $\boldsymbol{\theta}$ in (A17), we can then show under standard regularity conditions that:

Proposition 1 *Let (τ_1, τ_2) and $(\sigma_{z_1}^2, \sigma_{z_2}^2, \sigma_{z_1 z_2})$ denote the means, variances and covariance of z_1 and z_2. If Assumption 1 holds, then:*
(a) The difference between the asymptotic covariance matrices of the OLS and MD estimators of $\boldsymbol{\theta}^{\dagger}$, $\hat{\boldsymbol{\theta}}_{LS}^{\dagger}$ and $\hat{\boldsymbol{\theta}}_{MD}^{\dagger}$, respectively, is positive semidefinite of rank 1 at most, with a basis for its image given by

$$\left\{-[\mu_0 + (\tau_2 - \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\tau_1)\mu_2], 1, -\mu_1 + \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\mu_2, \mathbf{0}_{1\times 5}\right\}, \tag{9}$$

and a basis for its kernel by

$$\left[1, \mu_0 + (\tau_2 - \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\tau_1)\mu_2, 0, \mathbf{0}_{1\times 5}\right], \tag{10}$$

$$\left[\mu_1 + \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\mu_2, 0, \mu_0 + (\tau_2 - \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\tau_1)\mu_2, \mathbf{0}_{1\times 5}\right] \tag{11}$$

and

$$(\mathbf{0}_{5\times 3}, \mathbf{I}_5). \tag{12}$$

(b) The difference between the asymptotic covariance matrices of the IV and MD estimators of $\boldsymbol{\theta}^{\dagger}$, $\tilde{\boldsymbol{\theta}}_{IV}^{\dagger}$ and $\hat{\boldsymbol{\theta}}_{MD}^{\dagger}$, respectively, is positive semidefinite of rank 1 at most, with the same basis for image and kernel.
(c) The difference between the asymptotic covariance matrices of the OLS and IV estimators of $\boldsymbol{\theta}^{\dagger}$, $\hat{\boldsymbol{\theta}}_{LS}^{\dagger}$ and $\tilde{\boldsymbol{\theta}}_{IV}^{\dagger}$, respectively, is positive/negative semidefinite of rank 1 depending of condition (3), with exactly the same basis for image and kernel.

This proposition considerably sharpens the results in Arellano (1989a) for the special case of spherically symmetric disturbances by showing that the asymptotic efficiency gains concentrate in a single linear combination of the parameters of the first equation γ, α and β given by (9). In contrast, any other linear combination of the parameters orthogonal to this one does not generate any efficiency gains. Specifically, the parameters of the second equation and the residual variances are estimated just as efficiently by the three procedures.

[8] More generally, $\sigma^2 = \sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$ and $\omega = \ln\left(\sigma_1/\sqrt{\sigma_2^2 - \sigma_{12}^2/\sigma_1^2}\right)$.

3.2 Semiparametric estimation and efficiency bounds

The optimal instruments theory of Chamberlain (1987) implies that Arellano's (1989a) MD estimator achieves the SP efficiency bound which exploits the correct specification of the conditional mean and variance functions for y_1 and y_2 in the reduced form model (2) when the joint third- and fourth-order cumulants of u_1 and u_2 conditional on z_1 and z_2 are constant. However, if this last maintained assumption is true, then one can in principle obtain an even more efficient MD estimator of the model parameters after augmenting it with equations for the third- and fourth-order cumulants of the reduced-form residuals under the assumption that the joint cumulants of u_1 and u_2 conditional on z_1 and z_2 are constant up to the eighth-order.

In fact, the results in Bickel et al (1993) allow us to obtain the SP efficiency bound that exploits that the joint distribution of u_1 and u_2 is independent of z_1 and z_2. Moreover, we can also consider a restricted version of this SP bound under the maintained assumption that (7) holds, as in Hodgson and Vorkink (2003), which will be bigger in the usual positive semidefinite sense. Henceforth, we shall refer to this bound and its associated estimator by the abbreviation SS, reserving SP for the one which does not impose sphericity.

An interesting question in this context is the possibility that some but not all of the parameters of model (2) can be partially adaptively estimated, in the sense that their SP estimators are as asymptotically efficient as the infeasible ML estimators which exploit the information of the true distribution of the shocks, including the values of their shape parameters. The following proposition provides a precise answer to this question under sphericity for the restricted estimators that impose $\sigma_{12} = 0$:

Proposition 2 *If Assumption 1 holds, then:*
(a) The difference between the asymptotic covariance matrices of the restricted SS and infeasible ML estimators of $\boldsymbol{\theta}^{\dagger}$*,* $\hat{\boldsymbol{\theta}}^{\dagger}_{SS}$ *and* $\hat{\boldsymbol{\theta}}^{\dagger}_{ML}(\bar{\boldsymbol{\eta}})$*, respectively, is positive semidefinite of rank 1 at most, with a basis for its image given by* $(\mathbf{0}_{1\times 7}, 1)$*, and a basis for its kernel by* $(\mathbf{I}_7, \mathbf{0}_{7\times 1})$*.*
(b) The difference between the asymptotic covariance matrices of the restricted SP and infeasible ML estimators of $\boldsymbol{\theta}^{\dagger}$*,* $\hat{\boldsymbol{\theta}}^{\dagger}_{SP}$ *and* $\hat{\boldsymbol{\theta}}^{\dagger}_{ML}(\bar{\boldsymbol{\eta}})$*, respectively, is positive semidefinite of rank 5 at most, with a basis for its image given by* $(1, \mathbf{0}_{1\times 7})$*,* $(0, -1, \mu_1 + \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\mu_2, \mathbf{0}_{1\times 5})$*,* $(\mathbf{0}_{1\times 3}, 1, \mathbf{0}_{1\times 4})$ *and* $(\mathbf{0}_{2\times 6}, \mathbf{I}_2)$*, and a basis for its kernel by* $(0, \mu_1 + \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\mu_2, 1, \mathbf{0}_{1\times 5})$ *and* $(\mathbf{0}_{2\times 4}, \mathbf{I}_2, \mathbf{0}_{2\times 2})$*.*
(c) The difference between the asymptotic covariance matrices of the MD and SP estimators of $\boldsymbol{\theta}^{\dagger}$*,* $\hat{\boldsymbol{\theta}}^{\dagger}_{MD}$ *and* $\hat{\boldsymbol{\theta}}^{\dagger}_{SP}$*, respectively, is positive semidefinite of rank 4 at most, with a basis for its image given by* $(\mathbf{0}_{2\times 1}, \mathbf{I}_2, \mathbf{0}_{2\times 5})$ *and* $(\mathbf{0}_{2\times 4}, \mathbf{I}_2, \mathbf{0}_{2\times 2})$*, and a basis for its kernel by* $(\mathbf{0}_{2\times 6}, \mathbf{I}_2)$*,* $(1, \mu_0 + \mu_1\tau_1 + \mu_2\tau_2, \tau_1, \mathbf{0}_{1\times 5})$ *and* $(\mathbf{0}_{1\times 3}, 1, \tau_1, \tau_2, \mathbf{0}_{1\times 2})$*.*

The first part of the proposition implies that all the structural model parameters except the overall residual scale σ^2 can be (partially) adaptively estimated by the SS estimator, as expected from Proposition 12 in Fiorentini and Sentana (2021).

More interestingly, the second part of the proposition implies that in addition to μ_1 and μ_2, the coefficient of the linear projection of y_1 onto a constant and z_1, which is given by

$$\beta + (\mu_1 + \sigma_{z_1 z_2}\sigma_{z_1}^{-2}\mu_2)\alpha,$$

will be adaptively estimated by the restricted SP estimator. In this respect, a very important by-product of this proposition is that the model parameters that can be partially adaptively estimated often continue to be consistently estimated under distributional misspecification of the innovations, as shown by Fiorentini and Sentana (2019, 2021) in the context of multivariate location-scale models such as (2). We will revisit this issue in the Monte Carlo section.

Finally, the last part of the proposition says that the variances of the structural-form residuals, as well as the intercepts in the reduced-form regressions of y_1 and y_2 on a constant and the demeaned values of z_1 and z_2, which are given by $\gamma + \tau_1(\beta + \alpha\mu_1) + \tau_2(\alpha\mu_2)$ and $\mu_0 + \tau_1\mu_1 + \tau_2\mu_2$, respectively, are asymptotically equally efficiently estimated by the MD and SP estimators. More importantly, it also says that the efficiency gains are concentrated in the four slope coefficients of the two structural equations.

It would be tedious but otherwise straightforward to extend Propositions 1 and 2 to the case in which the distribution of the shocks conditional on z_1 and z_2 is not spherical as a function of the four third-order and five fourth-order cumulants of u_1 and u_2. In fact, there is one important instance in which those higher-order cumulants would be unnecessary for the comparisons. Specifically, we can use Proposition 13.2 in Fiorentini and Sentana (2021) to prove that, subject to regularity, both the parameters of the unrestricted covariance matrix of the reduced-form residuals and the intercepts in the reduced-form regressions of y_1 and y_2 on a constant and the demeaned values of z_1 and z_2 will be as efficiently estimated by the IV estimator and the unrestricted SP estimator, while the slopes will always be adaptively estimated, just as in the second part of Proposition 2 above. The reason is twofold. First, the information matrix, the feasible parametric efficiency bound, the SP bound, and the usual Gaussian sandwich formula become block-diagonal between those reduced-form parameters and the four structural slope coefficients α, β, μ_1 and μ_2. In turn, this block-diagonality leads to a saddle-point characterisation of the asymptotic efficiency of the SP estimator of $\boldsymbol{\theta}$, with the slope coefficients being adaptive and the others only reaching the efficiency of the Gaussian PMLE.

3.3 Efficiency gains from the equality constraint

It is also of interest to analyze the effects of imposing the covariance restriction $\sigma_{12} = 0$ on the different estimators we have considered:

Proposition 3 *If Assumption 1 holds, then:*
(a) The difference between the asymptotic covariance matrices of the unrestricted and restricted infeasible ML estimators of $\boldsymbol{\theta}^{\dagger}$*,* $\tilde{\boldsymbol{\theta}}^{\dagger}_{ML}$ *and* $\hat{\boldsymbol{\theta}}^{\dagger}_{ML}$*, respectively, is positive semidefinite of rank 1 at most, with the basis for its image given by (9), and a basis for its kernel by (10), (11) and (12).*
(b) The difference between the asymptotic covariance matrices of the unrestricted and restricted SS estimators of $\boldsymbol{\theta}^{\dagger}$*,* $\tilde{\boldsymbol{\theta}}^{\dagger}_{SS}$ *and* $\hat{\boldsymbol{\theta}}^{\dagger}_{SS}$*, respectively, is positive semidefinite of rank 1 at most, with the basis for its image given by (9), and a basis for its kernel by (10), (11) and (12).*
(c) The difference between the asymptotic covariance matrices of the unrestricted and restricted SP estimators of $\boldsymbol{\theta}^{\dagger}$*,* $\tilde{\boldsymbol{\theta}}^{\dagger}_{SP}$ *and* $\hat{\boldsymbol{\theta}}^{\dagger}_{SP}$*, respectively, is positive semidefinite of rank 1 at most, with the basis for its image given by (9), and a basis for its kernel by (10), (11) and (12).*

Therefore, when one uses "efficient" estimators, the imposition of the valid equality constraint $\sigma_{12} = 0$ always leads to (weak) efficiency gains for exactly the same linear combination of the parameters of the first structural equation for which optimal MD leads to an efficiency gain relative to both OLS and IV. In fact, it is straightforward to generalise (a) so that it applies to the feasible parametric ML estimators of $\boldsymbol{\theta}^{\dagger}$ which simultaneously estimate the finite vector of shape parameters $\boldsymbol{\eta}$, as well as to the ML estimators of these parameters themselves. This is in contrast to the seemingly counterintuitive result in Arellano (1989a), which simply reflects the fact that OLS does not use the optimal MD weighting matrix in the non-normal case.

3.4 Finite mixtures as sieves

Finally, we study the extent to which PMLEs based on finite mixtures of normals with an increasing number of components could constitute a proper sieves-type SP procedure, as we argued in the introduction.

We do so first when the shocks to model (2) conditional on z_1 and z_2 follow a bivariate Student t with 0 means, unit standard deviations, no correlation and 5 degrees of freedom but whose parameters are estimated by finite scale mixture-based log-likelihood functions with $K = 2, 3$ and 4 components. For comparison purposes, we consider four different benchmarks that impose the restriction $\sigma_{12} = 0$: (i) the MLE based on the correctly specified log-likelihood function that fixes the number of degrees of freedom to 5, (ii) the SS estimator, (ii) the OLS estimator, and (iv) the optimal MD estimator.

We compute the expected value of the Hessian and outer product of the score of the scale mixture-based PMLEs by means of large sample averages of the analytical expressions in Fiorentini and Sentana (2021) evaluated at the true values of the mean and variance parameters in $\boldsymbol{\theta}$ and the pseudo true values of the shape parameters, which we numerically obtain from samples of millions of simulated observations.

The results, which we report in Table 1, show that the scale mixture-based PMLEs of all the model parameters except the overall residual scale σ^2 quickly approach the asymptotic efficiency of the infeasible MLE despite the fact that no finite scale mixture of normals can approximate the unbounded higher-order moments, tail behaviour or non-linear tail dependence of a multivariate Student t. In fact, although panel (a) in Figure 3 of Gallant and Tauchen (1999) clearly illustrates that a more complex misspecified model does not necessarily lead to more efficient estimators because one is not simply adding new elements to the score, but also changing the pseudo-true values of the shape parameters at which one evaluates the original components of the score, we find that the efficiency improvements occur monotonically.[9] As a result, it seems that the covariance matrix of the errors in the least squares projection of the scores of the true model onto the scores of the mixture-based log-likelihood becomes smaller and smaller as K increases (see Proposition 7 in Calzolari, Fiorentini and Sentana (2004)).

In contrast, the asymptotic variances of the scale mixture-based PMLEs of σ^2 coincides with the asymptotic variances of the OLS estimators irrespective of the number of components, which reflects (i) the block diagonality of the different asymptotic covariance matrices in Proposition 12.2 of Fiorentini and Sentana (2021) because the determinant of (8) is precisely σ^4, and (ii) the fact that the ML estimators of the mean in a scale mixture of K gammas is numerically the same regardless of K, as explained in Fiorentini and Sentana (2023).

We then conduct a similar exercise when u_1 and u_2 conditional on z_1 and z_2 follow a bivariate asymmetric Student t with 0 means, unit standard deviations, no correlation, negative tail dependence and the same μ_{22} as in the symmetric case. We estimate the unrestricted model parameters using general finite mixture-based log-likelihood functions with $K = 2, 3$ and 4 components, and consider as benchmarks the following three unrestricted estimators: infeasible MLE, SP, and IV. In this case, we compute the expected value of the Hessian and outer product of the score of the mixture-based PMLEs using large sample averages of the theoretical expressions in Amengual, Fiorentini and Sentana (2023) evaluated at the true values of the mean and variance parameters and the pseudo true values of the shape parameters obtained from very large samples of simulated observations.

The results we report in Table 2 show that the mixture-based PMLEs of the slope parameters approach the asymptotic efficiency of the infeasible MLE despite the fact that no finite mixture of normals can approximate the unbounded higher-order moments, tail behaviour or non-linear tail dependence of a multivariate asymmetric Student t. Again, we find that the efficiency improvements occur monotonically. In contrast, the asymptotic variances of the mixture-based PMLEs of the intercepts and covariance matrix of the reduced form in mean-deviation form

[9] In this respect, the efficieny gains of any $K > 1$ relative to $K = 1$ should be easy to prove formally because the ML estimators of the unconditional mean and covariance matrix of the mixture model coincide regardless of K.

coincide with the asymptotic variances of the corresponding IV estimators irrespective of the number of components, which reflects the fact that the ML estimators of the mean vector and covariance matrix in mixtures of K normals are numerically the same for any $K \geq 1$ (see also the discussion at the end of section 3.2).

4 Monte Carlo analysis

In previous sections, we have derived several asymptotic results regarding the relative efficiency of the LS, IV and MD estimators, as well as the finite mixture-based PMLEs, the SS estimators, and the feasible and infeasible MLEs. In this section, in contrast, we make use of an extensive Monte Carlo simulation exercise to asses their small sample behaviour.

4.1 Design

We consider three different parameter configurations:

a. $\mu_{22} = 3$ and $\rho_{y_2 z_2 . z_1} = (\mu_{22} - 1)^{-\frac{1}{2}} = 1/\sqrt{2} \simeq 0.71$, which is such that the IV and OLS estimators of α and β have the same asymptotic efficiency (see the solid line in Figure 1);

b. $\mu_{22} = 3$ and $\rho_{y_2 z_2 . z_1} = 2^{-\frac{1}{2}} \sqrt{\mu_{22}/(\mu_{22} - 1)} = \sqrt{3}/2 \simeq 0.87$, which corresponds to the dotted line in Figure 1; and

c. $\mu_{22} = 7/3$ and $\rho_{y_2 z_2 . z_1} = (\mu_{22} - 1)^{-\frac{1}{2}} = \sqrt{3}/2 \simeq 0.87$, which is another case of equal efficiency of IV and OLS, but with lower co-kurtosis.[10]

As for the distribution of the structural shocks, we consider four non-Gaussian possibilities in which (u_1, u_2) follow a:

1. Student t distribution with $\nu = 5$ or $\nu = 5.5$ degrees of freedom corresponding to $\mu_{22} = 3$ and $\mu_{22} = 7/3$, respectively;

2. scale mixture of two normals in which the higher variance component has probability $\lambda = .05$ and the ratio of the variances is either $\varkappa = 0.094$ or $\varkappa = 0.122$ corresponding to $\mu_{22} = 3$ and $\mu_{22} = 7/3$, respectively;

3. asymmetric Student t distribution with negative tail dependence $\mathbf{b} = (-1, -1)'$ but degrees of freedom $\nu = 9.65$ or $\nu = 10.38$, respectively;

[10] We do not consider the case in which $\mu_{22} = 7/3$ and $\rho_{y_2 z_2 . z_1} = .5\sqrt{\mu_{22}/(\mu_{22} - 1)}$ because the efficiency of IV relative to OLS for α is just 1.02 in that case.

4. location-scale mixture of two normals in which the higher variance component has probability $\lambda = .05$, μ_{22} is as in *1.*, and the marginal skewness of u_1 and u_2 is as in *3.*, which is achieved with

$$\boldsymbol{\delta} = \begin{pmatrix} -1.01 \\ -1.06 \end{pmatrix} \text{ or } \boldsymbol{\delta} = \begin{pmatrix} -1.16 \\ -1.24 \end{pmatrix} \quad \text{and} \quad \aleph_L = \begin{pmatrix} 0.32 & 0 \\ 0 & 0.32 \end{pmatrix} \text{ or } \begin{pmatrix} 0.38 & 0 \\ 0 & 0.38 \end{pmatrix},$$

respectively (see Appendix D for further details on this parametrisation).

For illustrative purposes, we display the joint densities and contours for standardised versions of these distributions in comparison to the bivariate spherical Gaussian distribution in Figures 4 and 5 for the spherically symmetric and general cases, respectively.

In all simulated samples the exogenous variables $\mathbf{z} = (z_1, z_2)'$ are generated according to a bivariate Student t distribution with 8 degrees of freedom with mean vector $\boldsymbol{\tau} = (1, 1)'$ and an identity variance covariance matrix.[11]

Next, for each choice of the partial correlation $\rho_{y_2 z_2 . z_1}$ mentioned above, we choose

$$R_2^2 = \frac{2\rho_{y_2 z_2 . z_1}}{1 + \rho_{y_2 z_2 . z_1}} \text{ and } \rho_{y_2 z_1} = \rho_{y_2 z_2} = \sqrt{\frac{R_2^2 - \rho^2_{y_2 z_2 . z_1}}{1 - \rho_{y_2 z_2 . z_1}}},$$

which guarantees that (i) $\rho^2_{y_2 z_2 . z_1} \leq R_2^2 \leq 1$, and (ii) the two slope coefficients of the second equations coincide. If we fix the variance of both y_1 and y_2 to 1 without loss of generality, these restrictions implicitly determine the variance of the error term of the second equation as $\sigma_2^2 = 1 - R_2^2$. We also impose the same balancing restriction on the slopes of the first equation by choosing

$$\alpha = \beta = \sqrt{\frac{(1 + \rho_{y_2 z_1}) R_1^2}{2}}.$$

Then, we fix R_1^2 to 0.5, which implies $\sigma_1^2 = 1/2$, an arbitrary choice that simply scales the asymptotic variances of all the different estimators of α and β by the same amount $(1 - R_1^2)$.[12] Finally, we choose the values of the intercepts γ and μ_0 so that $E(y_1) = E(y_2) = 1$ (see Appendix C for further details).

4.2 Simulation results

We simulate $10,000$ samples of length $N = 250$ and $N = 1,000$ for each of the above designs. For each simulated sample, we compute the IV, LS and optimal MD estimators, together with unrestricted and restricted versions of PMLE estimators that use either a discrete mixture of

[11] Notice that the choice of $\sigma_{z_1 z_2} = 0$ considerably simplifies some of the eigenvectors in Propositions 1, 2 and 3. For example the linear combination that according to Proposition 2.b can be adaptively estimated by the SP estimator and consistenly estimated by a distributionally misspecified ML estimator becomes $\beta + \mu_1 \alpha$.

[12] In design *a.*, we then have $R_2^2 = 2/3$, $\sigma_2^2 = 1/3$, $\gamma = 0.20$, $\alpha = \beta = 0.40$, $\mu_0 = 0.16$, and $\mu_1 = \mu_2 = 0.58$. In turn, in designs *b* and *c.*, $R_2^2 = 6/7$, $\sigma_2^2 = 1/7$, $\gamma = 0.22$, $\alpha = \beta = 0.39$, $\mu_0 = -0.31$, and $\mu_1 = \mu_2 = 0.66$.

two normals –UPML(mn), RPML(mn)– or a Student t distribution –UPML(t) and RPML(t). In both cases, we simultaneously estimate the shape parameters. Finally, we also compute a two-step SS estimator that starting from the consistent OLS estimator, $\boldsymbol{\theta}_{LS}$, carries out one BHHH iteration using the efficient spherically symmetric semiparametric score estimated nonparametrically. Specifically, we compute the standardised reduced form residuals

$$\hat{\mathbf{v}}^* = \hat{\boldsymbol{\Omega}}^{-\frac{1}{2}} \left[\mathbf{y} - \boldsymbol{\mu}\left(z_1, z_2, \boldsymbol{\theta}_{LS}\right)\right],$$

where $\hat{\boldsymbol{\Omega}}^{-\frac{1}{2}}$ denotes the inverse of the Cholesky decomposition of the sample covariance matrix of the reduced form residuals $[\mathbf{y} - \boldsymbol{\mu}\left(z_1, z_2, \boldsymbol{\theta}_{LS}\right)]$, define $\hat{\varsigma} = \hat{\mathbf{v}}^{*\prime}\hat{\mathbf{v}}^*$ and estimate nonparametrically the density of $\zeta = \varsigma^{1/3}$, $g(\zeta)$, and its derivative, $g'(\zeta)$, using a Gaussian kernel with the usual Silverman (1986) "rule-of-thumb" bandwidth. The change of variable formula then yields

$$\delta(\varsigma) = \frac{-2}{3\zeta^2}\frac{g'(\zeta)}{g(\zeta)} + \frac{4}{3\zeta^3},$$

which we use to compute the semiparametric efficient score using expression (C30) in the Supplemental Appendix C of Fiorentini and Sentana (2021) by subtracting

$$\mathbf{W}_s(\boldsymbol{\theta}_{LS}) \left[\delta(\varsigma)\frac{\varsigma}{2} - 1 - \frac{2}{(4\kappa+2)}\left(\frac{\varsigma}{2} - 1\right)\right]$$

from the nonparametric score, where κ denotes the coefficient of multivariate excess kurtosis (see Mardia (1970) for details) and $\mathbf{W}_s(\boldsymbol{\theta})$ is defined in Appendix A.5.

We display the finite sample results by means of the box-plots in Figures 6 to 11, which concentrate on α and β, the two parameters of interest. Figures 6 to 8 show the Monte Carlo results for 250 observations for cases *a.*, *b.* and *c.*, respectively, while Figures 9 to 11 contain the results for $1,000$ observations in the same order.

Our findings indicate that OLS is better in finite samples than what the asymptotic theory suggests because the sample co-kurtosis coefficient is downward biased for μ_{22}. In fact, the asymptotic efficiency of the IV estimator of α relative to LS can only be observed in panels b and d of Figure 10 when the sample length is large and the distribution of the shocks is either a spherical or a general finite mixture of normals, which is when there seems to be a lower small sample bias for μ_{22}.

They also confirm that optimal MD dominates both OLS and IV in finite samples, but the need to estimate third- and fourth-order multivariate cumulants to compute the optimal weighting matrix handicaps it somewhat (see Altonji and Segal (1996) for analogous results in the context of optimal GMM estimators when the shocks are fat tailed)

Our results also indicate that non-Gaussian PML based on a restrictive parametric distribution like the Student t or a discrete scale mixture of normals works well when the true distribution is spherical, but it generates inconsistencies otherwise when we impose the constraint $\sigma_{12} = 0$. Notice, though, that the unrestricted estimators are always consistent for the slope parameters while the restricted estimators seem to be consistent for $\beta + \mu_1\alpha$ despite being inconsistent for both α and β, which is in line with our theoretical discussion following Proposition 2.

In turn, the performance of the two-step SS estimators is very similar to the performance of the corresponding parametric estimators, although their finite sample variances are larger than what the asymptotic theory predicts. Specifically, the consistency pattern of the restricted and unrestricted SS estimators is almost identical.

More importantly, we find that non-Gaussian PMLEs based on a flexible distribution like a general finite mixture of normals works well in practice regardless of the true distribution, systematically dominating MD. In addition, the version that imposes the valid covariance restriction $\sigma_{12} = 0$ is always more efficient than the unrestricted one.

5 Directions for further research

As we mentioned at the end of section 3.2, it would be useful to generalise our theoretical results dropping the assumption of spherical symmetry. Similarly, and although we have seen that our proposed finite mixture-based PMLEs get close to achieving the SP efficiency bound both under sphericity and in general, an obvious extension of our Monte Carlo experiments would be to consider standard two-step SP estimators that starting from a consistent estimator such as OLS carry out one BHHH iteration using the efficient SP score estimated non-parametrically without imposing spherical symmetry. The curse of dimensionality in estimating multivariate densities, though, might further reduce the theoretical advantages of this method in finite samples.

Another worthwhile exercise would be to extend the analysis in this paper to the general simultaneous equation model with an arbitrary numbers of endogenous variables and instrumental ones considered by Arellano (1989a). Aside from involving more complex analytical expressions than in the bivariate example we have considered, the main practical complication would be that the number of free parameters of a standardised multivariate mixture increases with the square of the cross-sectional dimension, as we explain in Appendix D.

Last, but not least, deriving a formal result that shows that finite Gaussian-mixture based PMLEs may provide a proper sieve ML estimator when the number of components increases at a suitable rate constitutes a particularly interesting avenue for further research.

References

Altonji, J.G and Segal, L.M. (1996): "Small-sample bias in GMM estimation of covariance structures", *Journal of Business and Economic Statistics*, 14, 353–366

Amengual, D., Fiorentini, G. and Sentana, E. (2022): "Tests for random coefficient variation in vector autoregressive models", in J.J. Dolado, L. Gambetti and C. Matthes (eds.) *Essays in honour of Fabio Canova, Advances in Econometrics* 44B, 1–35, Emerald.

Amengual, D., Fiorentini, G. and Sentana, E. (2023): "Information matrix tests for Gaussian mixtures", mimeo, CEMFI.

Arellano (1989a): "On the efficient estimation of simultaneous equations with covariance restrictions", *Journal of Econometrics* 42, 247–265.

Arellano (1989b): "An efficient GLS estimator of triangular models with covariance restrictions", *Journal of Econometrics* 42, 267–273.

Bahadur, R.R. (1958): "Examples of inconsistency of maximum likelihood estimates", *Sankhya* 20, 207–210.

Basu, D. (1955): "An inconsistency of the method of maximum likelihood", *Annals of Mathematical Statistics* 26, 144–145.

Berkson, J. (1980): "Minimum chi-square, not maximum likelihood!", *Annals of Statistics* 8, 457–487.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1993): *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins.

Calzolari, G., Fiorentini, G. and Sentana, E. (2004): "Constrained indirect estimation", *Review of Economic Studies* 71, 945–973.

Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions", *Journal of Econometrics* 34, 305–334.

Cramér, H. (1946): *Mathematical methods of statistics*, Princeton University Press.

Durbin, J. (1954): "Errors in variables", *Review International Statistical Institute* 22, 23–32.

Fisher, R. A. (1922): "On the mathematical foundations of theoretical statistics", *Philosophical Transactions of the Royal Society London Series A* 222, 309–368.

Fisher, R. A. (1925): "Theory of statistical estimation", *Proceedings of the Cambridge Philosophical Society* 22, 700–725.

Fiorentini, G. and Sentana, E. (2019): "Consistent non-Gaussian pseudo maximum likelihood estimators", *Journal of Econometrics* 213, 321–358.

Fiorentini, G. and Sentana, E. (2021): "Specification tests for non-Gaussian maximum likelihood estimators", *Quantitative Economics* 12, 683–742.

Fiorentini, G. and Sentana, E. (2022): "Discrete mixtures of normals pseudo maximum likelihood estimators of structural vector autoregressions", forthcoming in the *Journal of Econometrics.*

Fiorentini, G. and Sentana, E. (2023): "Consistent estimation with finite mixtures", mimeo, CEMFI.

Gallant, A.R. and Tauchen, G. (1999): "The relative efficiency of method of moments estimators", *Journal of Econometrics* 92, 149–172.

Gauss, C.F. (1809): *Theoria motus corporum coelestium*, Perthes.

Gouriéroux, C., Monfort A. and Trognon, A. (1984): "Pseudo maximum likelihood methods: theory", *Econometrica* 52, 681–700.

Hansen, L.P. (1982): "Large sample properties of generalized method of moments estimators", *Econometrica* 50, 1029–1054.

Hausman, J. (1978): "Specification tests in econometrics", *Econometrica* 46, 1273-1291.

Hodgson and Vorkink (2003): "Efficient estimation of conditional asset pricing models", *Journal of Business and Economic Statistics*, 21, 269–283.

Huber, P. J. (1967): "The behavior of maximum likelihood estimates under nonstandard conditions", in *Proceedings of the V Berkeley Symposium in Mathematical Statistics and Probability* 1, 221–233, University of California Press.

Islam, N. (1993): "Estimation of dynamic models from panel data", unpublished Ph.D. dissertation, Harvard University.

Kraft, C.H., and Le Cam, L.M. (1956): "A remark on the roots of the maximum likelihood equation", *Annals of Mathematical Statistics* 27, 1174–1177.

Laplace, P-S. (1774): "Mémoire sur la probabilité des causes par les évènements", *Mémoires de l'Academie Royale des Sciences Presentés par Divers Savan* 6, 621–656.

Legendre, A-M. (1805): *Nouvelles méthodes pour la détermination des orbites des comètes*, F. Didot.

Magnus, J.R. and Neudecker, H. (2019): *Matrix differential calculus with applications in statistics and econometrics*, 3rd edition, Wiley.

Malinvaud, E. (1970): *Statistical methods in econometrics*, 2nd edition, North Holland.

Mardia, K.V. (1970): "Measures of multivariate skewness and kurtosis with applications", *Biometrika* 57, 519–530.

Monés, M.A. and Ventura, E. (1996): "Saving decisions and fiscal incentives", *Applied Economics* 28, 1105–1117

Newey, W.K. and Steigerwald, D.G. (1997): "Asymptotic bias for quasi-maximum-likelihood estimators in conditional heteroskedasticity models", *Econometrica* 65, 587–99.

Neyman, J., and Scott, E.L. (1948): "Consistent estimation from partially consistent observations", *Econometrica* 16, 1–32.

Nguyen, T.T., Nguyen, H.D., Chamroukhi, F. and McLachlan, G.J. (2020): "Approximation by finite mixtures of continuous density functions that vanish at infinity", *Cogent Mathematics & Statistics* 7, 1750861.

Pearson, K. (1894): "Contributions to the mathematical theory of evolution", *Philosophical Transactions of the Royal Society London Series A* 185, 71–110.

Pearson, K. (1900): "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science Series 5* 50, 157–175.

Pollock, D.S.G. (1988): "The estimation of linear stochastic models with covariance restrictions", *Econometric Theory* 4, 403–427.

Rao, C.R. (1945): "Information and the accuracy attainable in the estimation of statistical parameters", *Bulletin of the Calcutta Mathematical Society* 37, 81–89.

Rao, C.R. (1961): "Asymptotic efficiency and limiting information", *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability* 1, 531–546, University of California Press.

Rothenberg, T.J. (1973): *Efficient estimation with a priori information*, Cowles Foundation Monograph 23, Yale.

Sentana, E. (2005): "Least squares predictions and mean-variance analysis", *Journal of Financial Econometrics* 3, 56–78.

Silverman, B.W. (1986): *Density estimation for statistics and data analysis*, Chapman & Hall.

White, H. (1982): "Maximum likelihood estimation of misspecified models", *Econometrica* 50, 1–25.

Wu, D-M. (1973): "Alternative tests of independence between stochastic regressors and disturbances", *Econometrica.* 41, 733–750.

Appendices

A Asymptotic covariance matrices

A.1 Instrumental Variables (IV)

Let $\mathbf{v}_i = (v_{1i}, v_{2i})'$ denote the reduced form innovations

$$\mathbf{v}_i = \mathbf{y}_i - \mathbf{C}\mathbf{z}_i = \mathbf{B}^{-1}\mathbf{u}_i,$$

where $\mathbf{y}_i = (y_{1i}, y_{2i})'$ and $\mathbf{z}_i = (z_{1i}, z_{2i})'$, so that $E(\mathbf{v}_i|\mathbf{z}_i) = \mathbf{0}$ and $V(\mathbf{v}_i|\mathbf{z}_i) = \mathbf{B}^{-1}\mathbf{\Sigma}\mathbf{B}'^{-1} = \mathbf{\Omega}$, with

$$\mathbf{B}'^{-1} = \begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}.$$

In this context, the unrestricted Gaussian PMLE of α and β coincides with the IV estimator that uses a constant, z_1 and z_2 as instruments in the first equation. To consider both equations at once, let $\boldsymbol{\vartheta} = (\boldsymbol{\theta}', \sigma_{12})'$ and

$$\mathbf{Z}^U_{di}(\boldsymbol{\vartheta}) = [\mathbf{Z}^U_{li}(\boldsymbol{\vartheta}), \mathbf{Z}^U_{si}(\boldsymbol{\vartheta})], \tag{A1}$$

where

$$\begin{aligned}
\mathbf{Z}^U_{li}(\boldsymbol{\vartheta}) &= \frac{\partial \boldsymbol{\mu}'_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} \mathbf{\Omega}^{-\frac{1}{2}\prime}(\boldsymbol{\vartheta}), \\
\mathbf{Z}^U_{si}(\boldsymbol{\vartheta}) &= \frac{1}{2}\frac{\partial vec'[\mathbf{\Omega}(\boldsymbol{\vartheta})]}{\partial \boldsymbol{\vartheta}} \left[\mathbf{\Omega}^{-\frac{1}{2}\prime}(\boldsymbol{\vartheta}) \otimes \mathbf{\Omega}^{-\frac{1}{2}\prime}(\boldsymbol{\vartheta})\right],
\end{aligned}$$

$$\frac{\partial \boldsymbol{\mu}'_i(\boldsymbol{\vartheta})}{\partial \boldsymbol{\vartheta}} = \begin{pmatrix} 1 & 0 \\ \mu_0 + \mu_1 z_{1i} + \mu_2 z_{2i} & 0 \\ z_{2i} & 0 \\ \alpha & 1 \\ \alpha z_{1i} & z_{1i} \\ \alpha z_{2i} & z_{2i} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \frac{\partial vec'[\mathbf{\Omega}(\boldsymbol{\vartheta})]}{\partial \boldsymbol{\vartheta}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 2(\alpha\sigma_2^2 + \sigma_{12}) & \sigma_2^2 & \sigma_2^2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ \alpha^2 & \alpha & \alpha & 1 \\ 2\alpha & 1 & 1 & 0 \end{pmatrix},$$

and

$$\mathbf{\Omega}^{-\frac{1}{2}}(\boldsymbol{\vartheta}) = \begin{pmatrix} \frac{1}{\sqrt{\sigma_1^2+\alpha^2\sigma_2^2+2\alpha\sigma_{12}}} & 0 \\ -\frac{\alpha\sigma_2^2+\sigma_{12}}{\sigma_1^2+\alpha^2\sigma_2^2+2\alpha\sigma_{12}} \Big/ \sqrt{\frac{\sigma_1^2\sigma_2^2-\sigma_{12}^2}{\sigma_1^2+\alpha^2\sigma_2^2+2\alpha\sigma_{12}}} & \sqrt{\frac{\sigma_1^2+\alpha^2\sigma_2^2+2\alpha\sigma_{12}}{\sigma_1^2\sigma_2^2-\sigma_{12}^2}} \end{pmatrix}$$

is the inverse of the (lower) Cholesky decomposition of $\mathbf{\Omega}$.

We can then exploit Proposition C2 in Supplementary Appendix C of Fiorentini and Sentana (2021) to obtain

$$AVar(\sqrt{n}\tilde{\boldsymbol{\vartheta}}_{IV}) = [\mathcal{A}_{U,\boldsymbol{\vartheta}\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})]^{-1}\mathcal{B}_{U,\boldsymbol{\vartheta}\boldsymbol{\vartheta}}(\boldsymbol{\vartheta},\boldsymbol{\varrho})[\mathcal{A}_{U,\boldsymbol{\vartheta}\boldsymbol{\vartheta}}(\boldsymbol{\vartheta})]^{-1}, \tag{A2}$$

where

$$\mathcal{A}_{U,\boldsymbol{\vartheta\vartheta}}(\boldsymbol{\vartheta}) = E\left[\mathbf{Z}_{di}^{U}(\boldsymbol{\vartheta})\mathcal{K}(\mathbf{0})\mathbf{Z}_{di}^{U\prime}(\boldsymbol{\vartheta})\right] \quad \text{and} \quad \mathcal{B}_{U,\boldsymbol{\vartheta\vartheta}}(\boldsymbol{\vartheta},\boldsymbol{\varrho}) = E\left[\mathbf{Z}_{di}^{U}(\boldsymbol{\vartheta})\mathcal{K}^{\mathbf{v}}(\boldsymbol{\vartheta},\boldsymbol{\varrho})\mathbf{Z}_{di}^{U\prime}(\boldsymbol{\vartheta})\right],$$

with

$$\mathcal{K}^{\mathbf{v}}(\boldsymbol{\vartheta},\boldsymbol{\varrho}) = V[\mathbf{e}_{di}(\boldsymbol{\vartheta},\mathbf{0})] = \begin{bmatrix} \mathbf{I}_2 & \boldsymbol{\Phi}^{\mathbf{v}}(\boldsymbol{\vartheta},\boldsymbol{\varrho}) \\ \boldsymbol{\Phi}^{\mathbf{v}\prime}(\boldsymbol{\vartheta},\boldsymbol{\varrho}) & \boldsymbol{\Upsilon}^{\mathbf{v}}(\boldsymbol{\vartheta},\boldsymbol{\varrho}) \end{bmatrix}, \tag{A3}$$

$\boldsymbol{\Phi}^{\mathbf{v}}(\boldsymbol{\vartheta}_0,\boldsymbol{\varrho}_0) = E[\mathbf{v}_i^* vec'(\mathbf{v}_i^*\mathbf{v}_i^{*\prime})]$, $\boldsymbol{\Upsilon}^{\mathbf{v}}(\boldsymbol{\vartheta}_0,\boldsymbol{\varrho}_0) = E[vec(\mathbf{v}_i^*\mathbf{v}_i^{*\prime} - \mathbf{I}_2)vec'(\mathbf{v}_i^*\mathbf{v}_i^{*\prime} - \mathbf{I}_2)]$ and $\mathbf{v}_i^* = \boldsymbol{\Omega}^{-1/2}\mathbf{v}_i$, so that $\boldsymbol{\Phi}^{\mathbf{v}}(\mathbf{0}) = \mathbf{0}$ and $\boldsymbol{\Upsilon}^{\mathbf{v}}(\mathbf{0}) = (\mathbf{I}_4 + \mathbf{K}_{22})$ if we use $\boldsymbol{\varrho} = \mathbf{0}$ to denote normality and $\mathbf{K}_{mn}$ for the commutation matrix of orders m and n (see e.g. Magnus and Neudecker (2019)).

Given that the assumption of constant conditional higher-order cumulants applies to the structural model, though, we need to relate the higher-order moments of the reduced form residuals to those of the structural ones. Defining

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{L}_2[\mathbf{B}^{-1}(\boldsymbol{\theta}) \otimes \mathbf{B}^{-1}(\boldsymbol{\theta})]\mathbf{D}_2 = \begin{bmatrix} 1 & 2\alpha & \alpha^2 \\ 0 & 1 & \alpha \\ 0 & 0 & 1 \end{bmatrix},$$

where $\mathbf{L}_2$ and $\mathbf{D}_2$ are the elimination and duplication matrices of order 2, respectively (see Magnus and Neudecker (2019)), we will have that

$$E[\mathbf{v}_i vec'(\mathbf{v}_i\mathbf{v}_i')] = -\mathbf{B}^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}(\boldsymbol{\theta})^{\frac{1}{2}}\boldsymbol{\Phi}^{\mathbf{u}}(\boldsymbol{\varrho})[\boldsymbol{\Sigma}(\boldsymbol{\theta})^{\frac{1}{2}\prime} \otimes \boldsymbol{\Sigma}(\boldsymbol{\theta})^{\frac{1}{2}\prime}]\mathbf{F}'(\boldsymbol{\theta})$$

and

$$E[vec(\mathbf{v}_i\mathbf{v}_i' - \mathbf{I}_2)vec'(\mathbf{v}_i\mathbf{v}_i' - \mathbf{I}_2)] = \mathbf{F}(\boldsymbol{\theta})[\boldsymbol{\Sigma}(\boldsymbol{\theta})^{\frac{1}{2}} \otimes \boldsymbol{\Sigma}(\boldsymbol{\theta})^{\frac{1}{2}}]\boldsymbol{\Upsilon}^{\mathbf{u}}(\boldsymbol{\varrho})[\boldsymbol{\Sigma}(\boldsymbol{\theta})^{\frac{1}{2}\prime} \otimes \boldsymbol{\Sigma}(\boldsymbol{\theta})^{\frac{1}{2}\prime}]\mathbf{F}'(\boldsymbol{\theta}),$$

where $\boldsymbol{\Phi}^{\mathbf{u}}(\boldsymbol{\varrho}_0) = E[\mathbf{u}_i^* vec'(\mathbf{u}_i^*\mathbf{u}_i^{*\prime})]$, $\boldsymbol{\Upsilon}^{\mathbf{u}}(\boldsymbol{\varrho}_0) = E[vec(\mathbf{u}_i^*\mathbf{u}_i^{*\prime} - \mathbf{I}_2)vec'(\mathbf{u}_i^*\mathbf{u}_i^{*\prime} - \mathbf{I}_2)]$ and $\mathbf{u}_i^* = \boldsymbol{\Sigma}^{-1/2}\mathbf{u}_i$.

After some tedious calculations, it is straightforward to prove that

$$AVar(\sqrt{n}\tilde{\alpha}_{IV}) = \frac{\sigma_1^2\sigma_{z_1}^2}{\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}$$

and

$$AVar(\sqrt{n}\tilde{\beta}_{IV}) = \frac{\sigma_1^2(\mu_1^2\sigma_{z_1}^2 + \mu_2^2\sigma_{z_2}^2 + 2\mu_1\mu_2\sigma_{z_1z_2})}{\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}.$$

For our purposes, it is convenient to rewrite these expressions as

$$AVar(\sqrt{n}\tilde{\alpha}_{IV}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)}{(1-R_2^2)\rho_{y_2z_2.z_1}^2}$$

and

$$AVar(\sqrt{n}\tilde{\beta}_{IV}) = \frac{R_2^2(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)}{(1-R_2^2)\rho_{y_2z_2.z_1}^2},$$

where R_1^2 and R_2^2 are the population coefficients of determination in equations (1) and (2), respectively, and $\rho_{y_2 z_2 . z_1}$ the correlation coefficient between y_2 and z_2 after partialling out the effect of z_1.

A.2 Ordinary Least Squares (LS)

As mentioned in Section 2, the restricted Gaussian PMLE that imposes $\sigma_{12} = 0$ coincides with the OLS estimator of the first equation. To consider both equations at once, let

$$\mathbf{Z}_{di}^{R}(\boldsymbol{\theta}) = (\mathbf{I}_8, \mathbf{0}_{8\times 1})\mathbf{Z}_{di}^{U}(\boldsymbol{\vartheta}, 0). \tag{A4}$$

Then, analogous calculations to the ones in the previous subsection imply that

$$AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{LS}) = \mathcal{A}_{R,\boldsymbol{\theta\theta}}^{-1}(\boldsymbol{\theta})\mathcal{B}_{R,\boldsymbol{\theta\theta}}(\boldsymbol{\theta},\boldsymbol{\varrho})\mathcal{A}_{R,\boldsymbol{\theta\theta}}^{-1}(\boldsymbol{\theta}), \tag{A5}$$

where

$$\mathcal{A}_{R,\boldsymbol{\theta\theta}}(\boldsymbol{\theta}) = E\left[\mathbf{Z}_{di}^{R}(\boldsymbol{\theta})\mathcal{K}^{\mathbf{v}}(\mathbf{0})\mathbf{Z}_{di}^{R\prime}(\boldsymbol{\theta})\right] \quad \text{and} \quad \mathcal{B}_{R,\boldsymbol{\theta\theta}}(\boldsymbol{\theta},\boldsymbol{\varrho}) = E\left[\mathbf{Z}_{di}^{R}(\boldsymbol{\theta})\mathcal{K}^{\mathbf{v}}(\boldsymbol{\theta},\boldsymbol{\varrho})\mathbf{Z}_{di}^{R\prime}(\boldsymbol{\theta})\right],$$

After some straightforward calculations, it is easy to show that

$$AVar(\sqrt{n}\hat{\alpha}_{LS}) = \frac{\sigma_1^2\sigma_{z_1}(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)\mu_2^2}{[\mu_2^2\sigma_{z_1z_2}^2 - \sigma_{z_1}^2(\sigma_2^2 + \mu_2^2\sigma_{z_2}^2)]^2} + \frac{\sigma_1^2\sigma_2^2\sigma_{z_1}^4\mu_{22}}{[\mu_2^2\sigma_{z_1z_2}^2 - \sigma_{z_1}^2(\sigma_2^2 + \mu_2^2\sigma_{z_2}^2)]^2}$$

and

$$\begin{aligned} AVar(\sqrt{n}\hat{\beta}_{LS}) \quad = \quad & \frac{\sigma_1^2\{\sigma_{z_2}^2\mu_2^2[\sigma_{z_1}^4\mu_1^2 + 2\sigma_{z_1}(\sigma_2^2 + \sigma_{z_1z_2}\mu_1\mu_2) - \sigma_{z_1z_2}^2\mu_2^2] + \sigma_{z_1}^2\sigma_{z_2}^4\mu_2^4\}}{[\sigma_2^2\sigma_{z_1}^2 + \mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)]^2} \\ & + \frac{\sigma_1^2(\sigma_2^2 + \mu_1\mu_2\sigma_{z_1z_2})[\sigma_2^2\sigma_{z_1}^2 - \sigma_{z_1z_2}(\sigma_{z_1}^2\mu_1\mu_2 + 2\sigma_{z_1z_2}\mu_2^2)]}{[\sigma_2^2\sigma_{z_1}^2 + \mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)]^2} \\ & + \frac{\sigma_1^2\sigma_2^2(\sigma_{z_1}^2\mu_1 + \sigma_{z_1z_2}\mu_2)^2\mu_{22}}{[\sigma_2^2\sigma_{z_1}^2 + \mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)]^2}. \end{aligned}$$

Again, it is convenient to rewrite these expressions as

$$AVar(\sqrt{n}\hat{\alpha}_{LS}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)[\mu_{22}(1-\rho_{y_2z_2.z_1}^2) + \rho_{y_2z_2.z_1}^2]}{1-R_2^2}$$

and

$$AVar(\sqrt{n}\hat{\beta}_{LS}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)[1 + (\mu_{22}-1)(R_2^2 - \rho_{y_2z_2.z_1}^2)]}{1-R_2^2},$$

A.3 Optimum Minimum Distance (MD)

Let $\mathbf{c} = vec(\mathbf{C})$ and $\boldsymbol{\omega} = vech(\boldsymbol{\Omega})$ denote the parameters of the unrestricted reduced form model. From equations (5)-(6), we will have that

$$\begin{array}{lll} c_{10} = \gamma + \alpha\mu_0, & c_{20} = \mu_0, & \omega_{11} = \sigma_{11} + \alpha^2\sigma_{22} + 2\alpha\sigma_{12}, \\ c_{11} = \beta + \alpha\mu_1, & c_{21} = \mu_1, & \omega_{12} = \alpha\sigma_{22} + \sigma_{12}, \\ c_{12} = \alpha\mu_2, & c_{22} = \mu_2, & \omega_{22} = \sigma_{22}. \end{array},$$

Let $\tilde{\boldsymbol{\phi}}_{LS}= (\tilde{c}_{10}, \tilde{c}_{11}, \tilde{c}_{12}, \tilde{c}_{20}, \tilde{c}_{21}, \tilde{c}_{22}, \tilde{\omega}_{11}, \tilde{\omega}_{12}, \tilde{\omega}_{22})'$ denote their unrestricted Gaussian PML estimators, which coincide with equation by equation OLS. To obtain the asymptotic distributions of these estimators, we need the first derivatives of the conditional mean vector and covariance matrix with respect to the unrestricted reduced form parameters, which are given by

$$\frac{\partial \mathbf{C}\mathbf{z}_i}{\partial \mathbf{c}'} = \mathbf{z}_i' \otimes \mathbf{I}_2 \text{ and } \frac{\partial vec[\boldsymbol{\Omega}(\boldsymbol{\theta})]}{\partial \boldsymbol{\omega}'} = \mathbf{D}_2.$$

In this notation, the contribution to the Gaussian log-likelihood scores for $\mathbf{c}$ and $\boldsymbol{\omega}$ corresponding to observation i will be given by

$$\mathbf{s}_{\mathbf{c}i}(\mathbf{c},\boldsymbol{\omega}) = \mathbf{z}_i \otimes \boldsymbol{\Omega}^{-1}(\boldsymbol{\theta})\mathbf{v}_i(\mathbf{c})$$

and

$$\mathbf{s}_{\boldsymbol{\omega}i}(\mathbf{c},\boldsymbol{\omega}) = \frac{1}{2}\mathbf{D}_2' vec[\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\mathbf{v}_i(\mathbf{c})\mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}].$$

Consequently, the outer product of the scores will be

$$\mathbf{s}_{\mathbf{c}i}(\mathbf{c},\boldsymbol{\omega})\mathbf{s}_{\mathbf{c}i}'(\mathbf{c},\boldsymbol{\omega}) = \mathbf{z}_i\mathbf{z}_i' \otimes \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\mathbf{v}_i(\mathbf{c})\mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1},$$

$$\mathbf{s}_{\boldsymbol{\omega}i}(\mathbf{c},\boldsymbol{\omega})\mathbf{s}_{\mathbf{c}i}'(\mathbf{c},\boldsymbol{\omega}) = \frac{1}{2}\mathbf{D}_2' vec[\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\mathbf{v}_i(\mathbf{c})\mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}][\mathbf{z}_i' \otimes \mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}]$$

and

$$\begin{aligned} \mathbf{s}_{\boldsymbol{\omega}i}(\mathbf{c},\boldsymbol{\omega})\mathbf{s}_{\boldsymbol{\omega}i}'(\mathbf{c},\boldsymbol{\omega}) &= \frac{1}{4}\mathbf{D}_2' vec[\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\mathbf{v}_i(\mathbf{c})\mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}] \\ &\quad \times vec'[\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\mathbf{v}_i(\mathbf{c})\mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} - \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}]\mathbf{D}_2. \end{aligned}$$

Similarly, we can easily adapt the expressions in Amengual, Fiorentini and Sentana (2022) to write the contribution of observation i to the Hessian matrix $\mathbf{h}_{\mathbf{c},\boldsymbol{\omega}i}(\mathbf{c},\boldsymbol{\omega})$ as

$$= -\left\{ \begin{array}{cc} (\mathbf{z}_i\mathbf{z}_i' \otimes \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}) & [\mathbf{z}_i\mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}) \otimes \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}]\mathbf{D}_2 \\ \mathbf{D}_2'[\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\mathbf{v}_i(\mathbf{c})\mathbf{z}_i' \otimes \boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}] & \mathbf{D}_2'\{\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} \otimes [\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}\mathbf{v}_i(\mathbf{c})\mathbf{v}_i'(\mathbf{c})\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1} - \frac{1}{2}\boldsymbol{\Omega}(\boldsymbol{\theta})^{-1}]\}\mathbf{D}_2 \end{array} \right\}.$$

Thus, we have all the ingredients to compute $AVar(\sqrt{n}\tilde{\boldsymbol{\phi}}_{LS})$ using the standard sandwich formula in White (1982) and Gouriéroux, Monfort and Trognon (1984).

On this basis, we can show that the asymptotic variance of Malinvaud's (1970) optimum MD estimator will be given by

$$AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{MD}) = \left\{ \frac{\partial \boldsymbol{\phi}'(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left[AVar(\sqrt{n}\tilde{\boldsymbol{\phi}}_{LS}) \right]^{-1} \frac{\partial \boldsymbol{\phi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right\}^{-1}, \tag{A6}$$

where

$$\boldsymbol{\phi}(\boldsymbol{\theta}) = \begin{pmatrix} c_{10} - \gamma - \alpha\mu_0 \\ c_{11} - \beta - \alpha\mu_1 \\ c_{12} - \alpha\mu_2 \\ c_{20} - \mu_0 \\ c_{21} - \mu_1 \\ c_{22} - \mu_2 \\ \omega_{11} - \sigma_{11} - \alpha^2\sigma_{22} \\ \omega_{12} - \alpha\sigma_{22} \\ \omega_{22} - \sigma_{22} \end{pmatrix}.$$

Specifically, we obtain that

$$AVar(\sqrt{n}\hat{\alpha}_{MD}) = \frac{\sigma_1^2\sigma_{z_1}^2\mu_{22}}{\sigma_{z_1}^2\sigma_2^2 + (\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)\mu_2^2\mu_{22}}$$

and

$$AVar(\sqrt{n}\hat{\beta}_{MD}) = \frac{\sigma_1^2[\sigma_2^2 + (\sigma_{z_1}^2\mu_1^2 + \sigma_{z_2}^2\mu_2^2 + 2\sigma_{z_1z_2}\mu_1\mu_2)\mu_{22}]}{\sigma_{z_1}^2\sigma_2^2 + (\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)\mu_2^2\mu_{22}},$$

which, rewritten in terms of the population coefficients of determination, become

$$AVar(\sqrt{n}\hat{\alpha}_{MD}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)\mu_{22}}{(1-R_2^2)[1+\rho_{y_2z_2.z_1}^2(\mu_{22}-1)]}$$

and

$$AVar(\sqrt{n}\hat{\beta}_{MD}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)[1+R_2^2(\mu_{22}-1)]}{(1-R_2^2)[1+\rho_{y_2z_2.z_1}^2(\mu_{22}-1)]}.$$

A.4 Maximum likelihood with spherical innovations

Invoking Proposition C1 in Supplementary Appendix C of Fiorentini and Sentana (2021), we can obtain the asymptotic variance of the ML estimator that imposes $\sigma_{12} = 0$ as

$$AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{ML}) = \mathcal{I}_R^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta}), \text{ where } \mathcal{I}_R(\boldsymbol{\theta}, \boldsymbol{\eta}) = E[\mathbf{Z}_i^R(\boldsymbol{\theta})\mathcal{M}(\boldsymbol{\eta})\mathbf{Z}_i^{R\prime}(\boldsymbol{\theta})], \tag{A7}$$

$$\mathbf{Z}_i^R(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Z}_{di}^R(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}, \ \mathcal{M}(\boldsymbol{\eta}) = \begin{pmatrix} \mathcal{M}_{ll}(\boldsymbol{\eta}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathcal{M}_{ss}(\boldsymbol{\eta}) & \mathcal{M}_{sr}(\boldsymbol{\eta}) \\ \mathbf{0} & \mathcal{M}_{sr}'(\boldsymbol{\eta}) & \mathcal{M}_{rr}(\boldsymbol{\eta}) \end{pmatrix},$$

$$\mathcal{M}_{ll}(\boldsymbol{\eta}) = \mathrm{M}_{ll}\mathbf{I}_2,$$

$$\mathcal{M}_{ss}(\boldsymbol{\eta}) = \mathrm{M}_{ss}(\mathbf{I}_4 + \mathbf{K}_{22}) + [\mathrm{M}_{ss} - 1]vec(\mathbf{I}_2)vec'(\mathbf{I}_2),$$

$$\mathcal{M}_{sr}(\boldsymbol{\eta}) = vec(\mathbf{I}_2)\mathrm{M}_{sr}$$

and

$$\mathcal{M}_{rr}(\boldsymbol{\eta}) = V[\,\mathbf{e}_{rt}(\boldsymbol{\phi})|\,\boldsymbol{\phi}] = -E[\partial\mathbf{e}_{rt}(\boldsymbol{\phi})/\partial\boldsymbol{\eta}'\big|\,\boldsymbol{\phi}],$$

with

$$\begin{aligned}
\mathrm{M}_{ll} &= E\left[\delta^2(\varsigma_i,\boldsymbol{\eta})\frac{\varsigma_i}{2}\right], \\
\mathrm{M}_{ss} &= 1 + E\left[\frac{\partial\delta(\varsigma_i,\boldsymbol{\eta})}{\partial\varsigma}\left(\frac{\varsigma_i}{2}\right)^2\right], \quad \text{and} \\
\mathrm{M}_{sr} &= -E\left[\frac{\varsigma_i}{2}\frac{\partial\delta(\varsigma_i,\boldsymbol{\eta})}{\partial\boldsymbol{\eta}'}\right].
\end{aligned}$$

Similarly, we can compute the asymptotic variance of the unrestricted ML estimator which also estimates σ_{12} as

$$AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{ML}) = \mathcal{I}_U^{-1}(\boldsymbol{\theta},\boldsymbol{\eta}), \text{ where } \mathcal{I}_U(\boldsymbol{\theta},\boldsymbol{\eta}) = E[\mathbf{Z}_i^U(\boldsymbol{\theta})\mathcal{M}(\boldsymbol{\eta})\mathbf{Z}_i^{U\prime}(\boldsymbol{\theta})], \tag{A8}$$

with

$$\mathbf{Z}_i^U(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Z}_{di}^U(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}.$$

As a consequence,

$$AVar(\sqrt{n}\hat{\alpha}_{ML}) = \frac{\sigma_1^2\sigma_{z_1}^2}{\mathrm{M}_{ss}\sigma_2^2\sigma_{z_1}^2 + \mathrm{M}_{ll}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}$$

and

$$AVar(\sqrt{n}\hat{\beta}_{ML}) = \frac{\sigma_1^2[\mathrm{M}_{ss}\sigma_2^2 + \mathrm{M}_{ll}(\mu_1^2\sigma_{z_1}^2 + \mu_2^2\sigma_{z_2}^2 + 2\mu_1\mu_2\sigma_{z_1z_2})]}{\mathrm{M}_{ll}[\mathrm{M}_{ss}\sigma_2^2\sigma_{z_1}^2 + \mathrm{M}_{ll}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)]}.$$

Analogous calculations using $\mathbf{Z}_t^U(\boldsymbol{\theta})$ in place of $\mathbf{Z}_t^R(\boldsymbol{\theta})$ for the unrestricted ML estimator yield

$$AVar(\sqrt{n}\tilde{\alpha}_{ML}) = \frac{\sigma_1^2\sigma_{z_1}^2}{\mathrm{M}_{ll}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}$$

and

$$AVar(\sqrt{n}\tilde{\beta}_{ML}) = \frac{\sigma_1^2(\mu_1^2\sigma_{z_1} + \mu_2^2\sigma_{z_2}^2 + 2\mu_1\mu_2\sigma_{z_1z_2})}{\mathrm{M}_{ll}\mu_2^2(\sigma_{z_1}\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}.$$

Once again, we can write these expressions as

$$AVar(\sqrt{n}\hat{\alpha}_{ML}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)}{(1-R_2^2)[(1-\rho_{y_2z_2.z_1}^2)\mathrm{M}_{ss} + \rho_{y_2z_2.z_1}^2\mathrm{M}_{ll}]}$$

and

$$AVar(\sqrt{n}\hat{\beta}_{ML}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)[R_2^2\mathrm{M}_{ll} + (1-R_2^2)\mathrm{M}_{ss}]}{(1-R_2^2)\mathrm{M}_{ll}[(1-\rho_{y_2z_2.z_1}^2)\mathrm{M}_{ss} + \rho_{y_2z_2.z_1}^2\mathrm{M}_{ll}]}$$

for the restricted estimator, and as

$$AVar(\sqrt{n}\tilde{\alpha}_{ML}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)}{(1-R_2^2)\rho_{y_2z_2.z_1}^2\mathrm{M}_{ll}}$$

and

$$AVar(\sqrt{n}\tilde{\beta}_{ML}) = \frac{R_2^2(1-R_1^2)(1-\rho^2_{y_2z_2.z_1})}{(1-R_2^2)\rho^2_{y_2z_2.z_1}\mathrm{M}_{ll}}$$

for the unrestricted one.

A.5 Spherically symmetric semiparametric estimator (SS)

From Proposition C3 in Supplementary Appendix C of Fiorentini and Sentana (2021), the spherically symmetric SP efficiency bound is given by

$$\mathring{\mathcal{S}}_j(\boldsymbol{\theta}) = \mathcal{I}_{j,\boldsymbol{\theta\theta}}(\boldsymbol{\phi}) - \mathbf{W}_s^j(\boldsymbol{\theta})\mathbf{W}_s^{j\prime}(\boldsymbol{\theta}) \cdot \left\{[2\mathrm{M}_{ss}-1] - \frac{2}{4\kappa+2}\right\}$$

where

$$\mathbf{W}_s^j(\boldsymbol{\theta}) = \mathbf{Z}_d^j(\boldsymbol{\theta})[\mathbf{0}', vec'(\mathbf{I}_2)]' \quad \text{for} \quad j = R, U,$$

and

$$\mathcal{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta},\boldsymbol{\eta}) = E\left[\mathbf{Z}_{dt}(\boldsymbol{\theta})\mathcal{M}_{dd}(\boldsymbol{\theta},\boldsymbol{\eta})\mathbf{Z}'_{dt}(\boldsymbol{\theta})\right].$$

Under suitable regularity conditions, we have that

$$AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{SS}) = [\mathring{\mathcal{S}}_R(\boldsymbol{\theta})]^{-1} \tag{A9}$$

and

$$AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{SS}) = [\mathring{\mathcal{S}}_U(\boldsymbol{\theta})]^{-1}. \tag{A10}$$

Tedious but otherwise straightforward calculations show that for the restricted estimator that imposes $\sigma_{12} = 0$ we obtain

$$AVar(\sqrt{n}\hat{\alpha}_{SS}) = AVar(\sqrt{n}\hat{\alpha}_{ML}) \text{ and } AVar(\sqrt{n}\hat{\beta}_{SS}) = AVar(\sqrt{n}\hat{\beta}_{ML}),$$

while for the unrestricted one we get

$$AVar(\sqrt{n}\tilde{\alpha}_{SS}) = AVar(\sqrt{n}\tilde{\alpha}_{ML}) \text{ and } AVar(\sqrt{n}\tilde{\beta}_{SS}) = AVar(\sqrt{n}\tilde{\beta}_{ML}).$$

A.6 Maximum likelihood with general innovations

If we use Proposition D3 in Supplementary Appendix D of Fiorentini and Sentana (2021), we can obtain the asymptotic variance of the ML estimator that imposes $\sigma_{12} = 0$ by computing

$$AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{ML}) = \mathcal{I}_{GR}^{-1}(\boldsymbol{\theta},\boldsymbol{\varrho}), \text{ where } \mathcal{I}_{GR}(\boldsymbol{\theta},\boldsymbol{\varrho}) = E[\mathbf{Z}_i^{GR}(\boldsymbol{\theta})\mathcal{M}(\boldsymbol{\varrho})\mathbf{Z}_i^{GR\prime}(\boldsymbol{\theta})],$$

where

$$\mathbf{Z}_{di}^{GR}(\boldsymbol{\theta}) = [\mathbf{Z}_{li}^R(\boldsymbol{\theta}), \mathbf{Z}_{si}^{GR}(\boldsymbol{\theta})], \tag{A11}$$

$$\mathbf{Z}_{si}^{GR}(\boldsymbol{\theta}) = \frac{\partial vec'[\boldsymbol{\Omega}^{\frac{1}{2}}(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} \left[\mathbf{I}_2 \otimes \boldsymbol{\Omega}^{-\frac{1}{2}\prime}(\boldsymbol{\theta})\right],$$

$$\frac{\partial vec'[\boldsymbol{\Omega}^{\frac{1}{2}}(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \frac{\alpha\sigma_2^2}{\sqrt{\sigma_1^2+\alpha\sigma_2^2}} & \frac{\sigma_1^2\sigma_2^2}{(\sigma_1^2+\alpha\sigma_2^2)^{3/2}} & 0 & -\frac{\alpha}{\sigma_1^2}\left(\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\alpha\sigma_2^2}\right)^{3/2} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \frac{1}{2\sqrt{\sigma_1^2+\alpha\sigma_2^2}} & -\frac{\alpha\sigma_2^2}{2(\sigma_1^2+\alpha\sigma_2^2)^{3/2}} & 0 & \frac{\alpha^2\sigma_2^2}{2(\sigma_1^2+\alpha^2\sigma_1^2\sigma_2^2)}\sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\alpha\sigma_2^2}} \\ \frac{\alpha^2}{2\sqrt{\sigma_1^2+\alpha\sigma_2^2}} & \frac{2\alpha\sigma_1^2+\alpha^3\sigma_2^2}{2(\sigma_1^2+\alpha\sigma_2^2)^{3/2}} & 0 & \frac{\sigma_1^2}{2(\sigma_1^2\sigma_2^2+\alpha^2\sigma_2^2)}\sqrt{\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\alpha\sigma_2^2}} \end{bmatrix},$$

and

$$\mathcal{M}(\boldsymbol{\varrho}) = \begin{bmatrix} \mathcal{M}_{ll}(\boldsymbol{\varrho}) & \mathcal{M}_{ls}(\boldsymbol{\varrho}) & \mathcal{M}_{lr}(\boldsymbol{\varrho}) \\ \mathcal{M}'_{ls}(\boldsymbol{\varrho}) & \mathcal{M}_{ss}(\boldsymbol{\varrho}) & \mathcal{M}_{sr}(\boldsymbol{\varrho}) \\ \mathcal{M}'_{lr}(\boldsymbol{\varrho}) & \mathcal{M}'_{sr}(\boldsymbol{\varrho}) & \mathcal{M}_{rr}(\boldsymbol{\varrho}) \end{bmatrix},$$

with

$$\mathcal{M}_{ll}(\boldsymbol{\varrho}) = V[\mathbf{e}_{lt}(\boldsymbol{\phi})|\boldsymbol{\phi}] = E\left[\partial^2 \ln f(\boldsymbol{\varepsilon}_t^*;\boldsymbol{\varrho})/\partial\boldsymbol{\varepsilon}^*\partial\boldsymbol{\varepsilon}^{*\prime}\middle|\,\boldsymbol{\varrho}\right],$$

$$\mathcal{M}_{ls}(\boldsymbol{\varrho}) = E[\mathbf{e}_{lt}(\boldsymbol{\phi})\mathbf{e}_{st}(\boldsymbol{\phi})'|\boldsymbol{\phi}] = E\left[\partial^2 \ln f(\boldsymbol{\varepsilon}_t^*;\boldsymbol{\varrho})/\partial\boldsymbol{\varepsilon}^*\partial\boldsymbol{\varepsilon}^{*\prime}\cdot(\boldsymbol{\varepsilon}_t^{\prime *}\otimes\mathbf{I}_2)\middle|\,\boldsymbol{\varrho}\right],$$

$$\mathcal{M}_{ss}(\boldsymbol{\varrho}) = V[\mathbf{e}_{st}(\boldsymbol{\phi})|\boldsymbol{\phi}] = E\left[(\boldsymbol{\varepsilon}_t^*\otimes\mathbf{I}_2)\cdot\partial^2 \ln f(\boldsymbol{\varepsilon}_t^*;\boldsymbol{\varrho})/\partial\boldsymbol{\varepsilon}^*\partial\boldsymbol{\varepsilon}^{*\prime}\cdot(\boldsymbol{\varepsilon}_t^{*\prime}\otimes\mathbf{I}_2)|\boldsymbol{\varrho}\right] - \mathbf{K}_{22},$$

$$\mathcal{M}_{lr}(\boldsymbol{\varrho}) = E[\mathbf{e}_{lt}(\boldsymbol{\phi})\mathbf{e}'_{rt}(\boldsymbol{\phi})|\boldsymbol{\phi}] = -E\left[\partial^2 \ln f(\boldsymbol{\varepsilon}_t^*;\boldsymbol{\varrho})/\partial\boldsymbol{\varepsilon}^*\partial\boldsymbol{\varrho}'|\boldsymbol{\varrho}\right],$$

$$\mathcal{M}_{sr}(\boldsymbol{\varrho}) = E[\mathbf{e}_{st}(\boldsymbol{\phi})\mathbf{e}'_{rt}(\boldsymbol{\phi})|\boldsymbol{\phi}] = -E\left[(\boldsymbol{\varepsilon}_t^*\otimes\mathbf{I}_2)\partial^2 \ln f(\boldsymbol{\varepsilon}_t^*;\boldsymbol{\varrho})/\partial\boldsymbol{\varepsilon}^*\partial\boldsymbol{\varrho}'|\boldsymbol{\varrho}\right],$$

and

$$\mathcal{M}_{rr}(\boldsymbol{\varrho}) = V[\mathbf{e}_{rt}(\boldsymbol{\phi})|\boldsymbol{\phi}] = -E\left[\partial^2 \ln f(\boldsymbol{\varepsilon}_t^*;\boldsymbol{\varrho})/\partial\boldsymbol{\varrho}\partial\boldsymbol{\varrho}'|\boldsymbol{\phi}\right].$$

Analogously, we can obtain $AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{ML}) = \mathcal{I}_U^{-1}(\boldsymbol{\theta},\boldsymbol{\varrho})$ by exploiting the expressions for the derivatives of the unrestricted model that we obtained when we discussed the IV estimators.

A.7 Semiparametric estimator (SP)

We can make use of Proposition D3 in Supplementary Appendix D of Fiorentini and Sentana (2021), which indicates that the SP efficiency bound for $j = R, U$ will be given by

$$\ddot{\mathcal{S}}_j(\boldsymbol{\phi}) = \mathcal{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta},\boldsymbol{\varrho}) - \mathbf{Z}_d^{Gj}(\boldsymbol{\theta})\left[\mathcal{M}_{dd}(\boldsymbol{\varrho}) - \mathcal{K}(0)\mathcal{K}^{\mathbf{v}+}(\boldsymbol{\varrho})\mathcal{K}(0)\right]\mathbf{Z}_d^{Gj\prime}(\boldsymbol{\theta}), \tag{A12}$$

where $+$ denotes the Moore-Penrose inverse, with

$$\mathcal{M}_{dd}(\boldsymbol{\varrho}) = \begin{pmatrix} \mathcal{M}_{ll}(\boldsymbol{\varrho}) & \mathcal{M}_{ls}(\boldsymbol{\varrho}) \\ \mathcal{M}'_{ls}(\boldsymbol{\varrho}) & \mathcal{M}_{ss}(\boldsymbol{\varrho}) \end{pmatrix}$$

and the matrix of third and fourth order central moments $\mathcal{K}^{\mathbf{v}}(\boldsymbol{\varrho})$ in (A3). Then, under suitable regularity conditions, we will have that

$$AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{SP}) = [\ddot{\mathcal{S}}_R(\boldsymbol{\theta})]^{-1} \tag{A13}$$

and

$$AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{SP}) = [\ddot{\mathcal{S}}_U(\boldsymbol{\theta})]^{-1}. \tag{A14}$$

The expression for $\mathcal{K}^{\mathbf{u}}(\boldsymbol{\varrho})$ simplifies considerably in the spherically symmetric case because

$$E(\mathbf{u}_i^*\mathbf{u}_i^{*\prime} \otimes \mathbf{u}_i^*) = \mathbf{0}, \tag{A15}$$

$$E(\mathbf{u}_i^*\mathbf{u}_i^{*\prime} \otimes \mathbf{u}_i^*\mathbf{u}_i^{*\prime}) = E[vec(\mathbf{u}_i^*\mathbf{u}_i^{*\prime})vec'(\mathbf{u}_i^*\mathbf{u}_i^{*\prime})] = (\kappa+1)[(\mathbf{I}_4+\mathbf{K}_{22})+vec\,(\mathbf{I}_2)\,vec'\,(\mathbf{I}_2)]. \tag{A16}$$

As a result, after some tedious calculations we obtain that for the estimator that imposes the restriction $\sigma_{12} = 0$,

$$AVar(\sqrt{n}\hat{\alpha}_{SP}) = \frac{\sigma_1^2\sigma_{z_1}^2(1+\kappa)}{\sigma_2^2\sigma_{z_1}^2 + \mathrm{M}_{ll}(1+\kappa)\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}$$

and

$$AVar(\sqrt{n}\hat{\beta}_{SP}) = \frac{\sigma_1^2[\sigma_2^2 + (1+\kappa)\mathrm{M}_{ll}(\mu_1^2\sigma_{z_1}^2 + \mu_2^2\sigma_{z_2}^2 + 2\mu_1\mu_2\sigma_{z_1z_2})]}{\mathrm{M}_{ll}[\sigma_2^2\sigma_{z_1}^2 + (1+\kappa)\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)]};$$

while for the unrestricted one,

$$AVar(\sqrt{n}\tilde{\alpha}_{SP}) = \frac{\sigma_1^2\sigma_{z_1}^2}{\mathrm{M}_{ll}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}$$

and

$$AVar(\sqrt{n}\tilde{\beta}_{SP}) = \frac{\sigma_1^2(\mu_1^2\sigma_{z_1}^2 + \mu_2^2\sigma_{z_2}^2 + 2\mu_1\mu_2\sigma_{z_1z_2})}{\mathrm{M}_{ll}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}.$$

Once again, we can rewrite these expressions as

$$AVar(\sqrt{n}\hat{\alpha}_{SP}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)(1+\kappa)}{(1-R_2^2)[(1-\rho_{y_2z_2.z_1}^2) + \rho_{y_2z_2.z_1}^2\mathrm{M}_{ll}(1+\kappa)]}$$

and

$$AVar(\sqrt{n}\hat{\beta}_{SP}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)(1+\kappa)[1-R_2^2+R_2^2\mathrm{M}_{ll}(1+\kappa)]}{(1-R_2^2)[(1-\rho_{y_2z_2.z_1}^2) + \rho_{y_2z_2.z_1}^2\mathrm{M}_{ll}(1+\kappa)]\mathrm{M}_{ll}(1+\kappa)},$$

in the restricted case, and as

$$AVar(\sqrt{n}\tilde{\alpha}_{SP}) = \frac{(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)}{\mathrm{M}_{ll}(1-R_2^2)\rho_{y_2z_2.z_1}^2}$$

and

$$AVar(\sqrt{n}\tilde{\beta}_{SP}) = \frac{R_2^2(1-R_1^2)(1-\rho_{y_2z_2.z_1}^2)}{\mathrm{M}_{ll}(1-R_2^2)\rho_{y_2z_2.z_1}^2}$$

when σ_{12} is also estimated.

A.8 Reparametrisations

The results in the previous subsections can be used to derive the asymptotic distribution of alternative parametrisations. Specifically, for estimators that impose $\sigma_{12} = 0$, the asymptotic covariance of the reparametrisation in (8) is simply

$$AVar(\sqrt{n}\hat{\boldsymbol{\theta}}^{\dagger}) = \mathbf{J}_{\boldsymbol{\theta}^{\dagger}\boldsymbol{\theta}} AVar(\sqrt{n}\hat{\boldsymbol{\theta}})\mathbf{J}'_{\boldsymbol{\theta}^{\dagger}\boldsymbol{\theta}},$$

where

$$\mathbf{J}_{\boldsymbol{\theta}^{\dagger}\boldsymbol{\theta}} = \frac{\partial \boldsymbol{\theta}^{\dagger}}{\partial \boldsymbol{\theta}'} = \begin{bmatrix} \mathbf{I}_6 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\sigma_1^2} & -\frac{1}{2\sigma_2^2} \\ \mathbf{0} & \frac{\sigma_2}{2\sigma_1} & \frac{\sigma_1}{2\sigma_2} \end{bmatrix}. \tag{A17}$$

In turn, for unconstrained estimators that also estimate σ_{12}, so that $\boldsymbol{\vartheta}^{\dagger} = (\boldsymbol{\vartheta}', \psi_{12})'$, we would have

$$AVar(\sqrt{n}\tilde{\boldsymbol{\vartheta}}^{\dagger}) = \mathbf{J}_{\boldsymbol{\vartheta}^{\dagger}\boldsymbol{\vartheta}} AVar(\sqrt{n}\tilde{\boldsymbol{\vartheta}})\mathbf{J}'_{\boldsymbol{\vartheta}^{\dagger}\boldsymbol{\vartheta}}$$

with

$$\mathbf{J}_{\boldsymbol{\vartheta}^{\dagger}\boldsymbol{\vartheta}} = \frac{\partial \boldsymbol{\vartheta}^{\dagger}}{\partial \boldsymbol{\vartheta}'} = \begin{bmatrix} \mathbf{I}_6 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_1^2\sigma_2^2 - 2\sigma_{12}^2}{2\sigma_1^2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} & -\frac{\sigma_1^2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} & \frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ \mathbf{0} & \frac{\sigma_2^2}{2\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} & \frac{\sigma_1^2}{2\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} & -\frac{\sigma_{12}}{2\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \\ \mathbf{0} & -\frac{\sigma_{12}}{\sigma_1^4} & 0 & \frac{1}{\sigma_1^2} \end{bmatrix}. \tag{A18}$$

B Proofs of Propositions

Proof of Proposition 1

Computing in Mathematica the spectral decomposition of $AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{LS}) - AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{MD})$ using the expressions (A5) and (A6), we find that it has only one eigenvalue different from zero, namely,

$$\frac{(\mu_{22}-1)^2\mu_2^2\sigma_1^2\sigma_2^2\{\sigma_{z_1z_2}^2(1+\tau_1^2)\mu_2^2 + [1+\mu_1^2+(\mu_0+\tau_2\mu_2)^2]\sigma_{z_1}^4\}(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}{[\mu_{22}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2) - \sigma_2^2\sigma_{z_1}^2][\sigma_2^2\sigma_{z_1}^2 - \mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)]^2}$$
$$-\frac{2(\mu_{22}-1)^2\mu_2^2\sigma_1^2\sigma_2^2\sigma_{z_1z_2}\mu_2[\tau_1(\mu_0+\tau_2\mu_2)-\mu_1]\sigma_{z_1}(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)}{[\mu_{22}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2) - \sigma_2^2\sigma_{z_1}][\sigma_2^2\sigma_{z_1} - \mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2 - \sigma_{z_1z_2}^2)]^2},$$

which is non-negative, with

$$\left(\frac{\sigma_{z_1}(1+\mu_2\tau_2) - \sigma_{z_1z_2}\mu_2\tau_1}{\sigma_{z_1}\mu_1 + \sigma_{z_1z_2}\mu_2}, -\frac{\sigma_{z_1}}{\sigma_{z_1}^2\mu_1 + \sigma_{z_1z_2}\mu_2}, 1, \mathbf{0}_{1\times5}\right)' \tag{B19}$$

as associated eigenvector.

Analogously, after computing the spectral decomposition of $AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{IV}) - AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{MD})$ using the expressions (A2) and (A6), we find that it has only one eigenvalue different from zero, namely,

$$\frac{\sigma_1^2\sigma_2^2\{\sigma_{z_1z_2}^2(1+\tau_1^2)\mu_2^2-2\sigma_{z_1z_2}\mu_2[\tau_1(\mu_0+\tau_2\mu_2)-\mu_1]\sigma_{z_1}^2+[1+\mu_1^2+(\mu_0+\tau_2\mu_2)^2]\sigma_{z_1}^4\}}{[\mu_{22}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)-\sigma_2^2\sigma_{z_1}^2]\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)},$$

which is non-negative, with (B19) its associated eigenvector once again.

Finally, doing the same for $AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{IV})-AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{MD})$ by combining (A2) and (A5), we find that it has only one eigenvalue different from zero, namely,

$$\frac{(\mu_{22}-1)^2\mu_2^2\sigma_{z_1}^2\sigma_{z_2}^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)\{\sigma_{z_1z_2}^2(1+\tau_1^2)\mu_2^2-2\sigma_{z_1z_2}\mu_2[(\mu_0+\mu_2\tau_2)\tau_1-\mu_1]\sigma_{z_1}^2\}}{[\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)-\sigma_1^2\sigma_{z_1}^2]^2[\mu_{22}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)-\sigma_2^2\sigma_{z_1}^2]}$$
$$+\frac{(\mu_{22}-1)^2\mu_2^2\sigma_{z_1}^2\sigma_{z_2}^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)\{[1+\mu_1^2+\sigma_{z_1}^2(\mu_0+\mu_2\tau_2)^2]\}}{[\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)-\sigma_1^2\sigma_{z_1}^2]^2[\mu_{22}\mu_2^2(\sigma_{z_1}^2\sigma_{z_2}^2-\sigma_{z_1z_2}^2)-\sigma_2^2\sigma_{z_1}^2]},$$

which can be positive or negative depending on μ_{22}, and with the same eigenvector. $\square$

Proof of Proposition 2

Computing in Mathematica the spectral decomposition of $AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{SS}^{\dagger})-AVar[\sqrt{n}\hat{\boldsymbol{\theta}}_{ML}^{\dagger}(\bar{\boldsymbol{\eta}})]$ using (A17), the expression in (A9) and exploiting the fact that

$$AVar[\sqrt{n}\hat{\boldsymbol{\theta}}_{ML}(\bar{\boldsymbol{\eta}})]=[\mathcal{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta},\bar{\boldsymbol{\eta}})]^{-1},$$

where $\mathcal{I}_{\boldsymbol{\theta\theta}}(\boldsymbol{\theta},\boldsymbol{\eta})$ denotes the block of the information matrix of the mean and variance parameters, we find that it has only one eigenvalue different from zero, with associated eigenvector

$$(\mathbf{0}_{1\times 7},1)'.$$

Similarly, we find that the spectral decomposition of $AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{SP})-AVar[\sqrt{n}\hat{\boldsymbol{\theta}}_{ML}(\bar{\boldsymbol{\eta}})]$ using also (A13), has five eigenvalues different from zero. By looking at the orthogonal basis for its null space, which is given by

$$\left(0,\sigma_{z_1}^2\mu_1+\sigma_{z_1z_2}\mu_2,\sigma_{z_1}^2,\mathbf{0}_{1\times 5}\right)'$$

and

$$(\mathbf{0}_{2\times 4},\mathbf{I}_2,\mathbf{0}_{2\times 2})',$$

we can immediately see that the parameters that are estimated adaptively are μ_1, μ_2, and the linear combination of α and β indicated by the first eigenvector. In turn, a basis for its image is given by

$$(1,\mathbf{0}_{1\times 7})',$$

$$\left(0,-\sigma_{z_1}^2,\mu_1\sigma_{z_1}^2+\sigma_{z_1z_2}\mu_2,\mathbf{0}_{1\times 5}\right)',$$

$$(\mathbf{0}_{1\times3}, 1, \mathbf{0}_{1\times4})'$$

and

$$(\mathbf{0}_{1\times6}, \mathbf{I}_2)'.$$

Finally, using an entirely analogous procedure with (A13) and (A6), we find that the spectral decomposition of $AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{MD}) - AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{SP})$ has four eigenvalues different from zero, with a basis for its image given by $(\mathbf{0}_{2\times1}, \mathbf{I}_2, \mathbf{0}_{2\times5})$ and $(\mathbf{0}_{2\times4}, \mathbf{I}_2, \mathbf{0}_{2\times2})$, and a basis for its kernel by $(\mathbf{0}_{2\times6}, \mathbf{I}_2)$, $(1, \mu_0 + \mu_1\tau_1 + \mu_2\tau_2, \tau_1, \mathbf{0}_{1\times5})$ and $(\mathbf{0}_{1\times3}, 1, \tau_1, \tau_2, \mathbf{0}_{1\times2})$, as can be easily checked by premultiplying the difference between the covariance matrices by an 8×8 matrix whose rows concatenate those two basis and postmultiplying it by its transpose. □

Proof of Proposition 3

Computing in Mathematica the spectral decomposition of $AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{ML}) - AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{ML})$ using the expressions (A7) and (A8), we find that it has only one eigenvalue different from zero, namely,

$$\frac{\sigma_1^2\sigma_2^2\mathrm{M}_{ss}\{\sigma^2_{z_1z_2}\mu_2^2(1+\tau_1^2) - 2\sigma_{z_1z_2}\sigma^2_{z_1}\mu_2[(\mu_0+\mu_2\tau_2)\tau_1 - \mu_1] + \sigma^4_{Z_1}[1+\mu_1^2+(\mu_0+\mu_2\tau_2)^2]\}}{\mu_2^2\mathrm{M}_{ll}(\sigma^2_{z_1}\sigma^2_{z_2} - \sigma^2_{z_1z_2})[\mathrm{M}_{ss}\sigma_2^2\sigma^2_{z_1} + \mu_2^2\mathrm{M}_{ll}(\sigma^2_{z_1}\sigma^2_{z_2} - \sigma^2_{z_1z_2})]}$$

with associated eigenvector (B19).

Using (A9) and (A10), we find that the same turns out to be true for $AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{SS}) - AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{SS})$.

Finally, if we do the same for $AVar(\sqrt{n}\tilde{\boldsymbol{\theta}}_{NP}) - AVar(\sqrt{n}\hat{\boldsymbol{\theta}}_{NP})$ using (A13) and (A14), we also find that it has only one eigenvalue different from zero, namely

$$\frac{\sigma_1^2\sigma_2^2\{\sigma^2_{z_1z_2}\mu_2^2(1+\tau_1^2) - 2\sigma_{z_1z_2}\sigma^2_{z_1}\mu_2[(\mu_0+\mu_2\tau_2)\tau_1 - \mu_1] + \sigma^4_{Z_1}[1+\mu_1^2+(\mu_0+\mu_2\tau_2)^2]\}}{\mu_2^2\mathrm{M}_{ll}(\sigma^2_{z_1}\sigma^2_{z_2} - \sigma^2_{z_1z_2})[\sigma_2^2\sigma^2_{z_1} + (1+\kappa)\mu_2^2\mathrm{M}_{ll}(\sigma^2_{z_1}\sigma^2_{z_2} - \sigma^2_{z_1z_2})]},$$

and that its image is given by the same eigenvector as in the previous cases. □

C Simplifying the DGP

C.1 Standardised variables

We start by assuming that:

$$\begin{pmatrix} y_1 \\ y_2 \\ z_1 \\ z_2 \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{y_1y_2} & \rho_{y_1z_1} & \rho_{y_1z_2} \\ \rho_{y_1y_2} & 1 & \rho_{y_2z_1} & \rho_{y_2z_2} \\ \rho_{y_1z_1} & \rho_{y_2z_1} & 1 & \rho_{z_1z_2} \\ \rho_{y_1z_2} & \rho_{y_2z_2} & \rho_{z_1z_2} & 1 \end{pmatrix}\right],$$

where the correlation matrix is positive definite.

In this notation, the coefficients of the least squares projection of y_1 onto y_2 and z_1 are

$$\begin{aligned}\begin{pmatrix}\alpha\\ \beta\end{pmatrix} &= \begin{pmatrix}1 & \rho_{y_2z_1}\\ \rho_{y_2z_1} & 1\end{pmatrix}^{-1}\begin{pmatrix}\rho_{y_1y_2}\\ \rho_{y_1z_1}\end{pmatrix}\\ &= \frac{1}{1-\rho^2_{y_2z_1}}\begin{pmatrix}\rho_{y_1y_2}-\rho_{y_1z_1}\rho_{y_2z_1}\\ \rho_{y_1z_1}-\rho_{y_1y_2}\rho_{y_2z_1}\end{pmatrix},\end{aligned}$$

the corresponding projection errors

$$u_1 = y_1 - \alpha y_2 - \beta z_1 = y_1 - \frac{\rho_{y_1y_2}-\rho_{y_1z_1}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}}y_2 - \frac{\rho_{y_1z_1}-\rho_{y_1y_2}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}}z_1$$

and the residual variance

$$\begin{aligned}V(u_1) &= 1-\begin{pmatrix}\rho_{y_1y_2} & \rho_{y_1z_1}\end{pmatrix}\begin{pmatrix}1 & \rho_{y_2z_1}\\ \rho_{y_2z_1} & 1\end{pmatrix}^{-1}\begin{pmatrix}\rho_{y_1y_2}\\ \rho_{y_1z_1}\end{pmatrix}\\ &= 1-\frac{\rho^2_{y_1y_2}+\rho^2_{y_1z_1}-2\rho_{y_2z_1}\rho_{y_1y_2}\rho_{y_1z_1}}{1-\rho^2_{y_2z_1}},\end{aligned}$$

so that the R^2 becomes

$$R_1^2 = \frac{\rho^2_{y_1y_2}+\rho^2_{y_1z_1}-2\rho_{y_2z_1}\rho_{y_1y_2}\rho_{y_1z_1}}{1-\rho^2_{y_2z_1}}.$$

In turn, the coefficients of the least squares projection of y_2 onto z_1 and z_2 are

$$\begin{aligned}\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix} &= \begin{pmatrix}1 & \rho_{z_1z_2}\\ \rho_{z_1z_2} & 1\end{pmatrix}^{-1}\begin{pmatrix}\rho_{y_2z_1}\\ \rho_{y_2z_2}\end{pmatrix}\\ &= \frac{1}{1-\rho^2_{z_1z_2}}\begin{pmatrix}\rho_{y_2z_1}-\rho_{y_2z_2}\rho_{z_1z_2}\\ \rho_{y_2z_2}-\rho_{y_2z_1}\rho_{z_1z_2}\end{pmatrix},\end{aligned}$$

the corresponding projection errors

$$u_2 = y_2 - \mu_1 z_1 - \mu_2 z_2 = y_2 - \frac{\rho_{y_2z_1}-\rho_{y_2z_2}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}}z_1 - \frac{\rho_{y_2z_2}-\rho_{y_2z_1}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}}z_2$$

and the residual variance

$$\begin{aligned}V(u_2) &= 1-\begin{pmatrix}\rho_{y_2z_1} & \rho_{y_2z_2}\end{pmatrix}\begin{pmatrix}1 & \rho_{z_1z_2}\\ \rho_{z_1z_2} & 1\end{pmatrix}^{-1}\begin{pmatrix}\rho_{y_2z_1}\\ \rho_{y_2z_2}\end{pmatrix}\\ &= 1-\frac{\rho^2_{y_2z_1}+\rho^2_{y_2z_2}-2\rho_{z_1z_2}\rho_{y_2z_1}\rho_{y_2z_2}}{1-\rho^2_{z_1z_2}},\end{aligned}$$

so that the R^2 becomes

$$R_2^2 = \frac{\rho^2_{y_2z_1}+\rho^2_{y_2z_2}-2\rho_{z_1z_2}\rho_{y_2z_1}\rho_{y_2z_2}}{1-\rho^2_{z_1z_2}}.$$

Finally, the covariance between the previous projection errors is

$$\begin{aligned}
& E[(y_1 - \alpha y_2 - \beta z_1)(y_2 - \mu_1 z_1 - \mu_2 z_2)] \\
= \; & E\left[\left(y_1 - \frac{\rho_{y_1y_2} - \rho_{y_1z_1}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} y_2 - \frac{\rho_{y_1z_1} - \rho_{y_1y_2}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} z_1\right)\right. \\
& \left.\left(y_2 - \frac{\rho_{y_2z_1} - \rho_{y_2z_2}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} z_1 - \frac{\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} z_2\right)\right] \\
= \; & \rho_{y_1y_2} - \frac{\rho_{y_2z_1} - \rho_{y_2z_2}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} \rho_{y_1z_1} - \frac{\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} \rho_{y_1z_2} \\
& - \frac{\rho_{y_1y_2} - \rho_{y_1z_1}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} + \frac{\rho_{y_1y_2} - \rho_{y_1z_1}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} \frac{\rho_{y_2z_1} - \rho_{y_2z_2}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} \rho_{y_2z_1} \\
& + \frac{\rho_{y_1y_2} - \rho_{y_1z_1}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} \frac{\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} \rho_{y_2z_2} - \frac{\rho_{y_1z_1} - \rho_{y_1y_2}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} \rho_{y_2z_1} \\
& + \frac{\rho_{y_1z_1} - \rho_{y_1y_2}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} \frac{\rho_{y_2z_1} - \rho_{y_2z_2}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} + \frac{\rho_{y_1z_1} - \rho_{y_1y_2}\rho_{y_2z_1}}{1-\rho^2_{y_2z_1}} \frac{\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} \rho_{z_1z_2} \\
= \; & \frac{(\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2})}{1-\rho^2_{z_1z_2}} \frac{[\rho_{y_1y_2}(\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2}) + \rho_{y_1z_1}(\rho_{z_1z_2} - \rho_{y_2z_2}\rho_{y_2z_1}) - \rho_{y_1z_2}(1-\rho^2_{y_2z_1})]}{1-\rho^2_{y_2z_1}}
\end{aligned}$$

Therefore, for y_2 to be exogenous in the first equation, we need either

$$\mu_2 = \frac{\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2}}{1-\rho^2_{z_1z_2}} = 0,$$

which seems very restrictive, or

$$\rho_{y_1z_2} = \frac{\rho_{y_1y_2}(\rho_{y_2z_2} - \rho_{y_2z_1}\rho_{z_1z_2}) + \rho_{y_1z_1}(\rho_{z_1z_2} - \rho_{y_2z_1}\rho_{y_2z_2})}{1-\rho^2_{y_2z_1}} = \frac{1-\rho^2_{z_1z_2}}{1-\rho^2_{y_2z_1}} \mu_2 \rho_{y_1y_2} + \delta\rho_{y_1z_1}, \quad \text{(C20)}$$

where δ is the coefficient of z_1 in the least squares projection of z_2 onto y_2 and z_1, whose coefficients are given by

$$\begin{pmatrix} \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} 1 & \rho_{y_2z_1} \\ \rho_{y_2z_1} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \rho_{y_2z_2} \\ \rho_{z_1z_2} \end{pmatrix} = \frac{1}{1-\rho^2_{y_2z_1}} \begin{pmatrix} \rho_{y_2z_2} - \rho_{z_1z_2}\rho_{y_2z_1} \\ \rho_{z_1z_2} - \rho_{y_2z_2}\rho_{y_2z_1} \end{pmatrix}.$$

Therefore, if we assume $\mu_2 \neq 0$, then we need to choose $\rho_{y_1z_2}$ so that (C20) holds.

C.2 Original variables

Let us now consider the least squares projection of y_2^o onto a constant, z_1^o and z_2^o, which is given by

$$y_2^o = \mu_0^o + \mu_1^o z_1^o + \mu_2^o z_2^o + u_1^o.$$

We can then individually centre and standardise each of the variables involved as follows

$$y_2 = \frac{y_2^o - \mu_0^o - \mu_1^o E(z_1^o) - \mu_2^o E(z_2^o)}{\sqrt{\mu_1^{o2} V(z_1^o) + \mu_2^{o2} V(z_2^o) + 2\mu_1^o\mu_2^o Cov(z_1^o, z_2^o) + V(u_1^o)}},$$

$$z_1 = \frac{z_1^o - E(z_1^o)}{\sqrt{V(z_1^o)}}, \quad \text{and} \quad z_2 = \frac{z_2^o - E(z_2^o)}{\sqrt{V(z_2^o)}},$$

which leads to the following transformed equation

$$y_2 = \mu_1 z_1 + \mu_2 z_2 + u_1,$$

where

$$\begin{aligned}
\mu_1 &= \mu_1^o \sqrt{\frac{V(z_1^o)}{\mu_1^{o2} V(z_1^o) + \mu_2^{o2} V(z_2^o) + 2\mu_1^o \mu_2^o Cov(z_1^o, z_2^o) + V(u_1^o)}}, \\
\mu_2 &= \mu_2^o \sqrt{\frac{V(z_2^o)}{\mu_1^{o2} V(z_1^o) + \mu_2^{o2} V(z_2^o) + 2\mu_1^o \mu_2^o Cov(z_1^o, z_2^o) + V(u_1^o)}},
\end{aligned}$$

and

$$V(u_1) = \frac{V(u_1^o)}{\mu_1^{o2} V(z_1^o) + \mu_2^{o2} V(z_2^o) + 2\mu_1^o \mu_2^o Cov(z_1^o, z_2^o) + V(u_1^o)} = 1 - R_2^2$$

The coefficients μ_1 and μ_2 are sometimes called the standardised regression coefficients, as they explain the ceteris paribus change in y_2^o (measured in standard deviation units) resulting from a unit standard deviation change in z_1^o or z_2^o.

Thus, once we standardise the three variables involved, the crucial ingredients of the first equation are the coefficient of determination R_2^2, the correlation between the regressors $\rho_{z_1 z_2}$ and the partial correlations between y_2 and each of the regressors, which are given by

$$\begin{aligned}
\rho_{y_2 z_1 \cdot z_2} &= \frac{E[(y_2 - \rho_{y_2 z_2} z_2)(z_1 - \rho_{z_1 z_2} z_2)]}{\sqrt{V(y_2 - \rho_{y_2 z_2} z_2) V(z_1 - \rho_{z_1 z_2} z_2)}} = \frac{\rho_{y_2 z_1} - \rho_{z_1 z_2} \rho_{y_2 z_2}}{\sqrt{\left(1 - \rho_{y_2 z_2}^2\right)\left(1 - \rho_{z_1 z_2}^2\right)}} = \mu_1 \sqrt{\frac{1 - \rho_{z_1 z_2}^2}{1 - \rho_{y_2 z_2}^2}}, \\
\rho_{y_2 z_2 \cdot z_1} &= \frac{E[(y_2 - \rho_{y_2 z_1} z_1)(z_2 - \rho_{z_1 z_2} z_1)]}{\sqrt{V(y_2 - \rho_{y_2 z_1} z_1) V(z_2 - \rho_{z_1 z_2} z_1)}} = \frac{\rho_{y_2 z_2} - \rho_{z_1 z_2} \rho_{y_2 z_1}}{\sqrt{\left(1 - d^2\right)\left(1 - \rho_{z_1 z_2}^2\right)}} = \mu_2 \sqrt{\frac{1 - \rho_{z_1 z_2}^2}{1 - \rho_{y_2 z_1}^2}}.
\end{aligned}$$

In fact, there are only three underlying parameters that determine these four quantities: $\rho_{y_2 z_1}$, $\rho_{y_2 z_2}$ and $\rho_{z_1 z_2}$ because

$$\begin{aligned}
\rho_{y_2 z_1 \cdot z_2}^2 &= \frac{R_2^2 - \rho_{y_2 z_2}^2}{1 - \rho_{y_2 z_2}^2}, \\
\rho_{y_2 z_2 \cdot z_1}^2 &= \frac{R_2^2 - \rho_{y_2 z_1}^2}{1 - \rho_{y_2 z_1}^2},
\end{aligned}$$

or alternatively

$$\begin{aligned}
\rho_{y_2 z_2}^2 &= \frac{R_2^2 - \rho_{y_2 z_1 \cdot z_2}^2}{1 - \rho_{y_2 z_1 \cdot z_2}^2}, \\
\rho_{y_2 z_1}^2 &= \frac{R_2^2 - \rho_{y_2 z_2 \cdot z_1}^2}{1 - \rho_{y_2 z_2 \cdot z_1}^2}.
\end{aligned}$$

Thus, we can either select $\rho_{y_2 z_1}$, $\rho_{y_2 z_2}$ and $\rho_{z_1 z_2}$, or we can select R_2^2, $\rho_{y_2 z_1 \cdot z_2}^2$ and $\rho_{y_2 z_2 \cdot z_1}^2$.

D On multivariate discrete mixture of normals

Consider the following mixture of two multivariate normals

$$\mathbf{u}_t \sim \begin{cases} N(\boldsymbol{\nu}_1, \boldsymbol{\Gamma}_1) & \text{with probability} \quad \lambda, \\ N(\boldsymbol{\nu}_2, \boldsymbol{\Gamma}_2) & \text{with probability} \quad 1-\lambda. \end{cases} \tag{D21}$$

Let s_t denote a Bernoulli variable which takes the value 1 with probability λ and 0 with probability $1-\lambda$. As is well known, the unconditional mean vector and covariance matrix of the observed variables are:

$$E(\mathbf{u}_t) = \boldsymbol{\tau} = E[E(\mathbf{u}_t|s_t)] = \lambda\boldsymbol{\nu}_1 + (1-\lambda)\boldsymbol{\nu}_2,$$

$$V(\mathbf{u}_t) = \boldsymbol{\Psi} = V[E(\mathbf{u}_t|s_t)] + E[V(\mathbf{u}_t|s_t)] = \lambda(1-\lambda)\boldsymbol{\delta}\boldsymbol{\delta}' + \lambda\boldsymbol{\Sigma}_1 + (1-\lambda)\boldsymbol{\Sigma}_2,$$

where $\boldsymbol{\delta} = \boldsymbol{\nu}_1 - \boldsymbol{\nu}_2$.

Therefore, this random vector, which we will denote as $\mathbf{u}_t^*$, will be standardised if and only if

$$\lambda\boldsymbol{\mu}_1 + (1-\lambda)\boldsymbol{\mu}_2 = \mathbf{0}$$

and

$$\lambda(1-\lambda)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' + \lambda\boldsymbol{\Sigma}_1 + (1-\lambda)\boldsymbol{\Sigma}_2 = \mathbf{I}.$$

For example, in the bivariate case, if we let $\boldsymbol{\Psi}_L\boldsymbol{\Psi}_L'$ denote the Cholesky decomposition of $\boldsymbol{\Psi}$, we can write

$$\mathbf{u}_t = \boldsymbol{\pi} + \boldsymbol{\Psi}_L\mathbf{u}_t^*, \quad \text{where } \boldsymbol{\pi} = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Psi}_L = \begin{bmatrix} \psi_{11} & 0 \\ \psi_{21} & \psi_{22} \end{bmatrix}.$$

Additionally, let

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}, \text{ and } \aleph_L = \begin{bmatrix} \varkappa_{11} & 0 \\ \varkappa_{21} & \varkappa_{22} \end{bmatrix},$$

so that the vector of shape parameters of $\mathbf{u}_t^*$ becomes $\boldsymbol{\varrho} = (\delta_1, \delta_2, \varkappa_{11}, \varkappa_{21}, \varkappa_{22}, \lambda)'$.

Let us initially assume that $\boldsymbol{\nu}_1 = \boldsymbol{\nu}_2 = \mathbf{0}$, so that $\boldsymbol{\delta} = \mathbf{0}$. Let $\boldsymbol{\Gamma}_{1L}\boldsymbol{\Gamma}_{1L}'$ and $\boldsymbol{\Gamma}_{2L}\boldsymbol{\Gamma}_{2L}'$ denote the Cholesky decompositions of the covariance matrices of the two components. Then, we can write

$$\lambda\boldsymbol{\Gamma}_1 + (1-\lambda)\boldsymbol{\Gamma}_2 = \boldsymbol{\Gamma}_{1L}[\lambda\mathbf{I}_2 + (1-\lambda)\boldsymbol{\Gamma}_{1L}^{-1}\boldsymbol{\Gamma}_{2L}\boldsymbol{\Gamma}_{2L}'\boldsymbol{\Gamma}_{1L}^{-1\prime}]\boldsymbol{\Gamma}_{1L}' = \boldsymbol{\Gamma}_{1L}(\lambda\mathbf{I}_2 + (1-\lambda)\aleph_L\aleph_L')\boldsymbol{\Gamma}_{1L}'.$$

Thus, it is not difficult to see that choosing

$$\boldsymbol{\Gamma}_1 = [\lambda\mathbf{I}_2 + (1-\lambda)\aleph_L\aleph_L']^{-1} \quad \text{and} \quad \boldsymbol{\Gamma}_2 = \boldsymbol{\Gamma}_{1L}\aleph_L\aleph_L'\boldsymbol{\Gamma}_{1L}'$$

or, equivalently,

$$\mathbf{\Gamma}_{1L} = [\lambda \mathbf{I}_2 + (1-\lambda)\aleph_L \aleph_L']^{-\frac{1}{2}} \text{ and } \mathbf{\Gamma}_{2L} = \mathbf{\Gamma}_{1L}\aleph_L$$

we can indeed obtain a bivariate standardised vector $\mathbf{u}_t^*$.

Now consider the case $\boldsymbol{\delta} \neq \mathbf{0}$, and let $\mathbf{\Upsilon} = \lambda(1-\lambda)\boldsymbol{\delta}\boldsymbol{\delta}' + \mathbf{I}_2$. Then, it is easy to see that

$$\boldsymbol{\nu}_1^* = \mathbf{\Upsilon}^{-\frac{1}{2}}\boldsymbol{\nu}_1, \ \ \boldsymbol{\nu}_2^* = \mathbf{\Upsilon}^{-\frac{1}{2}}\boldsymbol{\nu}_2, \ \ \mathbf{\Gamma}_1^* = \mathbf{\Upsilon}_1^{-\frac{1}{2}}\mathbf{\Gamma}_1\mathbf{\Upsilon}'^{-\frac{1}{2}}, \text{ and } \mathbf{\Gamma}_2^* = \mathbf{\Upsilon}^{-\frac{1}{2}}\mathbf{\Gamma}_2\mathbf{\Upsilon}'^{-\frac{1}{2}}$$

continue to generate another standardised vector.

In summary, we can generate a standardised bivariate mixture as

$$\mathbf{u}_t^* = \mathbf{\Upsilon}^{-\frac{1}{2}}\left\{(s_i-\lambda)\boldsymbol{\delta} + s_i + [\mathbf{\Gamma}_{1L} - s_i(\mathbf{\Gamma}_{1L} - \mathbf{\Gamma}_{2L})]\mathbf{z}_t\right\},$$

where $\mathbf{z}_t \sim N(\mathbf{0}, \mathbf{I}_2)$. The intuition is as follows. First, note that $(s_t-\lambda)\boldsymbol{\delta}$ is a shifted and scaled Bernoulli random variable with 0 mean and variance $\lambda(1-\lambda)\boldsymbol{\delta}\boldsymbol{\delta}'$. But since

$$[\mathbf{\Gamma}_{1L} - s_i(\mathbf{\Gamma}_{1L} - \mathbf{\Gamma}_{2L})]\mathbf{z}_t$$

is a discrete scale mixture of normals with 0 unconditional mean and unit unconditional variance that is orthogonal to $(s_t-\lambda)\boldsymbol{\delta}$, the sum of the two random variables will have variance $\mathbf{I}_2 + \lambda(1-\lambda)\boldsymbol{\delta}\boldsymbol{\delta}'$, which explains the $\mathbf{\Upsilon}^{-\frac{1}{2}}$ in front of the curly brackets.

Therefore, two equivalent ways of defining and simulating $\mathbf{u}_t$ with mean $\boldsymbol{\tau}$ and variance $\mathbf{\Psi}$ are

$$\mathbf{u}_t = \boldsymbol{\tau} + \mathbf{\Psi}_L\mathbf{u}_t^*, \text{ where } \mathbf{u}_t^* = \begin{cases} N[\boldsymbol{\nu}_1^*(\boldsymbol{\eta}), \mathbf{\Gamma}_1^*(\boldsymbol{\eta})] \text{ with probability } \lambda \\ N[\boldsymbol{\nu}_2^*(\boldsymbol{\eta}), \mathbf{\Gamma}_2^*(\boldsymbol{\eta})] \text{ with probability } 1-\lambda \end{cases} \tag{D22}$$

and

$$\mathbf{u}_t = \begin{cases} N(\boldsymbol{\nu}_1, \mathbf{\Gamma}_{1L}\mathbf{\Gamma}_{1L}') \text{ with probability } \lambda \\ N(\boldsymbol{\nu}_2, \mathbf{\Gamma}_{2L}\mathbf{\Gamma}_{2L}') \text{ with probability } 1-\lambda \end{cases}$$

where

$$\boldsymbol{\nu}_i = \boldsymbol{\nu}_i(vec'(\boldsymbol{\pi}), vech'(\mathbf{\Psi}_L), vec'(\boldsymbol{\delta}), vech'(\aleph_L), \lambda)$$

and

$$\mathbf{\Gamma}_{iL} = \mathbf{\Gamma}_{iL}(vec'(\boldsymbol{\pi}), vech'(\mathbf{\Psi}_L), vec'(\boldsymbol{\delta}), vech'(\mathbf{\Xi}_L), \lambda)$$

for $i = 1, 2$. In this set up, the means of the components will be given by $\boldsymbol{\nu}_1 = (\nu_1^1, \nu_2^1)'$ with

$$\nu_1^1 = \tau_1 + \frac{(1-\lambda)\psi_{11}\delta_1}{\sqrt{1+\lambda(1-\lambda)\delta_1^2}}$$

and

$$\nu_2^1 = \tau_2 + \frac{(1-\lambda)\psi_{21}\delta_1}{\sqrt{1+\lambda(1-\lambda)\delta_1^2}} + \frac{(1-\lambda)\psi_{22}\delta_2}{1+\lambda(1-\lambda)\delta_1^2}\sqrt{\frac{1+\lambda(1-\lambda)\delta_1^2}{1+\lambda(1-\lambda)(\delta_1^2+\delta_2^2)}},$$

and $\boldsymbol{\nu}_2 = (\nu_1^2, \nu_2^2)'$ with

$$\nu_1^2 = \tau_1 - \frac{\lambda\psi_{11}\delta_1}{\sqrt{1+\lambda(1-\lambda)\delta_1^2}}$$

and

$$\nu_2^2 = \tau_2 - \frac{\lambda\psi_{11}\delta_1}{\sqrt{1+\lambda(1-\lambda)\delta_1^2}} - \frac{\lambda\psi_{22}\delta_2}{1+\lambda(1-\lambda)\delta_1^2}\sqrt{\frac{1+\lambda(1-\lambda)\delta_1^2}{1+\lambda(1-\lambda)(\delta_1^2+\delta_2^2)}}.$$

As for the the Cholesky decompositions of the covariance matrices of the two components, namely

$$\boldsymbol{\Gamma}_{1L} = \begin{bmatrix} \gamma_{11}^1 & 0 \\ \gamma_{21}^1 & \gamma_{22}^1 \end{bmatrix} \text{ and } \boldsymbol{\Gamma}_{2L} = \begin{bmatrix} \gamma_{11}^2 & 0 \\ \gamma_{21}^2 & \gamma_{22}^2 \end{bmatrix},$$

we will have

$$\gamma_{11}^1 = \frac{1}{\sqrt{[1+\lambda(1-\lambda)\delta_1^2][\lambda+(1-\lambda)\varkappa_{11}^2]}}\psi_{11},$$

$$\gamma_{22}^1 = \sqrt{\frac{[1+\lambda(1-\lambda)\delta_1^2][\lambda+(1-\lambda)\varkappa_{11}^2]}{[1+\lambda(1-\lambda)(\delta_1^2+\delta_2^2)]\{\lambda[(\varkappa_{11}^2+\varkappa_{21}^2)(1-\lambda)-\lambda]+(1-\lambda)\lambda\varkappa_{22}^2+(1-\lambda)^2\varkappa_{11}^2\varkappa_{22}^2\}}}\psi_{22},$$

$$\begin{aligned}\gamma_{21}^1 &= \gamma_{11}^1\frac{\psi_{21}}{\psi_{11}} - \gamma_{22}^1\frac{(1-\lambda)\varkappa_{11}\varkappa_{21}}{\lambda+(1-\lambda)\varkappa_{11}^2} \\ &\quad -\gamma_{22}^1(1-\lambda)\lambda\delta_1\delta_2\frac{\sqrt{\lambda[(\varkappa_{11}^2+\varkappa_{21}^2)(1-\lambda)-\lambda]+(1-\lambda)\lambda\varkappa_{22}^2+(1-\lambda)^2\varkappa_{11}^2\varkappa_{22}^2}}{[1+\lambda(1-\lambda)\delta_1^2][\lambda+(1-\lambda)\varkappa_{11}^2]},\end{aligned}$$

$$\begin{aligned}\gamma_{11}^2 &= \varkappa_{11}\gamma_{11}^1, \\ \gamma_{22}^2 &= \varkappa_{22}\gamma_{22}^1,\end{aligned}$$

and

$$\begin{aligned}\gamma_{21}^2 &= \gamma_{11}^2\frac{\psi_{21}}{\psi_{11}} - \gamma_{22}^1\frac{\lambda\varkappa_{21}}{[\lambda+(1-\lambda)\varkappa_{11}^2]\varkappa_{22}} \\ &\quad -\gamma_{22}^1(1-\lambda)\lambda\delta_1\delta_2\varkappa_{11}\frac{\sqrt{\varkappa_{11}^2\varkappa_{22}^2+(1-\lambda)[\varkappa_{22}^2+\varkappa_{21}^2+\varkappa_{11}^2(1-\varkappa_{22}^2)]-\lambda\varkappa_{11}^2(\varkappa_{22}^2-\lambda)+\lambda^2}}{[1+\lambda(1-\lambda)\delta_1^2][\lambda+(1-\lambda)\varkappa_{11}^2]\varkappa_{22}}.\end{aligned}$$

Similar calculations can be applied for general n, the only difference being that the number of free parameters of the standardised mixture increases with the square of the cross-sectional dimension.

Tables

Table 1: Asymptotic variances of alternative estimators

Parameter	OLS	PML–SMN$_K$ $K=2$	PML–SMN$_K$ $K=3$	PML–SMN$_K$ $K=4$	SS	ML	MD
			Mean parameters of equation 1a				
γ	1.268	0.931	0.905	0.902	0.901	0.901	1.201
α	1.500	0.782	0.731	0.725	0.723	0.723	1.125
β	1.000	0.656	0.631	0.627	0.627	0.627	0.875
			Mean parameters of equation 1b				
μ_0	1.000	0.792	0.775	0.772	0.771	0.771	1.000
μ_1	0.333	0.264	0.258	0.257	0.257	0.257	0.333
μ_2	0.333	0.264	0.258	0.257	0.257	0.257	0.333
			(Reparametrised) variance parameters of structural innovations				
ω	3.000	1.493	1.313	1.290	1.286	1.286	3.000
σ^2	0.833	0.833	0.833	0.833	0.833	0.300	0.833

Notes: DGP for structural innovations: bivariate Student t with 0 means, unit standard deviations, no correlation and 5 degrees of freedom. Parameter values: $\gamma = 0.204$, $\alpha = \beta = 0.398$, $\mu_0 = 0.155$, $\mu_1 = \mu_2 = 0.577$, $\sigma_1^2 = 1/2$, $\sigma_2^2 = 1/3$, $\mu_{Z_1} = \mu_{Z_1} = 1$, $\sigma^2_{z_1} = \sigma^2_{Z_1} = 1$, and $\sigma_{z_1 z_2} = 0$. OLS denotes the usual ordinary least squares estimator, PML–SMN$_K$ denotes Pseudo-ML based on a bivariate scale mixture of K normals, SS denotes the spherically symmetric SP estimator, ML denotes MLE which exploit the information of the true distribution of the shocks, including the degrees of freedom, and MD denotes the optimum minimum distance estimator. We compute the expected value of the Hessian and variance of the score of the finite mixture-based PMLEs by means of large sample averages of the analytical expressions in Fiorentini and Sentana (2021) evaluated at the true values of the mean and variance parameters in $\boldsymbol{\theta}$ and the pseudo true values of the shape parameters, which we numerically obtain from samples of millions of simulated observations.

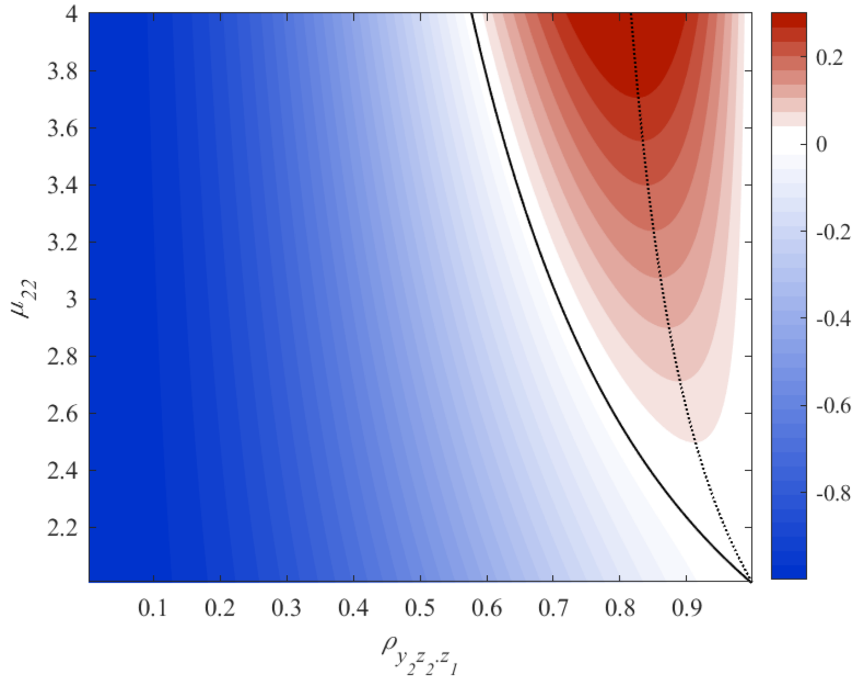
Table 2: Asymptotic variances of alternative estimators

Parameter	IV	PML–MN$_K$ $K=2$	PML–MN$_K$ $K=3$	PML–MN$_K$ $K=4$	SP	ML
			Slope parameters of equation 1a			
α	1.502	1.320	1.301	1.300	1.296	1.296
β	1.000	0.879	0.867	0.865	0.863	0.863
			Slope parameters of equation 1b			
μ_1	0.333	0.259	0.252	0.251	0.251	0.251
μ_2	0.333	0.259	0.252	0.251	0.251	0.251
			(Reparametrised) reduced form intercepts			
$E(y_1)$	0.553	0.553	0.553	0.553	0.553	0.499
$E(y_2)$	0.333	0.333	0.333	0.333	0.333	0.299
			Reduced form variance parameters			
ω_{11}	1.803	1.803	1.803	1.803	1.803	0.796
ω_{22}	0.950	0.950	0.950	0.950	0.950	0.308
ω_{12}	0.815	0.815	0.815	0.815	0.815	0.229

Notes: DGP for structural innovations: bivariate asymmetric Student t with 0 means, unit standard deviations, no correlation and shape parameters $\nu = 9.65$ and $b_i = -1$. Parameter values: $\gamma = 0.204$, $\alpha = \beta = 0.398$, $\mu_0 = 0.155$, $\mu_1 = \mu_2 = 0.577$, $\sigma_1^2 = 1/2$, $\sigma_2^2 = 1/3$, $\mu_{Z_1} = \mu_{Z_1} = 1$, $\sigma_{z_1}^2 = \sigma_{Z_1}^2 = 1$, and $\sigma_{z_1 z_2} = 0$. IV denotes the usual instrumental variables estimator, PML–MN$_K$ denotes Pseudo-ML based on a bivariate mixture of K normals, SP denotes the semiparametric estimator, ML denotes MLE which exploit the information of the true distribution of the shocks, including the degrees of freedom. Moreover, $E(y_1)$ and $E(y_2)$ are short-hand for $\gamma + \tau_1(\beta + \alpha\mu_1) + \tau_2(\alpha\mu_2)$ and $\mu_0 + \tau_1\mu_1 + \tau_2\mu_2$, respectively. We compute the expected value of the Hessian and variance of the score of the finite mixture-based PMLEs using large sample averages of the theoretical expressions in Amengual, Fiorentini and Sentana (2023) evaluated at the true values of the mean and variance parameters and the pseudo true values of the shape parameters obtained from very large samples of simulated observations.

Figures

Figure 1: Relative efficiency OLS/IV for α

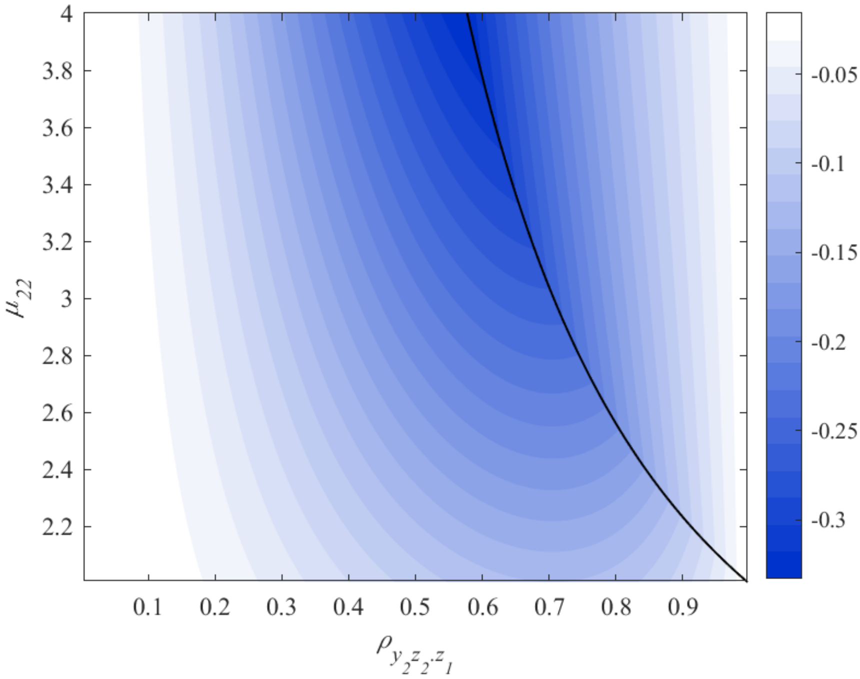


Notes: When the R^2 of equation (2) coincides with $\rho^2_{y_2 z_2 . z_1}$, the relative efficiency of the OLS/IV estimators of α is given by

$$\frac{V(\hat{\alpha}_{LS})}{V(\hat{\alpha}_{IV})} = [(1 - \rho^2_{y_2 z_2 . z_1})\mu_{22} + \rho^2_{y_2 z_2 . z_1}]\rho^2_{y_2 z_2 . z_1}.$$

The solid line denotes the boundary line $\mu_{22} = 1 + \rho^{-2}_{y_2 z_2 . z_1}$ while the dotted line denotes the locus of $(\rho_{y_2 z_2 . z_1}, \mu_{22})$ combinations for which the IV estimator of α reaches its maximum asymptotic efficiency relative to the corresponding OLS estimator, which is given by $\rho^2_{y_2 z_2 . z_1} = \frac{1}{2}\mu_{22}/(\mu_{22} - 1)$.

Figure 2: Relative efficiency MD/OLS-IV for α



Notes: When the R^2 of equation (2) coincides with $\rho^2_{y_2 z_2 . z_1}$, the relative efficiency of the MD/OLS and MD/IV estimators of α are given by

$$\frac{V(\hat{\alpha}_{MD})}{V(\hat{\alpha}_{LS})} = \frac{\mu_{22}}{[1+(\mu_{22}-1)\rho^2_{y_2 z_2 . z_1}][(1-\rho^2_{y_2 z_2 . z_1})\mu_{22}+\rho^2_{y_2 z_2 . z_1}]}$$

and

$$\frac{V(\hat{\alpha}_{MD})}{V(\hat{\alpha}_{IV})} = \frac{\mu_{22}\rho^2_{y_2 z_2 . z_1}}{1+(\mu_{22}-1)\rho^2_{y_2 z_2 . z_1}},$$

respectively. The solid line denotes the boundary line $\mu_{22} = 1 + \rho^{-2}_{y_2 z_2 . z_1}$.

Figure 3: Relative efficiency Student t ML/MD for α and β

Figure 3a: Relative efficiency ML/MD for α

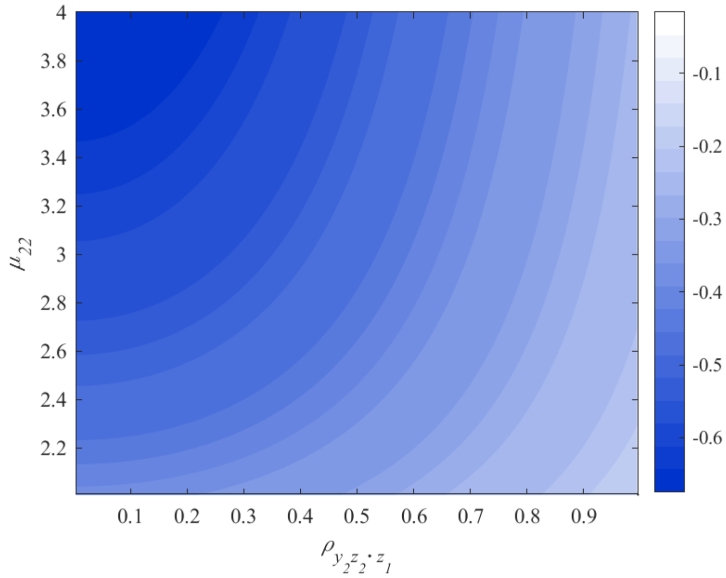
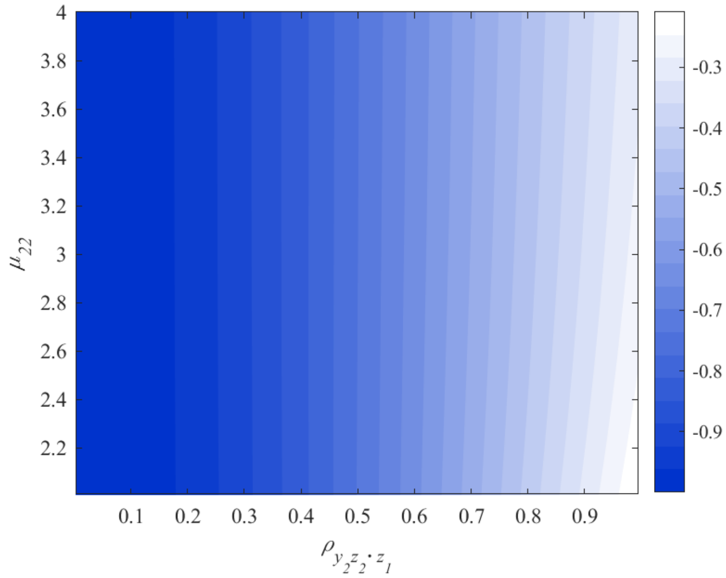


Figure 3b: Relative efficiency ML/MD for β



Notes: When the R^2 of equation (2) coincides with $\rho^2_{y_2 z_2 . z_1}$, the relative efficiency of the MD/OLS and MD/IV estimators of α and β are, respectively, given by

$$\frac{AVar(\sqrt{n}\hat{\alpha}_{MLt})}{AVar(\sqrt{n}\hat{\alpha}_{MD})} = \frac{1 + (\mu_{22} - 1)\rho^2_{y_2 z_2 . z_1}}{[(1 - \rho^2_{y_2 z_2 . z_1})\mathrm{M}_{ss} + \rho^2_{y_2 z_2 . z_1}\mathrm{M}_{ll}]\mu_{22}}$$

and

$$\frac{AVar(\sqrt{n}\hat{\beta}_{MLt})}{AVar(\sqrt{n}\hat{\beta}_{MD})} = \frac{\rho^2_{y_2 z_2 . z_1}}{[(1 - \rho^2_{y_2 z_2 . z_1})\mathrm{M}_{ss} + \rho^2_{y_2 z_2 . z_1}\mathrm{M}_{ll}]\mu_{22}},$$

where $\mathrm{M}_{ll} = \nu(2 + \nu)/[(\nu - 2)(\nu + 4)]$ and $\mathrm{M}_{ss} = (\nu + 2)/(\nu + 4)$ with $\nu = 2(2\mu_{22} - 1)/(\mu_{22} - 1)$.

Figure 4: Monte Carlo spherical data generating processes versus Gaussian distribution

Figure 4a: Bivariate standard Gaussian distribution

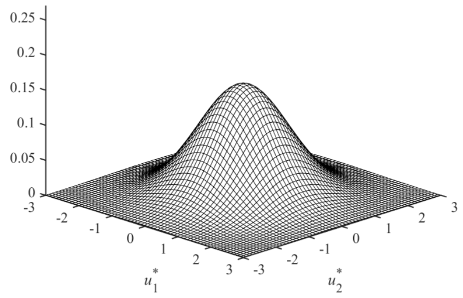


Figure 4b: Contours of a bivariate standard Gaussian distribution

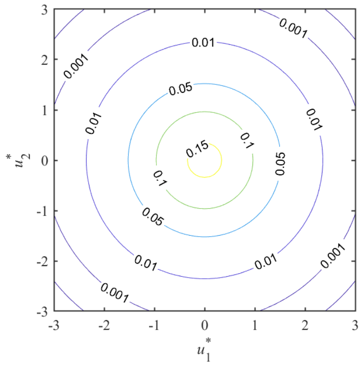


Figure 4c: Bivariate standard Student t distribution density

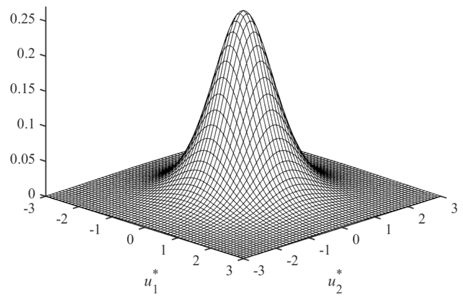


Figure 4d: Contours of a bivariate standard Student t density

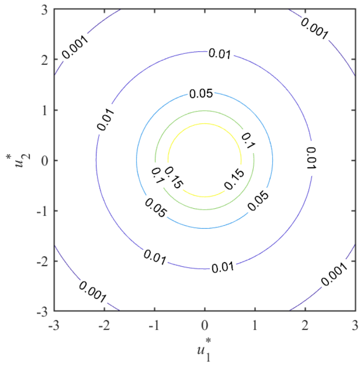


Figure 4e: Bivariate standard scale mixture of two normals

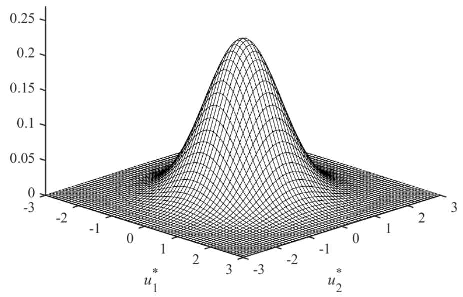
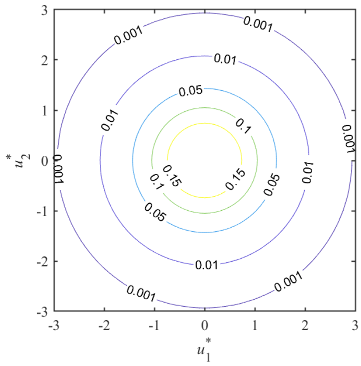


Figure 4f: Contours of a bivariate standard scale mixture of two normals



Notes: In all panels, $E(u_{1i}^*) = E(u_{2i}^*) = 0$, $V(u_{1i}^*) = V(u_{2i}^*) = 1$, and $cov(u_{1i}^*, u_{2i}^*) = 0$. Panels c-d: Student t distribution with $\nu = 5$ degrees of freedom. Panels e-f: Scale mixture of two normals with scale parameter $\varkappa = 0.09$ and mixing probability $\lambda = 0.05$.

Figure 5: Monte Carlo non-spherical data generating processes versus Gaussian distribution

Figure 5a: Bivariate standard Gaussian distribution density

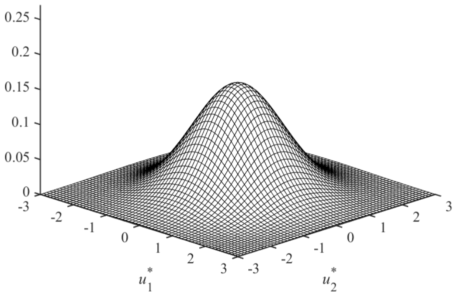


Figure 5b: Contours of a bivariate standard Gaussian distribution

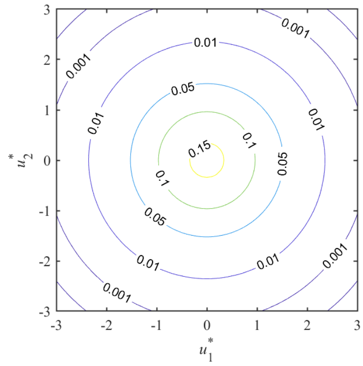


Figure 5c: Bivariate standard asymmetric Student t distribution density

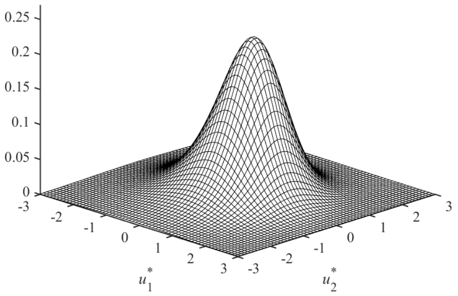


Figure 5d: Contours of a bivariate standard asymmetric Student t distribution

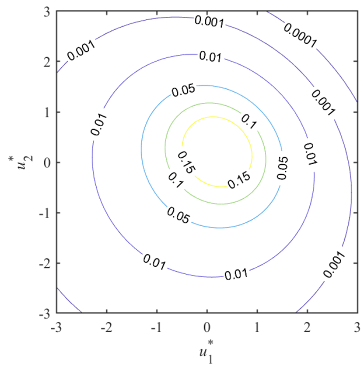


Figure 5e: Bivariate standard location-scale mixture of two normals

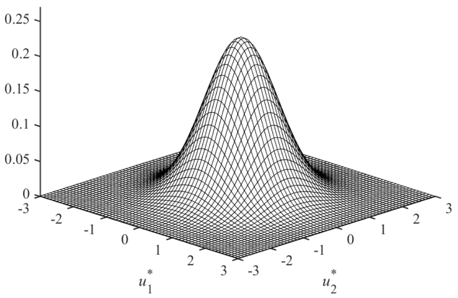
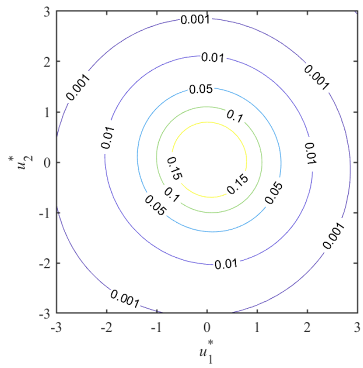


Figure 5f: Contours of a bivariate standard location-scale mixture of two normals



Notes: In all panels, $E(u_{1i}^*) = E(u_{2i}^*) = 0$, $V(u_{1i}^*) = V(u_{2i}^*) = 1$, and $cov(u_{1i}^*, u_{2i}^*) = 0$. Panels c-d: Asymmetric Student t density with $\nu = 9.65$ degrees of freedom, skewness parameters $b_i = -1$. Panels e-f: Location-scale mixture of two normals with mixing probability $\lambda = 0.05$, location vector $\boldsymbol{\delta} = -(1.01, 1.06)'$ and scale parameter $\varkappa = 0.32$ (see Appendix D for details).

Figure 6: Monte Carlo results: $T = 250$, $\mu_{22} = 3$ and $\rho_{y_2 z_2 . z_1} = (\mu_{22} - 1)^{-1}$

Figure 6a: Student t innovations

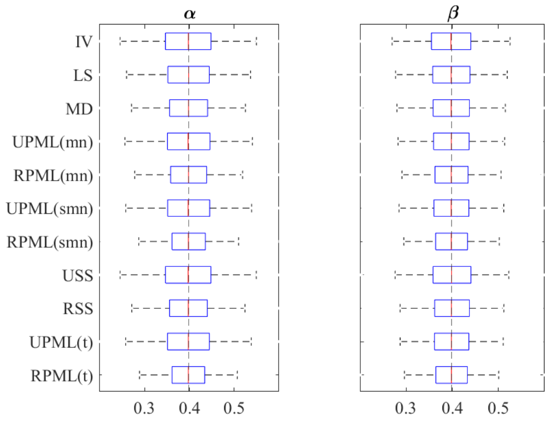


Figure 6b: Scale mixture of two normals

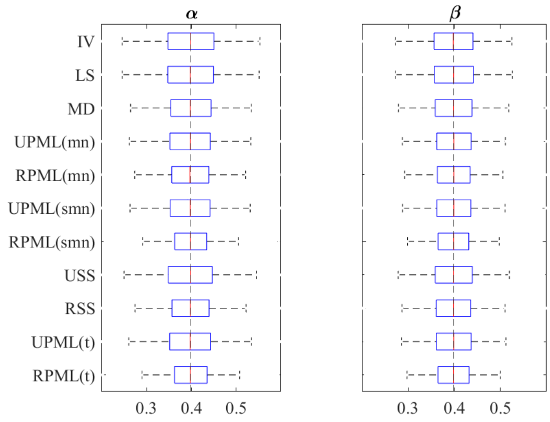


Figure 6c: Asymmetric Student t innovations

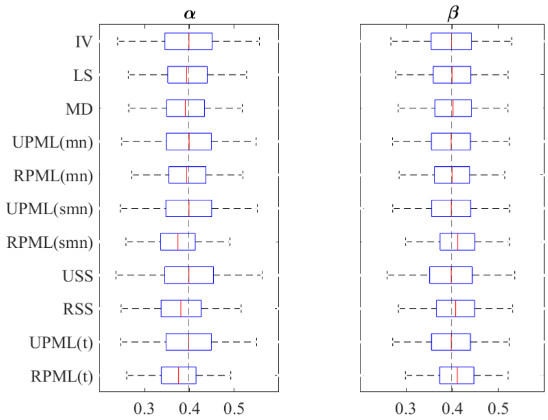
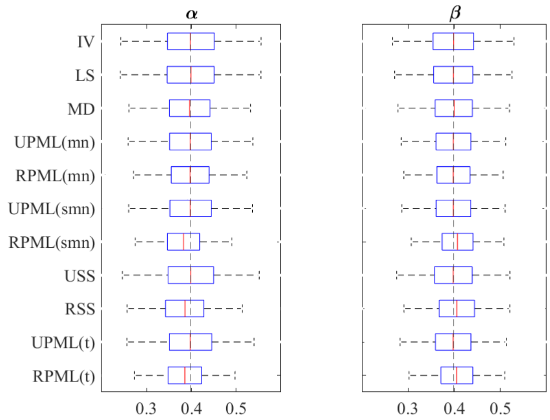


Figure 6d: Location-scale mixture of two normals



Notes: IV denotes the instrumental variables estimator, LS denotes the ordinary least squares estimator, MD denotes the optimum minimum distance estimator, UPML(mn) and RPML(mn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a mixture of two normals, UPML(smn) and RPML(smn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a scale mixture of two normals, USS and RSS denote the restricted ($\sigma_{12} = 0$) and unrestricted elliptically symmetric semiparametric estimators described in section 3, while UPML(t) and RPML(t) denote the restricted ($\sigma_{12} = 0$) and unrestricted feasible PML estimators based on a Student-t. DGPs: Panel a: Student t distribution with $\nu = 5$ degrees of freedom; Panel b: scale mixture of two normals with scale parameter $\varkappa = 0.09$ and mixing probability $\lambda = 0.05$; Panel c: asymmetric Student t density with $\nu = 9.65$ degrees of freedom, skewness parameters $b_i = -1$; and Panel d: location-scale mixture of two normals with mixing probability $\lambda = 0.05$, location vector $\boldsymbol{\delta} = -(1.01, 1.06)'$ and scale parameter $\varkappa = 0.32$ (see Appendix D for details). In all DGPs, we set $\sigma_1^2 = 1$ so that $R_1^2 = 1/2$. In order to have a tie between IV and LS, we set $\rho_{y_2 z_2 . z_1} = 1/\sqrt{2}$ so that $\sigma_2^2 = 1/3$ and, therefore, $\gamma = 0.20$, $\alpha = \beta = 0.40$, $\mu_0 = 0.15$ and $\mu_1 = \mu_2 = 0.58$.

Figure 7: Monte Carlo results: $T = 250$, $\mu_{22} = 3$ and $\rho_{y_2 z_2 . z_1} = \frac{1}{2}\mu_{22}(\mu_{22} - 1)^{-1}$

Figure 7a: Student t innovations

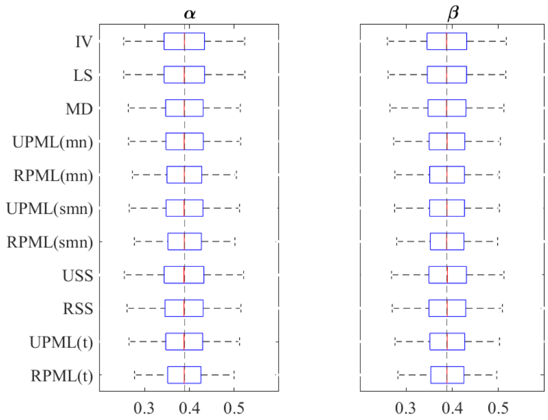


Figure 7b: Scale mixture of two normals

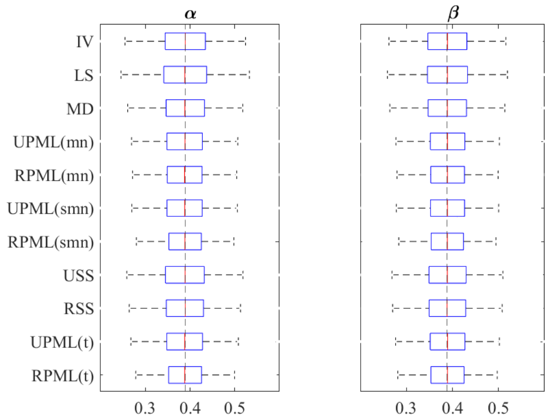


Figure 7c: Asymmetric Student t innovations

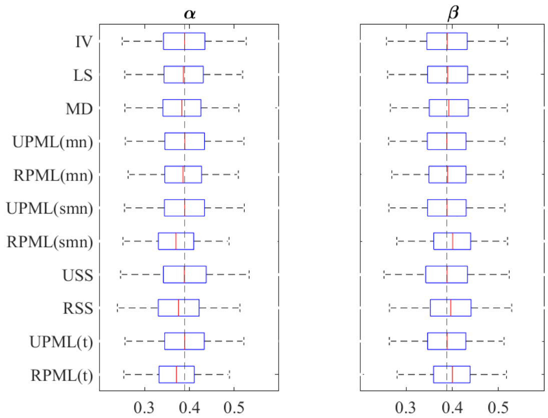
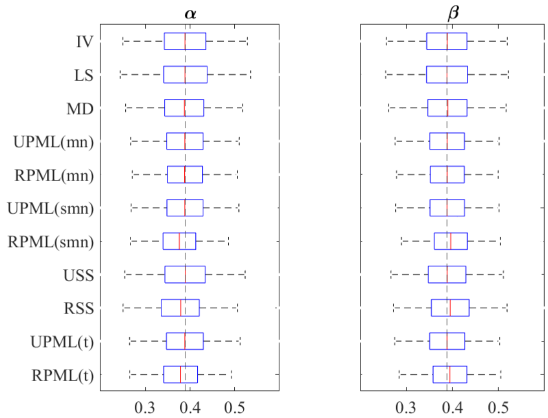


Figure 7d: Location-scale mixture of two normals



Notes: IV denotes the instrumental variables estimator, LS denotes the ordinary least squares estimator, MD denotes the optimum minimum distance estimator, UPML(mn) and RPML(mn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a mixture of two normals, UPML(smn) and RPML(smn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a scale mixture of two normals, USS and RSS denote the restricted ($\sigma_{12} = 0$) and unrestricted elliptically symmetric semiparametric estimators described in section 3, while UPML(t) and RPML(t) denote the restricted ($\sigma_{12} = 0$) and unrestricted feasible PML estimators based on a Student-t. DGPs: Panel a: Student t distribution with $\nu = 5$ degrees of freedom; Panel b: scale mixture of two normals with scale parameter $\varkappa = 0.09$ and mixing probability $\lambda = 0.05$; Panel c: asymmetric Student t density with $\nu = 9.65$ degrees of freedom, skewness parameters $b_i = -1$; and Panel d: location-scale mixture of two normals with mixing probability $\lambda = 0.05$, location vector $\boldsymbol{\delta} = -(1.01, 1.06)'$ and scale parameter $\varkappa = 0.32$ (see Appendix D for details). In all DGPs, we set $\sigma_1^2 = 1$ so that $R_1^2 = 1/2$. In order to have a maximum relative efficiency of IV versus LS, we set $\rho_{y_2 z_2 . z_1} = \sqrt{3}/2$ so that $\sigma_2^2 = 1/7$ and, therefore, $\gamma = 0.22$, $\alpha = \beta = 0.39$, $\mu_0 = 0.31$ and $\mu_1 = \mu_2 = 0.65$.

Figure 8: Monte Carlo results: $T = 250$, $\mu_{22} = 7/3$ and $\rho_{y_2 z_2 . z_1} = (\mu_{22} - 1)^{-1}$

Figure 8a: Student t innovations

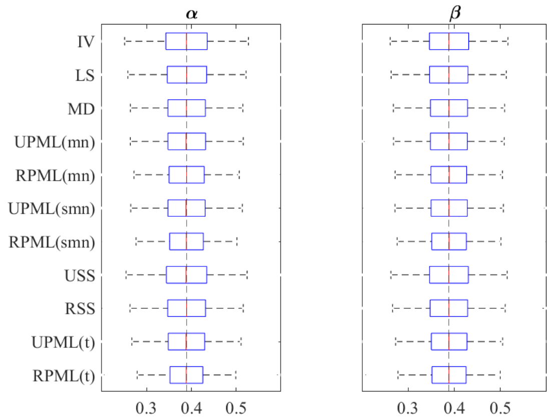


Figure 8b: Scale mixture of two normals

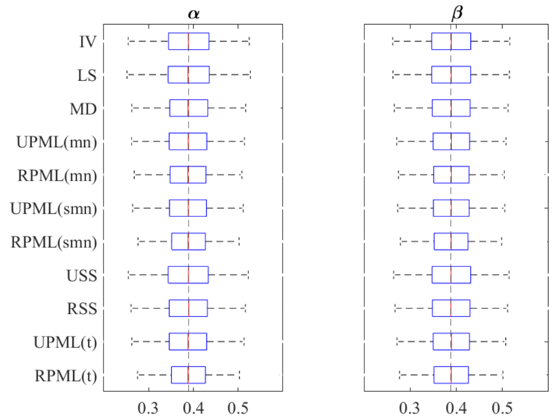


Figure 8c: Asymmetric Student t innovations

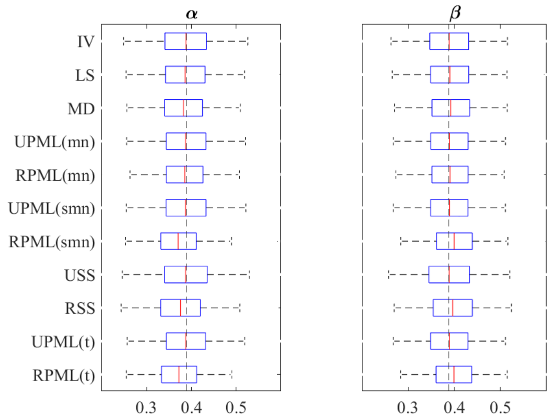
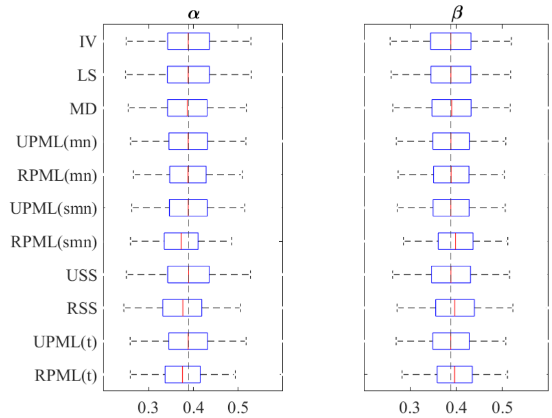


Figure 8d: Location-scale mixture of two normals



Notes: IV denotes the instrumental variables estimator, LS denotes the ordinary least squares estimator, MD denotes the optimum minimum distance estimator, UPML(mn) and RPML(mn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a mixture of two normals, UPML(smn) and RPML(smn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a scale mixture of two normals, USS and RSS denote the restricted ($\sigma_{12} = 0$) and unrestricted elliptically symmetric semiparametric estimators described in section 3, while UPML(t) and RPML(t) denote the restricted ($\sigma_{12} = 0$) and unrestricted feasible PML estimators based on a Student-t. DGPs: Panel a: Student t distribution with $\nu = 11/2$ degrees of freedom; Panel b: scale mixture of two normals with scale parameter $\varkappa = 0.12$ and mixing probability $\lambda = 0.05$; Panel c: asymmetric Student t density with $\nu = 10.38$ degrees of freedom, skewness parameters $b_i = -1$; and Panel d: location-scale mixture of two normals with mixing probability $\lambda = 0.05$, location vector $\boldsymbol{\delta} = -(1.16, 1.24)'$ and scale parameter $\varkappa = 0.38$ (see Appendix D for details). In all DGPs, we set $\sigma_1^2 = 1$ so that $R_1^2 = 1/2$. In order to have a tie between IV and LS, we set $\rho_{y_2 z_2 . z_1} = \sqrt{3}/2$ so that $\sigma_2^2 = 1/7$ and, therefore, $\gamma = 0.22$, $\alpha = \beta = 0.39$, $\mu_0 = 0.31$ and $\mu_1 = \mu_2 = 0.65$.

Figure 9: Monte Carlo results: $T = 1,000$, $\mu_{22} = 3$ and $\rho_{y_2 z_2.z_1} = (\mu_{22} - 1)^{-1}$

Figure 9a: Student t innovations

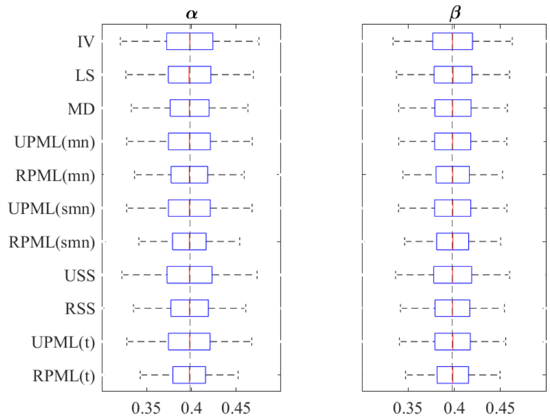


Figure 9b: Scale mixture of two normals

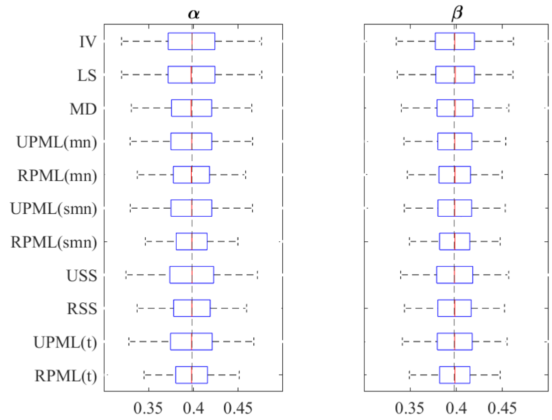


Figure 9c: Asymmetric Student t innovations

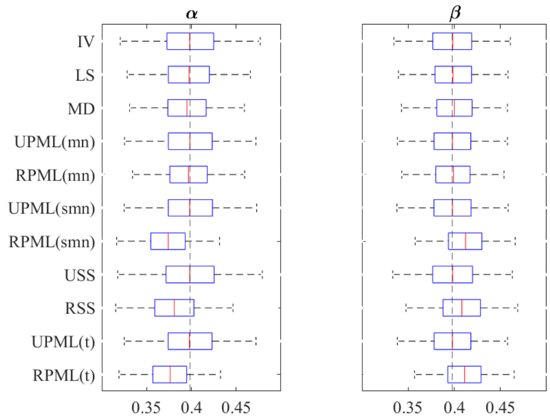
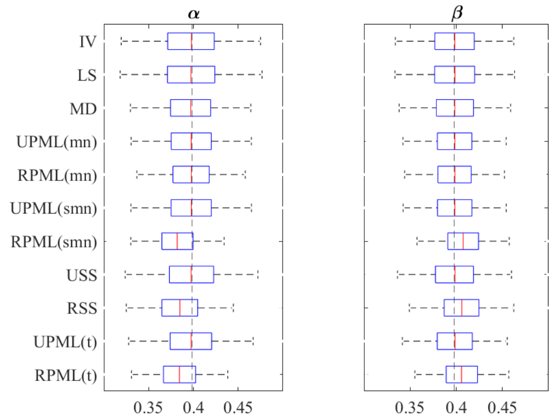


Figure 9d: Location-scale mixture of two normals



Notes: IV denotes the instrumental variables estimator, LS denotes the ordinary least squares estimator, MD denotes the optimum minimum distance estimator, UPML(mn) and RPML(mn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a mixture of two normals, UPML(smn) and RPML(smn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a scale mixture of two normals, USS and RSS denote the restricted ($\sigma_{12} = 0$) and unrestricted elliptically symmetric semiparametric estimators described in section 3, while UPML(t) and RPML(t) denote the restricted ($\sigma_{12} = 0$) and unrestricted feasible PML estimators based on a Student-t. DGPs: Panel a: Student t distribution with $\nu = 5$ degrees of freedom; Panel b: scale mixture of two normals with scale parameter $\varkappa = 0.09$ and mixing probability $\lambda = 0.05$; Panel c: asymmetric Student t density with $\nu = 9.65$ degrees of freedom, skewness parameters $b_i = -1$; and Panel d: location-scale mixture of two normals with mixing probability $\lambda = 0.05$, location vector $\boldsymbol{\delta} = -(1.01, 1.06)'$ and scale parameter $\varkappa = 0.32$ (see Appendix D for details). In all DGPs, we set $\sigma_1^2 = 1$ so that $R_1^2 = 1/2$. In order to have a tie between IV and LS, we set $\rho_{y_2 z_2.z_1} = 1/\sqrt{2}$ so that $\sigma_2^2 = 1/3$ and, therefore, $\gamma = 0.20$, $\alpha = \beta = 0.40$, $\mu_0 = 0.15$ and $\mu_1 = \mu_2 = 0.58$.

Figure 10: Monte Carlo results: $T = 1,000$, $\mu_{22} = 3$ and $\rho_{y_2 z_2 . z_1} = \frac{1}{2}\mu_{22}(\mu_{22} - 1)^{-1}$

Figure 10a: Student t innovations

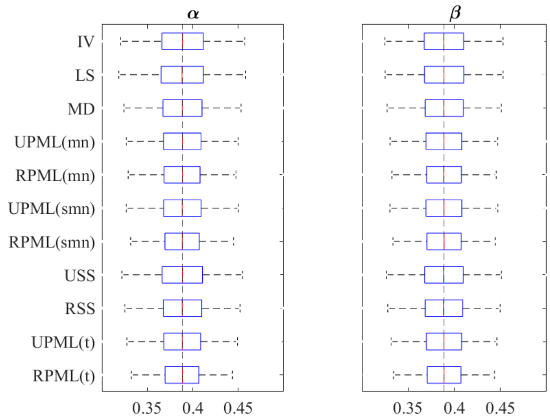


Figure 10b: Scale mixture of two normals

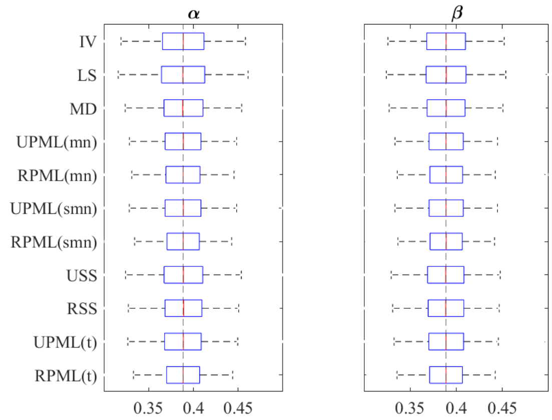


Figure 10c: Asymmetric Student t innovations

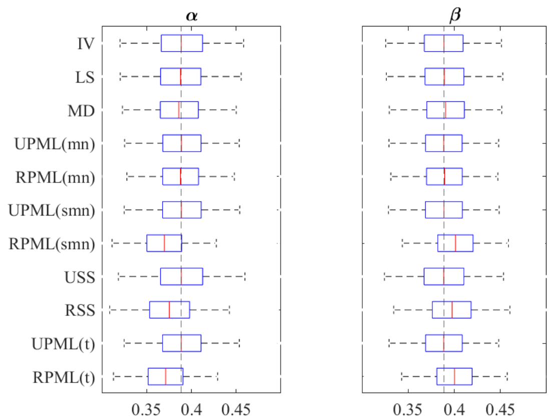
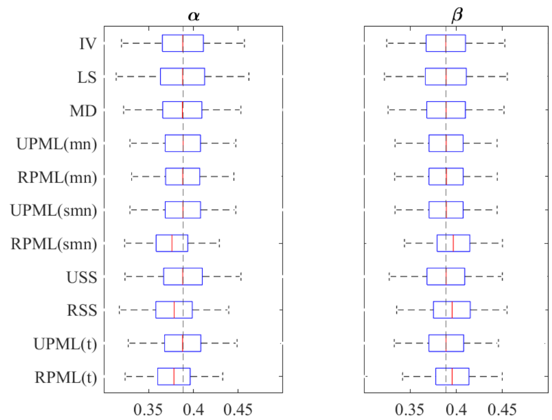


Figure 10d: Location-scale mixture of two normals



Notes: IV denotes the instrumental variables estimator, LS denotes the ordinary least squares estimator, MD denotes the optimum minimum distance estimator, UPML(mn) and RPML(mn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a mixture of two normals, UPML(smn) and RPML(smn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a scale mixture of two normals, USS and RSS denote the restricted ($\sigma_{12} = 0$) and unrestricted elliptically symmetric semiparametric estimators described in section 3, while UPML(t) and RPML(t) denote the restricted ($\sigma_{12} = 0$) and unrestricted feasible PML estimators based on a Student-t. DGPs: Panel a: Student t distribution with $\nu = 5$ degrees of freedom; Panel b: scale mixture of two normals with scale parameter $\varkappa = 0.09$ and mixing probability $\lambda = 0.05$; Panel c: asymmetric Student t density with $\nu = 9.65$ degrees of freedom, skewness parameters $b_i = -1$; and Panel d: location-scale mixture of two normals with mixing probability $\lambda = 0.05$, location vector $\boldsymbol{\delta} = -(1.01, 1.06)'$ and scale parameter $\varkappa = 0.32$ (see Appendix D for details). In all DGPs, we set $\sigma_1^2 = 1$ so that $R_1^2 = 1/2$. In order to have a maximum relative efficiency of IV versus LS, we set $\rho_{y_2 z_2 . z_1} = \sqrt{3}/2$ so that $\sigma_2^2 = 1/7$ and, therefore, $\gamma = 0.22$, $\alpha = \beta = 0.39$, $\mu_0 = 0.31$ and $\mu_1 = \mu_2 = 0.65$.

Figure 11: Monte Carlo results: $T = 1,000$, $\mu_{22} = 7/3$ and $\rho_{y_2 z_2 . z_1} = (\mu_{22} - 1)^{-1}$

Figure 11a: Student t innovations

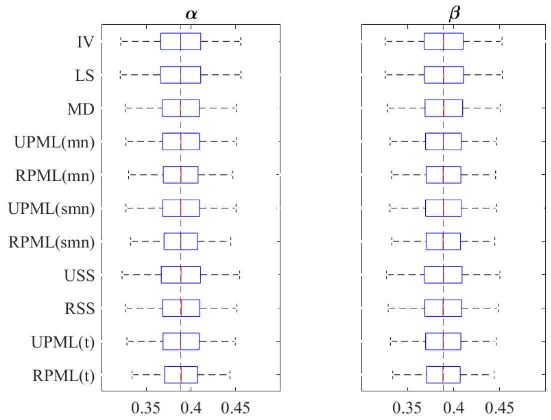


Figure 11b: Scale mixture of two normals

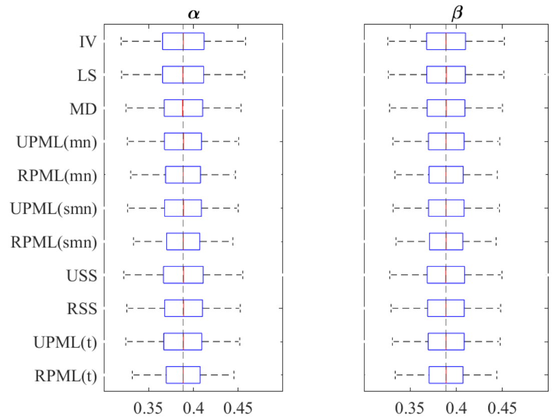


Figure 11c: Asymmetric Student t innovations

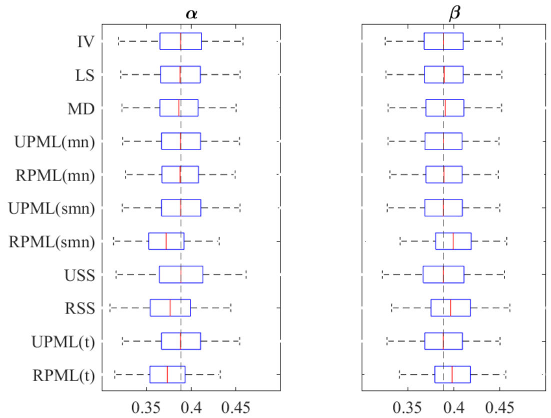
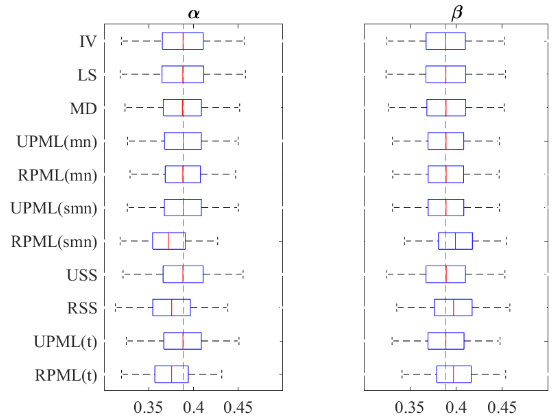


Figure 11d: Location-scale mixture of two normals



Notes: IV denotes the instrumental variables estimator, LS denotes the ordinary least squares estimator, MD denotes the optimum minimum distance estimator, UPML(mn) and RPML(mn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a mixture of two normals, UPML(smn) and RPML(smn) denote the restricted ($\sigma_{12} = 0$) and unrestricted PML estimators based on a scale mixture of two normals, USS and RSS denote the restricted ($\sigma_{12} = 0$) and unrestricted elliptically symmetric semiparametric estimators described in section 3, while UPML(t) and RPML(t) denote the restricted ($\sigma_{12} = 0$) and unrestricted feasible PML estimators based on a Student-t. DGPs: Panel a: Student t distribution with $\nu = 11/2$ degrees of freedom; Panel b: scale mixture of two normals with scale parameter $\varkappa = 0.12$ and mixing probability $\lambda = 0.05$; Panel c: asymmetric Student t density with $\nu = 10.38$ degrees of freedom, skewness parameters $b_i = -1$; and Panel d: location-scale mixture of two normals with mixing probability $\lambda = 0.05$, location vector $\boldsymbol{\delta} = -(1.16, 1.24)'$ and scale parameter $\varkappa = 0.38$ (see Appendix D for details). In all DGPs, we set $\sigma_1^2 = 1$ so that $R_1^2 = 1/2$. In order to have a tie between IV and LS, we set $\rho_{y_2 z_2 . z_1} = \sqrt{3}/2$ so that $\sigma_2^2 = 1/7$ and, therefore, $\gamma = 0.22$, $\alpha = \beta = 0.39$, $\mu_0 = 0.31$ and $\mu_1 = \mu_2 = 0.65$.