The information matrix test for Gaussian mixtures[*]

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>

January 2024
Revised: April 2024

Abstract

In incomplete data models the EM principle implies the moments the Information Matrix test assesses are the expectation given the observations of the moments it would assess were the underlying components observed. This principle also leads to interpretable expressions for their asymptotic covariance matrix adjusted for sampling variability in the parameter estimators under correct specification. Monte Carlo simulations for finite Gaussian mixtures indicate that the parametric bootstrap provides reliable finite sample sizes and good power against various misspecification alternatives. We confirm that 3-component Gaussian mixtures accurately describe cross-sectional distributions of per capita income in the 1960-2000 Penn World Tables.

Keywords: Expectation-Maximisation principle, Incomplete data, Hessian matrix, Outer product of the score.

JEL: C46, C52, O47

[*]We would like to thank Otilia Boldea, Paul Johnson, Jan Magnus and Olivier Scaillet, as well as audiences at the Barcelona Workshop in Financial Econometrics (ESADE, May 2023), the Workshop in memory of Stefano Fachin (Rome, February 2024), Duke, Florence, Montreal and JRC Ispra for useful comments, discussions and suggestions. We are also grateful to Maria Grazia Pittau for sharing her data with us, and to Tao Wang for excellent research assistance. 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

Finite mixture distributions play an important role in economics and many other disciplines, where they are often used to model unobserved heterogeneity. For example, they have been extensively employed for identifying "convergence clubs" of countries based on per capita GDP, as well as within-country clustering in household income and wealth distributions (see Johnson and Papageorgiou (2020) and Cowell and Flachaire (2015) for some recent surveys describing the use of mixtures in each of those areas).

Classical tests (i.e. Likelihood ratio, Wald and score or Lagrange Multiplier (LM)) for the number of components in a mixture are a devilish problem even if one assumes that the distribution of the components belongs to a specific parametric family because there are multiple paths converging to the null along which different parameters become increasingly underidentified (see Amengual, Bei, Carrasco and Sentana (2024) and the references therein for a detailed discussion of these unusual features when the null contains a single univariate Gaussian component).

By comparison, testing Gaussianity of the underlying components against a more flexible family of parametric distributions while maintaining that the number of components is correct would be relatively straightforward if one relied on the Expectation - Maximisation (EM) principle to obtain expressions for the scores and information matrix of the model under the alternative evaluated under the null along the lines of Almuzara, Amengual and Sentana (2019).

In this paper, in contrast, we consider a specification test for finite Gaussian mixtures which is not a priori targeted to either the number of components or their normality. Specifically, we follow Boldea and Magnus (2009), who suggested the information matrix (IM) test in their study of the score vector and Hessian matrix of the log-likelihood function of multivariate Gaussian mixtures. The IM test introduced by White (1982) directly assesses the IM equality, which states that the sum of the Hessian matrix and the outer product of the score vector should be zero in expectation when the estimated model is correctly specified.

Our approach, though, is rather different from Boldea and Magnus (2009), in that we rely on the EM principle to show that the moments underlying the IM test are the expectation given the observed data of the moments that the IM test would focus on if the underlying components were observed. But given that the influence functions underlying those moment tests effectively coincide with the list of all the distinct third- and fourth-order multivariate Hermite polynomials, as shown by Amengual, Fiorentini and Sentana (2024), the IM test for Gaussian mixtures is effectively testing that the expected value of those polynomials weighted by the posterior probability that each observation belongs to the corresponding component is simultaneously 0 for each and every underlying component of the mixture. This interpretation has two important advantages. First, it allows us to obtain the right number of degrees of freedom for the IM test, which in turn avoids the numerical calculation of Moore-Penrose inverses (see Boldea and Magnus (2024)). Second, it may prove particularly useful for the purposes of indicating in which specific directions modelling efforts to enrich finite mixture models should focus.

In fact, our approach to deriving the IM test and its interpretation applies to any model in which the observations can be viewed as incomplete data, in the sense of Dempster, Laird and Rubin (1977), so it has a much wider applicability. Examples include the limited dependent variable models that Gouriéroux, Monfort, Renault and Trognon (1987) and Smith (1987) tackled with the same approach. The EM principle also leads to interpretable expressions for the asymptotic covariance matrix of the scaled sample averages of the relevant influence functions adjusted for sampling variability in the parameter estimators under the null of correct specification.

Importantly, we explicitly address the widespread and often justified concern that the asymptotic distribution of the IM test offers a poor guide in finite samples (see Horowitz (1994) and the reference therein) by relying on bootstrap procedures. In this respect, our Monte Carlo simulations indicate that the parametric bootstrap, in combination with theoretical expressions for the asymptotic covariance matrices of the influence functions, provides reliable finite sample sizes and good power against various empirically relevant misspecification alternatives.

Finally, we apply our procedures to provide formal support to the empirical evidence in Pittau, Zelli and Johnson (2010), who argued that a Gaussian mixture with three components provides a very good fit for the cross-sectional distributions of per capita income in the Penn World Tables between 1960 and 2000.

The rest of the paper is organised as follows. In Section 2, we formally introduce the IM test, show its numerical invariance to reparametrisations, and derive its expression in a general context with incomplete data. Next, in Section 3, we apply our general result to finite mixtures of multivariate normals. Then, we present the results of some Monte Carlo exercises looking at the size and power of the tests in finite samples in Section 4, and assess the suitability of finite mixtures for cross-country distributions of GDP per capita in Section 5. We conclude in Section 6 mentioning some avenues for further research, with proofs and auxiliary results relegated to appendices.

2 The information matrix test

2.1 The test statistic

Consider a parametric model that fully characterises $\mathbf{y}$, a random vector of dimension M, as a function of $\boldsymbol{\phi}$, a p-dimensional vector of parameters, with p finite, by means of its probability distribution in the discrete case or its density in the continuous one, both of which we will simply call $f(\mathbf{y};\boldsymbol{\phi})$ henceforth.

Assuming for simplicity that sampling is random, the log-likelihood function of a sample of size N on $\mathbf{y}$ will be given by

$$L_N(\boldsymbol{\phi}) = \sum_{i=1}^{N} \ln f(\mathbf{y}_i;\boldsymbol{\phi}) = \sum_{i=1}^{N} l_i(\boldsymbol{\phi}).$$

Consequently, the average score and Hessian of this model will be given by

$$\bar{\mathbf{s}}_N(\boldsymbol{\phi}) = \frac{1}{N}\frac{\partial L_N(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial l_i(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i(\boldsymbol{\phi})$$

and

$$\bar{\mathbf{h}}_N(\boldsymbol{\phi}) = \frac{1}{N}\frac{\partial^2 L_N(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 l_i(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i(\boldsymbol{\phi}),$$

respectively. If we call $\hat{\boldsymbol{\phi}}_N$ the unrestricted maximum likelihood estimators of the parameters of interest, we will have that $\bar{\mathbf{s}}_N(\hat{\boldsymbol{\phi}}_N) = \mathbf{0}$ and $\bar{\mathbf{h}}_N(\hat{\boldsymbol{\phi}}_N)$ negative definite.

As Newey (1985) and Tauchen (1985) showed, the information matrix test can be regarded as a moment test based on the following influence functions:

$$vech[\mathbf{h}_i(\boldsymbol{\phi}) + \mathbf{s}_i(\boldsymbol{\phi})\mathbf{s}_i'(\boldsymbol{\phi})] = \mathbf{D}^{+}vec[\mathbf{h}_i(\boldsymbol{\phi}) + \mathbf{s}_i(\boldsymbol{\phi})\mathbf{s}_i'(\boldsymbol{\phi})], \tag{1}$$

where $\mathbf{D}^{+}$ is the Moore-Penrose inverse of the duplication matrix.

In practice, we need to evaluate the influence functions in (1) at $\hat{\boldsymbol{\phi}}_N$, so we need to compute the asymptotic covariance matrix of

$$\frac{\sqrt{N}}{N}\sum_{i=1}^{N} vech[\mathbf{h}_i(\hat{\boldsymbol{\phi}}_N) + \mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)\mathbf{s}_i'(\hat{\boldsymbol{\phi}}_N)]. \tag{2}$$

To do so, White (1982) relied on a standard first-order expansion of (1), which requires the calculation of third-order derivatives of $l_i(\boldsymbol{\phi})$. However, Chesher (1983) and Lancaster (1984) realised that in a likelihood context, one can use the generalised information matrix equality to obtain the expected value of the Jacobian of (1) with respect to $\boldsymbol{\phi}$ from the covariance matrix between (1) and $\mathbf{s}_i(\boldsymbol{\phi})$ evaluated at the true values of the parameters, $\boldsymbol{\phi}_0$. In effect, our *i.i.d.*

assumption means that we simply need to compute the residual covariance matrix from the least squares projection of (1) onto the linear span of $\mathbf{s}_i(\boldsymbol{\phi}_0)$, which is given by

$$\mathcal{R}(\boldsymbol{\phi}_0)-\mathcal{U}(\boldsymbol{\phi}_0)\mathcal{I}^{-1}(\boldsymbol{\phi}_0)\mathcal{U}'(\boldsymbol{\phi}_0), \tag{3}$$

where

$$\left[\begin{array}{cc} \mathcal{R}(\boldsymbol{\phi}_0) & \mathcal{U}(\boldsymbol{\phi}_0) \\ \mathcal{U}'(\boldsymbol{\phi}_0) & \mathcal{I}(\boldsymbol{\phi}_0) \end{array}\right] = V\left\{\begin{array}{c} vech[\mathbf{h}_i(\boldsymbol{\phi}_0)+\mathbf{s}_i(\boldsymbol{\phi}_0)\mathbf{s}_i'(\boldsymbol{\phi}_0)] \\ \mathbf{s}_i(\boldsymbol{\phi}_0) \end{array}\right\}.$$

Therefore, the infeasible IM test statistic will be given by the following quadratic form

$$N\left\{\frac{1}{N}\sum_{i=1}^{N} vech'[\mathbf{h}_i(\hat{\boldsymbol{\phi}}_N)+\mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)\mathbf{s}_i'(\hat{\boldsymbol{\phi}}_N)]\right\}[\mathcal{R}(\boldsymbol{\phi}_0)-\mathcal{U}(\boldsymbol{\phi}_0)\mathcal{I}^{-1}(\boldsymbol{\phi}_0)\mathcal{U}(\boldsymbol{\phi}_0)]^{+}$$
$$\times\left\{\frac{1}{N}\sum_{i=1}^{N} vech[\mathbf{h}_i(\hat{\boldsymbol{\phi}}_N)+\mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)\mathbf{s}_i'(\hat{\boldsymbol{\phi}}_N)]\right\}, \tag{4}$$

where we have relied on a Moore-Penrose generalised inverse because some of the influence functions in (1) may be an exact linear combination of $\mathbf{s}_i(\boldsymbol{\phi}_0)$ or appear multiple times.

Chesher (1983) and Lancaster (1984) suggested a feasible version of (4) as N times the R^2 in the regression of a vector of N ones onto $\mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)$ and $vech[\mathbf{h}_i(\hat{\boldsymbol{\phi}}_N)+\mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)\mathbf{s}_i'(\hat{\boldsymbol{\phi}}_N)]$ using an OLS routine robust to multicollinearity. Effectively, the inclusion of $\mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)$ as additional regressors makes the test statistic robust to the fact that the influence functions (1) are evaluated at $\hat{\boldsymbol{\phi}}_N$. Nevertheless, as explained by Horowitz (1994) and the references therein, this outer product regression has very poor finite sample properties, so in our work below we will rely on the parametric bootstrap applied to a feasible version of (4) which evaluates the theoretical expression (3) at the MLE $\hat{\boldsymbol{\phi}}_N$, as forcefully argued by Orme (1990). The theoretical results in Beran (1988) imply that if the usual Gaussian asymptotic approximation provides a reliable guide to the finite sample distribution of the sample version of the moments being tested, the bootstrapped critical values should not only be valid, but also their errors should be of a lower order of magnitude under additional regularity conditions that guarantee the validity of a higher-order Edgeworth expansion.

2.2 Numerical invariance to reparametrisations

Let us now study the effect on the IM test of reparametrising the model from $\boldsymbol{\phi}$ to $\boldsymbol{\varphi}$ by means of the one-to-one mapping $\boldsymbol{\varphi}=t(\boldsymbol{\phi})$, which we assume is a second-order continuous diffeomorphism in a neighbourhood of $\boldsymbol{\phi}_0$ whose inverse is given by $\boldsymbol{\phi}=r(\boldsymbol{\varphi})$.

As we show in the proof of the next proposition, the influence functions underlying the IM test of the reparametrised model will be

$$\left[\frac{\partial r'(\boldsymbol{\varphi})}{\partial\boldsymbol{\varphi}}\otimes\frac{\partial r'(\boldsymbol{\varphi})}{\partial\boldsymbol{\varphi}}\right] vec\left[\mathbf{h}_i(\boldsymbol{\phi})+\mathbf{s}_i(\boldsymbol{\phi})\mathbf{s}_i'(\boldsymbol{\phi})\right]+vec\left\{\mathbf{s}_i(\boldsymbol{\phi})\otimes\mathbf{I}_p\frac{\partial vec[\partial r'(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}]}{\partial\boldsymbol{\varphi}'}\right\}. \tag{5}$$

Then, we can show:

Lemma 1 *The infeasible IM test statistic in (4) which uses the influence functions (1) written in terms of $\boldsymbol{\phi}$ numerically coincides with the analogous IM test statistic that relies on the influence functions (5) written in terms of $\boldsymbol{\varphi}$.*

Intuitively, the sample average of the second summand in (5) is exactly zero when evaluated at $\hat{\boldsymbol{\varphi}}_N$, so effectively, the influence functions (5) are a linear transformation of (1). Besides, given that $\mathbf{s}_i(\boldsymbol{\phi})$ is one of the regressors, adding a linear combination of it to the regressand does not alter the residual covariance matrix.

Interestingly, the same numerical identity also holds for the feasible outer product of the score (OPS) version suggested by Chesher (1983) and Lancaster (1984) because they effectively use the sample second moments in computing the relevant residual covariance matrices. Naturally, the numerical invariance also applies to the alternative feasible version that replaces $\boldsymbol{\phi}_0$ by $\hat{\boldsymbol{\phi}}_N$ in the evaluation of the asymptotic covariance matrices.

Example: Assume that y is normally distributed with mean μ and variance σ^2 so that, in terms of the notation above, we would have $\boldsymbol{\phi} = (\mu, \sigma^2)'$,

$$\mathbf{s}(\boldsymbol{\phi}) = \left[\begin{array}{c} (y-\mu)/\sigma^2 \\ (y-\mu)^2/(2\sigma^4) - 1/\sigma^2 \end{array} \right]$$

and

$$\mathbf{h}(\boldsymbol{\phi}) = -\left[\begin{array}{cc} 1/\sigma^2 & (y-\mu)/\sigma^4 \\ (y-\mu)/\sigma^4 & (y-\mu)^2/(2\sigma^6) - 1/(2\sigma^4) \end{array} \right].$$

Now consider reparametrising the distribution of y in terms of its Sharpe ratio $\tau = \mu/\sigma$ and standard deviation $\psi = \sigma$, so that $\boldsymbol{\varphi} = (\tau, \psi)'$ and $r(\boldsymbol{\varphi}) = (\tau\psi, \psi^2)'$. Then, direct calculations deliver

$$\frac{\partial \ln g(y; \boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} = \left[\begin{array}{c} y/\psi - \tau \\ (y^2 - \tau\psi y + \psi^2)/\psi^3 \end{array} \right]$$

and

$$\frac{\partial \ln g(y; \boldsymbol{\varphi})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}'} = -\left[\begin{array}{cc} 1 & y/\psi^2 \\ y/\psi^2 & (3y^2 - 2\tau\psi y - \psi^2)/\psi^4 \end{array} \right].$$

Alternatively, starting from the score and Hessian written in terms of $\boldsymbol{\varphi}$, namely

$$\mathbf{s}[r(\boldsymbol{\varphi})] = \left[\begin{array}{c} (y-\tau\psi)/\psi^2 \\ (y-\tau\psi)^2/(2\psi^4) - 1/\psi^2 \end{array} \right]$$

and

$$\mathbf{h}_i[r(\boldsymbol{\varphi})] = -\left[\begin{array}{cc} 1/\psi^2 & (y_i-\tau\psi)/\psi^4 \\ (y_i-\tau\psi)/\psi^4 & (y_i-\tau\psi)^2/(2\psi^6) - 1/(2\psi^4) \end{array} \right],$$

and using the fact that

$$\frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} = \left[\begin{array}{cc} \psi & \tau \\ 0 & 2\psi \end{array} \right] \quad \text{and} \quad \frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}} = \left[\begin{array}{cccc} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 \end{array} \right],$$

we can easily verify through straightforward calculations that (5) holds.

2.3 The case of incomplete data

We follow Dempster, Laird and Rubin (1977) in using the term "incomplete data" to denote situations in which it is convenient to think of the observed data $\mathbf{y}$ as the output of a mapping $\mathbf{g}(.)$ from the complete sample space $\boldsymbol{Z}$ to the observed sample space $\boldsymbol{Y}$, so that the complete data $\boldsymbol{\zeta}$ is only known to lie in R, the subset of $\boldsymbol{Z}$ implicitly defined by the equation $\mathbf{y} = \mathbf{g}(\boldsymbol{\zeta})$.

Let $f(\boldsymbol{\zeta};\boldsymbol{\phi})$ denote the joint density of $\boldsymbol{\zeta}$ given a vector of parameters $\boldsymbol{\phi}$. We know from basic probability theory that

$$f(\mathbf{y};\boldsymbol{\phi}) = \int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}. \tag{6}$$

Throughout, we maintain the following regularity condition:

Assumption 1 *The boundary of R does not depend on the model parameters $\boldsymbol{\phi}$.*

Our next result provides a general approach to computing the information matrix test when the observations $\mathbf{y}$ can be viewed as incomplete data:

Proposition 1 *The influence functions (1) of the IM test of model (6) are*

$$E\left\{ vech\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\Bigg|\mathbf{y}\right\}, \tag{7}$$

with the expectation taken with respect to the conditional distribution of $\boldsymbol{\zeta}$ given $\mathbf{y}$ over R.

Proposition 1 implies we can write the influence functions underlying the IM test as the expected value conditional on the observed variables of the influence functions underlying the IM test of the complete log-likelihood. This interpretation is very convenient in those set ups in which the complete log-likelihood function adopts a particularly simple form, such as in the limited dependent variable models considered by Gouriéroux et al. (1987), who proved a special case of this expression when $f(\boldsymbol{\zeta};\boldsymbol{\phi})$ belongs to what they called a "bilinear" exponential family, and $\mathbf{y} = \mathbf{g}(\boldsymbol{\zeta})$. These include univariate probit and Tobit models among others, as well as their simultaneous equation versions studied by Smith (1987). The Gaussian mixtures in the next section provide another case in point.

To compute (4), though, we also need expressions for the different elements that appear in the theoretical expression (3).

Let $\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})$ denote a vector influence functions of the complete data $\boldsymbol{\zeta}$ such that

$$E_{\boldsymbol{\zeta}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})] = \mathbf{0}$$

when both the expectation and the influence function are evaluated at the same value of the model parameters, $\boldsymbol{\phi}$. In addition, let

$$\mathbf{m}(\mathbf{y};\boldsymbol{\phi}) = E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})].$$

The law of iterated expectations implies that $E_{\mathbf{y}}[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})] = \mathbf{0}$, which confirms the suitability of (7) to test for the correct specification of the likelihood model for the observed data. In this context, we can prove the following result, which generalises Lemma 4 in Gouriéroux et al. (1987), who focused on the case in which the latent influence functions $\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})$ coincide with $\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})/\partial\boldsymbol{\phi}$ when $f(\boldsymbol{\zeta};\boldsymbol{\phi})$ belongs to an exponential family:

Proposition 2

$$V_{\mathbf{y}}[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})] = V_{\boldsymbol{\zeta}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})] - E_{\mathbf{y}}\{V_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})]\} \tag{8}$$

and

$$\begin{aligned} E_{\mathbf{y}}\left[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right] &= -E_{\boldsymbol{y}}\left[\frac{\partial \mathbf{m}(\boldsymbol{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right] \\ = E_{\boldsymbol{\zeta}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right] &- E_{\mathbf{y}}\left\{cov_{\boldsymbol{\zeta}|\mathbf{y}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi}),\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right]\right\}. \end{aligned} \tag{9}$$

Thus, we can compute the different elements that appear in the theoretical expression (3) by applying Proposition 2 to the vector

$$\left\{vech'\left[\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right],\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right\}', \tag{10}$$

whose elements are the conditional expected values of

$$\left\{vech'\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right],\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right\}'. \tag{11}$$

Corollary 1 *The application of Proposition 2 to (10) yields*

$$\begin{aligned} \mathcal{I}(\boldsymbol{\phi}) &= V_{\mathbf{y}}\left[\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right] = V_{\boldsymbol{\zeta}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right] - E_{\mathbf{y}}\left\{V_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right]\right\} \\ &= -E_{\boldsymbol{\zeta}}\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}\right] - E_{\mathbf{y}}\left\{V_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right]\right\}, \end{aligned} \tag{12}$$

$$\begin{aligned} \mathcal{U}(\boldsymbol{\phi}) &= E_{\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right\} \\ &= cov_{\boldsymbol{\zeta}}\left\{vech\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right],\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right\} \\ &-E_{\mathbf{y}}\left[cov_{\boldsymbol{\zeta}|\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right],\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right\}\right], \end{aligned} \tag{13}$$

and

$$\begin{aligned} \mathcal{R}(\boldsymbol{\phi}) &= V_{\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\right\} \\ &= V_{\boldsymbol{\zeta}}\left\{vech\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\right\} \\ &-E_{\mathbf{y}}\left[V_{\boldsymbol{\zeta}|\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\right\}\right]. \end{aligned} \tag{14}$$

Once again, the advantage of this procedure is that, in many instances, the complete model is much simpler to work with than the observed one, something that we illustrate in the next section with normal mixtures.

3 Finite Gaussian mixtures

3.1 Definition

Let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_k, \ldots, \xi_K)$ denote a categorical random variable of dimension K, which is nothing other than a collection of K mutually exclusive Bernoulli random variables with $\Pr(\xi_k = 1) = \lambda_k$ such that $\sum_{k=1}^{K} \lambda_k = 1$. If $\boldsymbol{\varepsilon}|\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I}_M)$, $\boldsymbol{\nu}_k$ is an $M \times 1$ vector and $\boldsymbol{\Gamma}_k$ an $M \times M$ positive definite matrix with $\boldsymbol{\gamma}_k = vech(\boldsymbol{\Gamma}_k)$, then

$$\mathbf{y} = \sum_{k=1}^{K} \xi_k(\boldsymbol{\nu}_k + \boldsymbol{\Gamma}_k^{1/2}\boldsymbol{\varepsilon}) \tag{15}$$

is an M-variate, K-component mixture of normals, whose first two unconditional moments are

$$\boldsymbol{\tau} = E(\mathbf{y}) = \sum_{k=1}^{K} \lambda_k \boldsymbol{\nu}_k = E_{\boldsymbol{\xi}}[E_{\mathbf{y}|\boldsymbol{\xi}}(\mathbf{y})], \quad \text{and} \tag{16}$$

$$\boldsymbol{\Psi} = V(\mathbf{y}) = \sum_{k=1}^{K} \lambda_k[(\boldsymbol{\nu}_k\boldsymbol{\nu}_k') + \boldsymbol{\Gamma}_k] - \left(\sum_{k=1}^{K} \lambda_k \boldsymbol{\nu}_k\right)\left(\sum_{k=1}^{K} \lambda_k \boldsymbol{\nu}_k'\right) = E_{\boldsymbol{\xi}}[V_{\mathbf{y}|\boldsymbol{\xi}}(\mathbf{y})] + V_{\boldsymbol{\xi}}[E_{\mathbf{y}|\boldsymbol{\xi}}(\mathbf{y}\mathbf{y}')]. \tag{17}$$

The natural model parameters are the mean vectors and covariance matrices of the components $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_k, \ldots, \boldsymbol{\nu}_K)'$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k, \ldots, \gamma_K)'$, respectively, and their probabilities $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k, \ldots, \lambda_K)$, which are subject to the unit simplex restrictions $\lambda_k \geq 0 \; \forall k$ and $\sum_{k=1}^{K} \lambda_k = 1$. These restrictions can be imposed in different ways. For example, one could use the multinomial logit parametrisation

$$\lambda_k = \frac{e^{\pi_k}}{\sum_{l=1}^{K-1} e^{\pi_l} + 1} \; (k = 1, \ldots K-1); \quad \lambda_K = \frac{1}{\sum_{l=1}^{K-1} e^{\pi_l} + 1}, \tag{18}$$

or one could make

$$\lambda_k = \pi_k, \; k = 1, \ldots, K-1 \text{ and } \lambda_K = 1 - \textstyle\sum_{l=1}^{K-1} \pi_l \tag{19}$$

and impose the inequality restrictions $\pi_k \geq 0$ $(k = 1, \ldots, K-1)$ and $\sum_{l=1}^{K-1} \pi_l \leq 1$ in estimation. Nevertheless, many of the expressions below are considerably simpler if we work with the K elements of $\boldsymbol{\lambda}$ rather than the $K-1$ elements of $\boldsymbol{\pi}$. As a result, the Jacobian matrix $\partial\boldsymbol{\lambda}/\partial\boldsymbol{\pi}'$ will play an important role in the practical implementation of our IM tests, as we explain in section 4. However, the choice of parametrisation is inconsequential because Lemma 1 implies that the IM test statistics are numerically invariant.[1] For that reason, in a slight abuse of notation we shall use $\boldsymbol{\phi} = (\boldsymbol{\nu}', \boldsymbol{\gamma}', \boldsymbol{\lambda}')'$ to denote the model parameters.

[1] Trivially, the IM test that we derive below will also be numerically invariant to a relabelling of the components of the mixture, as this only involves a reordering of the parameters.

3.2 Influence functions

The log-density for $\mathbf{y}$ is given by

$$l(\mathbf{y};\boldsymbol{\phi}) = \ln\left\{\sum_{k=1}^{K}\lambda_k|\boldsymbol{\Gamma}_k|^{-1/2}\phi_M[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\right\}, \tag{20}$$

where $\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) = \boldsymbol{\Gamma}_k^{-1/2}(\mathbf{y}-\boldsymbol{\nu}_k)$, with $\boldsymbol{\theta}_k = (\boldsymbol{\nu}'_k, \boldsymbol{\gamma}'_k)'$, and $\phi_M(.)$ the M-variate spherical normal density. Theorem 1 in Boldea and Magnus (2009) contains detailed expressions for the score and Hessian of (20) when the mixing probabilities are parametrised as in (19) (see also Appendix D for details). But a simpler and more intuitive way of obtaining the required expressions for (1) is by using the EM-based formulas in Proposition 1, with the observed data being $\mathbf{y}_i$ for $i = 1, \ldots, N$ and the complete data $\boldsymbol{\zeta}_i = (\mathbf{y}'_i, \boldsymbol{\xi}'_i)$. Thus, we can show that:

Proposition 3 *The sum of the Hessian and the outer product of the scores corresponding to a single observation* $\mathbf{y}$ *is a block diagonal matrix whose only non-zero elements are*

$$\begin{aligned}
\partial\boldsymbol{\nu}_k\partial\boldsymbol{\nu}'_k \quad &: \quad w_k(\boldsymbol{\phi})\boldsymbol{\Gamma}_k'^{-1/2}[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]\boldsymbol{\Gamma}_k^{-1/2}, && (21)\\
\partial\boldsymbol{\nu}_k\partial\boldsymbol{\gamma}'_k \quad &: \quad w_k(\boldsymbol{\phi})\frac{1}{2}\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M \\
&\quad -w_k(\boldsymbol{\phi})[\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M, && (22)\\
\partial\boldsymbol{\nu}_k\partial\lambda_k \quad &: \quad w_k(\boldsymbol{\phi})\frac{1}{\lambda_k}\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k), && (23)\\
\partial\boldsymbol{\gamma}_k\partial\boldsymbol{\gamma}'_k \quad &: \quad w_k(\boldsymbol{\phi})\frac{1}{4}\mathbf{D}'_M(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M] \\
&\quad \times vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M \\
&\quad -w_{ki}(\boldsymbol{\phi})\frac{1}{2}\mathbf{D}'_M\{2[(\boldsymbol{\Gamma}_k^{-1}\otimes\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}] - (\boldsymbol{\Gamma}_k^{-1}\otimes\boldsymbol{\Gamma}_k^{-1})\}\mathbf{D}_M, && (24)\\
\partial\boldsymbol{\gamma}_k\partial\lambda_k \quad &: \quad w_{ki}(\boldsymbol{\phi})\frac{1}{2\lambda_k}\mathbf{D}'_M(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M], && (25)
\end{aligned}$$

where $w_k(\boldsymbol{\phi})$ *represents the posterior probability that* $\mathbf{y}$ *comes from the* k^{th} *component given the parameter values, so that*

$$w_k(\boldsymbol{\phi}) = E(\xi_k|\mathbf{y};\boldsymbol{\phi}) = \Pr(\xi_k = 1|\mathbf{y};\boldsymbol{\phi}) = \frac{\lambda_k|\boldsymbol{\Gamma}_k|^{-1/2}\phi_M[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]}{\sum_{l=1}^{K}\lambda_l|\boldsymbol{\Gamma}_l|^{-1/2}\phi_M[\boldsymbol{\varepsilon}_l^*(\boldsymbol{\theta}_l)]}. \tag{26}$$

However, not all those elements can be used as influence functions of the IM test. First, (23) will be zero at the ML estimators because this vector is proportional to the score with respect to $\boldsymbol{\nu}_k$, whose expression appears in Appendix C. Similarly, (21) and (25) will also be zero because they are linear combinations of the score vector with respect to $\boldsymbol{\gamma}_k$ presented in the same appendix. Therefore, we are left with (22) and (24), which contain $\frac{1}{2}M^2(M+1)$ and $\frac{1}{8}M(M+1)(M^2+M+2)$ distinct influence functions, respectively. Unfortunately, those expressions still include redundant elements, what suggests the use of generalised inverses (see Boldea and Magnus (2024)). Nevertheless, the calculation of the strictly necessary influence functions, its asymptotic covariance matrix and the correct number of degrees of freedom can be further simplified on the basis of the following result, which avoids generalised inverses:

Proposition 4 *1. The IM matrix test based on (22) and (24) evaluated at the MLEs of the model parameters numerically coincides with a moment test based on the influence functions:*

$$w_k(\boldsymbol{\phi})\left\{\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right\},\quad k=1,\ldots,K \tag{27}$$

evaluated at the same estimators, where

$$\mathbf{H}_j(\boldsymbol{\varepsilon}^*)=\left[\begin{array}{c}H_{j,0,\cdots,0}(\boldsymbol{\varepsilon}^*)\\ H_{j-1,1,\cdots,0}(\boldsymbol{\varepsilon}^*)\\ \vdots\\ H_{0,\cdots,0,j}(\boldsymbol{\varepsilon}^*)\end{array}\right]=\left[\begin{array}{c}H_j(\varepsilon_1^*)\\ H_{j-1}(\varepsilon_1^*)H_1(\varepsilon_2^*)\\ \vdots\\ H_j(\varepsilon_M^*)\end{array}\right]$$

is the $\binom{M+j-1}{j}$ vector containing the distinct multivariate Hermite polynomials of order j of a standardised random vector $\boldsymbol{\varepsilon}^*$ in Appendix B, which can be expressed as products of the corresponding univariate Hermite polynomials of its elements.

2. The asymptotic covariance matrix of (27) corrected for the sampling uncertainty in estimating the model parameters under the null is the residual covariance matrix in the multivariate theoretical regression of (27) on

$$w_k(\boldsymbol{\phi})\left\{\begin{array}{c}1\\ \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right\},\quad k=1,\ldots,K. \tag{28}$$

3. If the effective number of components is K, then the asymptotic distribution of the IM test will be a χ^2 random variable with degrees of freedom equal to

$$\frac{KM(M+1)(M+2)(M+7)}{24}. \tag{29}$$

Although the IM test is often regarded as a black box, Proposition 4 provides a simple and intuitive moment test interpretation in which the influence functions are the distinct multivariate Hermite polynomials of orders 3 and 4 of $\mathbf{y}$ standardised using the mean vector and covariance matrix of the k^{th} component of the mixture and weighted by the posterior probability that it belongs to that component. Thus, this result provides a direct generalisation of Proposition 1 in Amengual, Fiorentini and Sentana (2024), which corresponds to the special case of $K=1$.

To provide additional intuition, let us focus on the univariate case. It is easy to see that the sum of the Hessian and OPS yields

$$\partial\nu_k\partial\gamma_k^2 \;:\; w_k(\boldsymbol{\phi})\frac{1}{2\gamma_k^3}[\varepsilon^{*3}(\boldsymbol{\theta}_k)-3\varepsilon^*(\boldsymbol{\theta}_k)]=\frac{1}{2\gamma_k^3}E(\xi_k|y;\boldsymbol{\phi})H_3[\varepsilon^*(\boldsymbol{\theta}_k)], \tag{30}$$

$$(\partial\gamma_k^2)^2 \;:\; w_k(\boldsymbol{\phi})\frac{1}{4\gamma_k^4}[\varepsilon^{*4}(\boldsymbol{\theta}_k)-6\varepsilon^{*2}(\boldsymbol{\theta}_k)+3]=\frac{1}{4\gamma_k^4}E(\xi_k|y;\boldsymbol{\phi})H_4[\varepsilon^*(\boldsymbol{\theta}_k)], \tag{31}$$

so the influence functions the IM test checks coincide with the third and fourth Hermite polynomials of the observed variable y standardised as if it belonged to the k^{th} component of the mixture, as shown by White (1982) for $K = 1$, but weighted by $w_k(\boldsymbol{\phi})$, the posterior probability that it belongs to that component.

The ease of interpretation of the influence functions in Proposition 4 allows one to immediately derive tests that focus on a subset of them, such as those involving the third- or fourth-order Hermite polynomials of a single component, which may prove particularly useful for the purposes of indicating in which specific directions modelling efforts to enrich the estimated model should focus. By choosing the relevant elements of the residual covariance matrix, the computation of the corresponding test statistics would be straightforward.

3.3 The asymptotic covariance matrix

Proposition 4 states the asymptotic covariance matrix of the influence functions involved, but it does not explain how we can compute it. Given that we can obtain in closed form the covariance matrix of multivariate Hermite polynomials using the results in Rahman (2017), we can use the law of iterated variances implicit in (12), (13) and (14) to obtain expressions for the three elements of (3). Specifically, we can use expression (8) to write

$$\begin{aligned}\mathcal{R}_{kj}(\boldsymbol{\phi}) &= cov_{\mathbf{y}}\left[w_k(\boldsymbol{\phi})\left\{\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right\}, w_j(\boldsymbol{\phi})\left\{\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{array}\right\}\right] \qquad (32)\\ &= cov_{\boldsymbol{\zeta}}\left\{\xi_k\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right), \xi_j\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{array}\right)\right\}\\ &-E_{\mathbf{y}}\left[cov\left\{\xi_k\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right), \xi_j\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{array}\right)\right\}\middle|\mathbf{y}\right],\end{aligned}$$

where

$$\begin{aligned}&E_{\mathbf{y}}\left[cov\left\{\xi_k\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right), \xi_j\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{array}\right)\middle|\mathbf{y}\right\}\right]\\ &= E_{\mathbf{y}}\left[cov(\xi_k,\xi_j|\mathbf{y})\left(\begin{array}{cc}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_3'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_4'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_3'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_4'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{array}\right)\right],\end{aligned}$$

with $cov(\xi_k, \xi_j|\mathbf{y}) = [I(j = k)w_k(\boldsymbol{\phi}) - w_k(\boldsymbol{\phi})w_j(\boldsymbol{\phi})]$.

In turn, we also know that at the true values

$$\begin{aligned}&cov_{\boldsymbol{\zeta}}\left\{\xi_k\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right), \xi_j\left(\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{array}\right)\right\}\\ &= E_{\boldsymbol{\zeta}}\left[\xi_k\xi_j\left(\begin{array}{cc}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_3'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_4'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_3'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_4'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{array}\right)\right] = I(j = k)\lambda_k\left(\begin{array}{cc}\mathbf{M}_3 & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_4\end{array}\right)\end{aligned}$$

because $\xi_k\xi_j = 0$ when $k \neq j$, $\xi_k^2 = \xi_k$, $\boldsymbol{\varepsilon}_k^*(\boldsymbol{\theta}_k) = \boldsymbol{\varepsilon}$ when $\xi_k = 1$ from (15), which is independent of ξ_k, and the third and fourth multivariate Hermite polynomials of a standard normal variable have zero means, are uncorrelated, and have covariances matrices $\mathbf{M}_3$ and $\mathbf{M}_4$, respectively, which adopt a particularly simple form regardless of the model parameters (see e.g. Lemma 2 in Amengual, Fiorentini and Sentana (2024)). As a result, it must be the case that

$$\mathcal{R}_{kj}(\boldsymbol{\phi}) = I(j=k)\lambda_k \begin{pmatrix} \mathbf{M}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_4 \end{pmatrix}$$
$$-E_{\mathbf{y}}\left[[I(j=k)w_k(\boldsymbol{\phi})-w_k(\boldsymbol{\phi})w_j(\boldsymbol{\phi})]\begin{pmatrix} \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_3'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_4'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] \\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_3'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_4'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] \end{pmatrix}\right]. \qquad (33)$$

In principle, one might expect the sample version of (33) to be less noisy than the sample version of (32) in finite samples. Nevertheless, both expressions involve the same weighted averages of the sixth, seventh and eighth powers of the elements of $\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)$, the only difference being whether they are scaled by $[I(j=k)w_k(\boldsymbol{\phi}) - w_k(\boldsymbol{\phi})w_j(\boldsymbol{\phi})]$ or $w_k(\boldsymbol{\phi})w_j(\boldsymbol{\phi})$. In addition, a combination of the sample version of (33) with the theoretical values of $\mathbf{M}_3$ and $\mathbf{M}_4$ could lead to indefinite estimated covariance matrices. For that reason, our suggestion would be either to compute the above expressions analytically using quadrature, in which case both calculations yield the same result up to machine precision, or to rely on the centred or uncentred sample versions of (32), as Chesher (1983) and Lancaster (1984) suggested.

We can use a similar procedure to obtain the covariances of (27) with (28), which we can use to purge those influence functions from the sampling variability arising from the ML estimation of the mixture model parameters. Specifically, we can exploit the fact that

$$E_{\boldsymbol{\zeta}}\left\{\xi_k \begin{pmatrix} \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \end{pmatrix} \xi_j \begin{pmatrix} 1 & \mathbf{H}_1'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_2'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] \end{pmatrix}\right\} = \mathbf{0}$$

for all k and j to show that

$$\mathcal{U}_{kj}(\boldsymbol{\phi}) = cov_{\mathbf{y}}\left[w_k(\boldsymbol{\phi})\begin{pmatrix} \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \end{pmatrix}, w_j(\boldsymbol{\phi})\begin{pmatrix} 1 \\ \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] \\ \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] \end{pmatrix}\right]$$
$$=-E_{\mathbf{y}}\left\{cov(\xi_k, \xi_j|\mathbf{y})\begin{pmatrix} \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] & \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_1'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] & \mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_2'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] & \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_1'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] & \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_2'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \end{pmatrix}\right\},$$

but again, it is not clear which expression leads to less noisy estimates in finite samples.

Finally, we can use an entirely analogous procedure to compute

$$\mathcal{I}_{kj}(\boldsymbol{\phi}) = cov_{\mathbf{y}}\left[w_k(\boldsymbol{\phi})\begin{pmatrix}1\\ \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\\ \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{pmatrix}, w_j(\boldsymbol{\phi})\begin{pmatrix}1\\ \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{pmatrix}\right]$$

$$= I(j=k)\lambda_k\begin{pmatrix}1 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_M & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{M}_2\end{pmatrix}$$

$$-E_{\mathbf{y}}\left[cov(\xi_k,\xi_j|\mathbf{y})\begin{pmatrix}1 & \mathbf{H}_1'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_2'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] & \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_1'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_2'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\\ \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] & \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_1'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)] & \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\mathbf{H}_2'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_j)]\end{pmatrix}\right],$$

where $\mathbf{M}_2 = \mathbf{D}_M'(\mathbf{I}_{M^2} + \mathbf{K}_{MM})\mathbf{D}_M$ and $\mathbf{K}_{mn}$ is the commutation matrix of orders m and n (see e.g. Magnus and Neudecker (2019)). In this respect, one important thing to note is that the expected value of $w_k(\boldsymbol{\phi})$ is not 0 but λ_k, which explains why we should compute the expected value of the second moments of (28) rather than their covariance matrix. However, this is inconsequential because working with the second moment matrix of those K vectors is effectively adding a constant to the theoretical regression mentioned in Proposition 4, which makes no difference to the theoretical calculations because both (27) and the remaining elements of (28) have all 0 mean under the null. In fact, the same argument implies that in the list of regressors we can replace without loss of generality the K components corresponding to the zero-order Hermite polynomials times the posterior probabilities by the $K-1$ scores of the underlying parameters $\boldsymbol{\pi}$ that characterise the prior probabilities in (18) or (19). Intuitively, given that both regressands and regressors have 0 means under the null of correct specification, the regression residuals with and without constant are identical, and therefore so is their covariance matrix.

3.4 Computational considerations

3.4.1 Initial values

To maximise (20) numerically, it is usually convenient to start the recursions from sensibly chosen values. In this respect, the EM algorithm discussed by Dempster, Laird and Rubin (1977) allows us to obtain initial values as close to the MLEs as desired. The recursions are as follows:

$$\hat{\boldsymbol{\nu}}_k^{(h)} = \frac{1}{\hat{\lambda}_k^{(h)}}\frac{1}{N}\sum_{i=1}^{N} w_{ki}(\boldsymbol{\nu}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}, \boldsymbol{\lambda}^{(h-1)})\mathbf{y}_i, \tag{34a}$$

$$\hat{\boldsymbol{\Gamma}}_k^{(h)} = \frac{1}{\hat{\lambda}_k^{(h)}}\frac{1}{N}\sum_{i=1}^{N} w_{ki}(\boldsymbol{\nu}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}, \boldsymbol{\lambda}^{(h-1)})\mathbf{y}_i\mathbf{y}_i' - \hat{\boldsymbol{\nu}}_k^{(h)}\hat{\boldsymbol{\nu}}_k^{(h)\prime}, \quad \text{and} \tag{34b}$$

$$\hat{\lambda}_k^{(h)} = \frac{1}{N}\sum\nolimits_{i=1}^{N} w_{ki}(\boldsymbol{\nu}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}, \boldsymbol{\lambda}^{(h-1)}), \tag{34c}$$

Given that (26) is homogeneous of degree zero in $\boldsymbol{\lambda}$, in principle these posterior probabilities are compatible with values of $\boldsymbol{\lambda}$ outside the unit simplex. Nevertheless, a useful property of the EM algorithm is that it automatically imposes the relevant inequality restrictions on the estimators of $\boldsymbol{\lambda}$ because

$$\textstyle\sum_{k=1}^{K} w_k(\boldsymbol{\phi}) = 1 \text{ for all } \mathbf{y} \text{ and for all } \boldsymbol{\phi}.$$

Still, the EM algorithm might get stuck in at least two situations. First, when one starts the recursions up with $\boldsymbol{\nu} = \bar{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_K$ and $\boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}} \otimes \boldsymbol{\iota}_K$, where $\bar{\boldsymbol{\nu}}$ is $M \times 1$, $\bar{\boldsymbol{\gamma}}$ $M(M+1)/2$, and $\boldsymbol{\iota}_K$ a vector of K ones, in which case $w_k(\bar{\boldsymbol{\nu}} \otimes \boldsymbol{\iota}_K, \bar{\boldsymbol{\gamma}} \otimes \boldsymbol{\iota}_K, \boldsymbol{\lambda}) = \lambda_k$ for all k, so the parameter values will not get updated because priors and posteriors coincide. One way of avoiding this problem is to use a fast numerical clustering algorithm to choose the initial values of the $\boldsymbol{\nu}_k' s$ with which to start the EM recursions. The second undesirable situation arises when a linear combination of the mean vector of one component coincides with the same linear combination of $\mathbf{y}_i$ for some i. Given that the corresponding linear combination of $\mathbf{y}_i - \boldsymbol{\nu}_k$ will be zero in that case, if we choose it as the eigenvector associated to the smallest eigenvalue of $\boldsymbol{\Gamma}_k$, and take this to zero while λ_k goes to $1/N$, the log-likelihood function will become unbounded. To avoid those poles, we systematically impose that $\lambda_k \geq 2/N$ for all k.

Unfortunately, the EM algorithm slows down considerably in the neighbourhood of the optimum, so it makes sense to switch to a quadratically convergent algorithm based on first and possibly second derivatives or the expected values of the latter, whose analytical expressions we provide in Appendix C. In this context, it is convenient to work with the Cholesky decomposition of the $\boldsymbol{\Gamma}_k$ matrices to ensure that they remain positive definite.

3.4.2 Invariance to affine transformations

Consider the following full-rank affine transformation $\mathbf{x} = \mathbf{c} + \mathbf{D}\mathbf{y}$ with $|\mathbf{D}| \neq 0$. It is clear that the transformed random vector continues to be a finite mixture of K multivariate normals with mean vectors $\mathbf{c} + \mathbf{D}\boldsymbol{\nu}_k$ and covariance matrices $\mathbf{D}\boldsymbol{\Gamma}_k\mathbf{D}'$ $(k = 1, \ldots, K)$. Our next result shows that the IM statistic is numerically invariant to the values of $\mathbf{c}$ and $\mathbf{D}$:

Lemma 2 *The IM test statistics of model (15) and the analogous one for* $\mathbf{x}$ *numerically coincide.*

This numerical invariance is not only a desirable property in itself, but it also implies that the sample mean vector and covariance matrix of the observations do not affect the null distribution of our proposed test in finite samples. In fact, we can exploit Lemma 2 to simplify the calculation of the IM statistic as follows. First, as we explain in Appendix E, we can always reparametrise the model in terms of the unconditional mean vector and covariance matrix on the one hand,

and the shape parameters of a standardised version of the mixture distribution on the other. One computational advantage of this procedure is that we reduce the number of parameters to be estimated by $M(M+3)/2$ because the results in Day (1969) imply that the joint ML estimators of $\boldsymbol{\tau}$ and $\boldsymbol{\Psi}$ numerically coincide with the sample mean and covariance matrix (with denominator N) of the observations. As a result, the criterion function maximized with respect to the shape parameters $\boldsymbol{\tau}$, $\aleph$ and $\boldsymbol{\lambda}$ keeping $\boldsymbol{\tau}$ and $\boldsymbol{\Psi}$ fixed at those restricted ML estimators coincides with the criterion function maximized over all five groups of parameters.

4 Monte Carlo simulations

As stated in Proposition 4, the asymptotic distribution of our proposed IM test is χ^2 with degrees of freedom equal to (29). However, this asymptotic approximation might not be very reliable in finite samples. For that reason, we conduct some Monte Carlo experiments to study the finite sample sizes for $N = 100$, $N = 400$ and $N = 1600$. For each of the data generating processes (DGPs) we describe below, we generate $10,000$ samples under the null. When assessing size, we compare the OPS version of the statistic proposed by Chesher (1983) and Lancaster (1984) and employed by Boldea and Magnus (2009) with the feasible version of the theoretical expression (33) that replaces the true parameter values $\boldsymbol{\phi}_0$ with their MLEs $\hat{\boldsymbol{\phi}}_T$. In all cases, we consider not only asymptotic critical values but also a parametric bootstrap procedure in which we simulate $B = 99$ samples from the mixture model estimated under the null.

We also investigate the power properties of our test by considering three types of alternatives:

1. mixtures with the same number of non-Gaussian components,
2. mixtures with a larger number of Gaussian components, and
3. non-mixture distributions.

We do so by looking at the rejection rates from $2,500$ samples of size $N = 100$ and $N = 400$ that use the aforementioned bootstrap critical values to correct the finite sample size distortions, as forcefully argued by Horowitz and Savin (2000).

Given that the true model parameters are unknown, it is important to estimate them accurately. For that reason, we first run the EM algorithm up to a pre-specified convergence level starting with $\boldsymbol{\lambda} = K^{-1}\boldsymbol{\iota}_K$, $\boldsymbol{\Gamma}_k = dg[\hat{V}_T(\mathbf{y})]$, and initial values for $\boldsymbol{\nu}_k$ which maximise the log-likelihood function among those obtained from multiple runs of the k-means++ algorithm of Arthur and Vassilvitskii (2007) with random initial draws for the cluster centres. Next, we switch to a quadratically convergent quasi-Newton routine written in terms of the $\boldsymbol{\pi}'s$ in (19) and the Cholesky factors of the $\boldsymbol{\Gamma}_k's$ with a tighter convergence level, ensuring that we avoid the log-likelihood poles we mentioned in section 3.4.1 by imposing $2/N \leq \lambda_k \leq 1-2/N$ for all k. We can then use Propositions 3 and 4 to compute the feasible version of the IM test statistic in (4) that takes into account the sampling uncertainty in estimating the mixture model parameters under the null of correct specification. In this respect, Proposition 2 and Corollary 1 allow us to obtain "closed-form" expressions for the covariance matrix of the influence functions involved, as well as their covariances with the log-likelihood scores, and the information matrix, where by "closed-form" we mean "up to a definite integral".

In the univariate case, we consider Gaussian mixtures of two and three components as null hypotheses. For the 2-component case, we follow Robertson and Fryer (1969) in generating the mixture with the "bitangential" probability density function (pdf) in Figure 1a, which coincides with the borderline case between unimodal and bimodal densities. Specifically, we set the means and variances of the components to 1/4 and 1/2, and 1/256 and 3/64, respectively, with a mixing probability for the first component of 0.646. The rejection rates we obtain using asymptotic critical values (see Panel A of Table 1) confirm the need for finite sample size adjustments, especially for the OPS version of the IM test. As Orme (1990) indicated, the quality of the asymptotic approximation is much better when one uses the theoretical expressions for the weighting matrix instead, as can be seen for samples of size $N = 1,600$. In contrast, Panel B of Table 1, which contains the bootstrap-based rejection rates, gives a completely different picture: sizes are very accurate and almost all Monte Carlo rejection rates fall within the relevant 95% confidence set.[2] For these reasons, to assess power we focus on the bootstrap version of the IM test statistic that relies on the theoretical expression for the asymptotic covariance matrix evaluated at the MLEs.

As the first alternative hypothesis to the 2-component Gaussian mixture in Figure 1a, we consider a mixture of two asymmetric Student t's with the same means, variances and mixing probability as under the null, but with shape parameters $\eta_1 = \eta_2 = 1/12$, $\beta_1 = 5$ and $\beta_2 = -5$ (see Mencía and Sentana (2012) for details). In addition, we consider a symmetric mixture of three normals which represents a borderline case between unimodal and trimodal density. Specifically, we set the means of the underlying components to -0.47, 0.47 and 0, their variances to 0.047, 0.047 and 0.018, and the mixing probabilities for the first two components to 0.18. Finally, the empirical application to "convergence clubs" in cross-country GDP per capita in section 5 suggests a lognormal distribution with parameters $\mu = -1/4$ and $\sigma^2 = 1$ as our third alternative. Figures 1b-d show the corresponding densities (solid lines), as well as the pdf of the closest (in the usual Kullback - Leibler sense) mixture of two normals (dashed lines). As can be

[2] Given the number of replications, the 95% asymptotic confidence intervals for the Monte Carlo rejection probabilities under the null are (.80,1.20), (4.57,5.43) and (9.41,10.59) at the 1%, 5% and 10% levels, respectively.

seen from Panel C of Table 1, the IM test is able to detect with reasonable power these three deviations from the null, especially for the larger sample size.

As our second null hypothesis, we consider a mixture of three normals whose parameter values are in line with the estimates we obtain in the empirical application in the next section (see Figure 1e). Specifically, we set the means of the underlying components to 3, 1 and 1/4, their variances to 2/5, 1/5 and 1/100, and the mixing probabilities for the first and second components to 0.25 and 0.45, respectively. As Panels A and B of Table 2 indicate, the same qualitative comments apply regarding the size of the different versions of the IM test in finite samples, with the only exception that the asymptotic critical values continue to lead to significant size distortions even for samples of size $1,600$. Intuitively, the quality of the asymptotic approximation to the finite sample distribution of the parameter estimators is lower for the 3-component mixture than for the 2-component one for any given sample size.

The first alternative hypothesis we consider in this 3-component Gaussian case is a mixture of two asymmetric Student t's and a symmetric one with the same degrees of freedom and skewness parameters as in Figure 1b for the first two components, and with the same mixing probabilities we use for the null. In addition, we consider a mixture of four normals with means 4, 2, 1, and 1/3, variances 2, 1/2, 1/10 and 0.015, and mixing probabilities 0.075, 0.25 and 0.325 for the first three components. Finally, we retain the same lognormal as in the 2-component mixture as an example of a non-Gaussian distribution which does not correspond to some finite mixture. Figures 1f-h show the corresponding densities (solid lines) as well as the pdf of the closest mixture of three normals (dashed lines). The rejection rates reported in Panel C of Table 2 show that the IM test continues to have good power, although there is a clear decrease when the true distribution is lognormal relative to the 2-component case, which simply reflects that fact that a 3-component Gaussian mixture does a much better job in approximating the same log-normal distribution than a 2-component one.

Given that the bootstrap takes considerable more CPU time in the bivariate case, we only consider as null hypothesis the two-component Gaussian mixture in Boldea and Magnus (2009), which is fully characterised by

$$\boldsymbol{\nu}_1 = \mathbf{0},\ \boldsymbol{\nu}_2 = 5\boldsymbol{\iota}_2,\ \boldsymbol{\Gamma}_1 = \mathbf{I}_2,\ \boldsymbol{\Gamma}_2 = \mathbf{I}_2 + \boldsymbol{\iota}_2\boldsymbol{\iota}_2',$$

and a mixing probability of 1/2. The pdf and contours of this density are depicted in Figures 2a and 2e, respectively. In Panels A and B of Table 3 we report the rejection rates under the null based on asymptotic critical values and bootstrapped ones, respectively. The same comments as in the univariate examples apply, but with the OPS version performing noticeably worse in this case. Interestingly, the size distortions of the other versions of the IM test are of the same order of magnitude as in the univariate examples despite the higher number of estimated parameters and much higher number of influence functions involved. Presumably, the reason is that the two components are much more clearly separated in the Boldea and Magnus (2009) design than in the univariate design in Figure 1a, which makes both the asymptotic covariance matrix of the influence functions and the information matrix closer to being block diagonal.

As for the alternatives, we first consider a mixture of two asymmetric bivariate Student t's with the same means, variances and mixing probability as under the null, but with shape parameters $\eta_1 = \eta_2 = 1/16$, and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = -(1,1)'$ (see again Mencía and Sentana (2012) for details).[3] In addition, we consider a discrete mixture of three normals with

$$\boldsymbol{\nu}_1 = 3\iota_2,\ \boldsymbol{\nu}_2 = \boldsymbol{\iota}_2,\ \boldsymbol{\nu}_3 = \frac{1}{4}\iota_2,\ \boldsymbol{\Gamma}_1 = \frac{2}{5}\mathbf{I}_2,\ \boldsymbol{\Gamma}_2 = \frac{1}{5}\mathbf{I}_2,\ \boldsymbol{\Gamma}_3 = \frac{1}{10}\mathbf{I}_2,$$

and mixing probabilities 0.25 and 0.45 for the first two components. Finally, as an example of a bivariate non-Gaussian distribution which cannot be expressed as a finite mixture we simulate two independent (standardised) univariate skew normals with a skewness parameter such that its skewness and kurtosis coefficients are -0.85 and 3.71, respectively (see Azzalini (1985) for details). Figures 2b-d show the corresponding pdfs while in Figures 2f-h we report their contours (solid lines) as well as those of the Gaussian mixtures of two components that best match those densities in the usual Kullback-Leibler sense (dashed lines). The rejections rates displayed in Panel C of Table 3 indicate that the IM test is also able to detect deviations from the null in all these bivariate experiments. As can be seen, the highest power is obtained when the alternative is a mixture of three normals and the lowest under the bivariate skew normal alternative. In addition, power is always quite close to one for the larger sample size.

5 Empirical application

As mentioned in the introduction, Gaussian mixtures feature pre-eminently in the empirical literature on "convergence clubs" in cross-country GDP per capita comparisons. In this section, we revisit the empirical application in Pittau et al. (2010), who found that a Gaussian mixture with three components provides a very good fit for the distributions of per capita income in version 6.1 of the Penn World Tables for 1960, 65, 70, etc. all the way to the year 2000. This covers 102 countries, of which 90 have data over the entire sample span.

In addition, they found that the within-group variances of both the rich and poor groups

[3] If we chose the same shape parameters as in the univariate alternative in Figure 1b, then we would systematically obtain rejection rates close to 100% even for $N = 100$.

of countries decreased over time, while the distance between their means increased, especially between the middle-income and high-income groups.

Finally, they found that the sizes of the different groups fluctuated somewhat, but with little movements across components, as judged by the posterior probabilities. These features can be seen in Panel A of Table 4 in which we report the parameter estimates, and also in Figure 3, which displays the temporal evolution of those cross-sectional distributions.

However, the validity of the results in Pittau et al. (2010) and their interpretation crucially depend on finite Gaussian mixtures with three components providing an accurate description of those distributions. For that reason, we apply the IM test that we have studied in previous sections to their data set, whose p-values, both based on asymptotic critical values and 9,999 bootstrapped samples, we report in Panel B of Table 4. As can be seem, the null hypothesis of correct specification is never rejected, which provides formal empirical support to their claim.[4]

6 Conclusions and directions for further research

We explain how the EM principle applied to incomplete data can also be used to obtain the moments underlying the IM test as the expectation given the observed data of the moments tested if the complete data were observed. This principle also leads to interpretable expressions for the asymptotic covariance matrix of those influence functions adjusted for the sampling uncertainty in the parameter estimators under the null of correct model specification.

We then apply these results to finite mixtures of Gaussian random vectors, showing that the IM test statistic can be easily computed as a quadratic form in the sample means of the K vectors that contain the distinct third- and fourth-order multivariate Hermite polynomials of the observations standardised with respect to the vector of means and covariance matrix of each of the underlying components multiplied by the posterior probability of those components, with a weighting matrix which is the inverse of the residual covariance matrix in the regression of those influence functions on the K vectors that contains the distinct zero-, first-, and second-order multivariate Hermite polynomials of the same standardised variables multiplied again by the posterior probability of the components.

Our procedures could be trivially extended to deal with restricted Gaussian mixture models. For example, the delta method would immediately give us the score, Hessian and relevant influence functions and their asymptotic covariance matrix in a model in which the covariance matrices of the components were assumed to be the same.

[4] In contrast, the IM test applied to 2-component mixtures estimated with the same data systematically rejects at the 5% level.

Our Monte Carlo exercises clearly indicate that one can substantially reduce size distortions in finite samples by using the theoretical expressions for the aforementioned weighting matrix evaluated at the MLEs rather than the OPS version of the IM test statistic put forward by Chesher (1983) and Lancaster (1984), and that a parametric bootstrap procedure practically eliminates them. Our results also confirm the non-trivial power of the IM tests against many empirically plausible alternatives.

Nevertheless, the IM test is not consistent because it will show trivial power against admittedly contrived alternatives with the right number of components in which the distribution of some of the components is not Gaussian but the expected value of all their third- and fourth-order Hermite polynomials are 0.

Finally, we employ the IM test to confirm that a Gaussian mixture with three components provides a very good fit for the cross-sectional distributions of per capita income in the Penn World Tables between 1960 and 2000, as argued by Pittau et al. (2010).

From a theoretical point of view, it would interesting to extend the Bartlett identities tests proposed by Chesher, Dhaene, Gouriéroux and Scaillet (1999) to incomplete data situations. In the context of finite Gaussian mixtures, in particular, we would expect the influence functions to coincide with the fifth- and higher-order multivariate Hermite polynomials of the observations standardised with respect to the vector of means and covariance matrix of each of the underlying components multiplied by the posterior probability of those components.

The IM tests that we present in this paper can also be extended in at least three empirically relevant directions. First, we could deal with switching regression models in which the linear regression coefficients depend of a set of predetermined variables $\mathbf{x}$. The main difference is that for each component of the mixture, we would have influence functions related to the conditional heteroskedasticity of the (multivariate) regression, the conditional skewness of its residuals, as well as their unconditional asymmetry and kurtosis. Numerical quadrature, though, would no longer be feasible unless we make an assumption about the marginal distribution of the predetermined regressors, something which is plausible in autoregressive processes. Second, we could allow the probabilities of the different regimes to be a function of some exogenous indicators using a multinomial logit model. And third, we could allow the regimes to have a Markovian structure, as in Hamilton (1989), which would force us to rely on a smoother rather than a filter, as in Almuzara et al. (2019). We are currently pursuing these interesting research avenues.

References

Almuzara, T., D. Amengual and E. Sentana (2019): “Normality tests for latent variables”, *Quantitative Economics* 10, 981–1017.

Amengual, D., Bei, X., Carrasco, M. and Sentana, E. (2024): “Score-type tests for normal mixture models”, forthcoming in the *Journal of Econometrics.*

Amengual, D., Fiorentini, G. and Sentana, E. (2024): “Multivariate Hermite polynomials and information matrix tests”, forthcoming in *Econometrics and Statistics.*

Arthur, D. and Vassilvitskii, S. (2007): “K-means++: the advantages of careful seeding”, SODA ‘07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035.

Azzalini, A. (1985): “A class of distributions which includes the normal ones”, *Scandinavian Journal of Statistics* 12, 171–178.

Barndorff-Nielsen, O. and Petersen, B.V. (1979): “The bivariate Hermite polynomials up to order six”, *Scandinavian Journal of Statistics* 6, 127–128.

Beran, R. (1988): “Prepivoting test statistics: a bootstrap view of asymptotic refinements”, *Journal of the American Statistical Association* 83, 687–697.

Boldea, O. and Magnus, J.R. (2009): “Maximum likelihood estimation of the multivariate normal mixture model”, *Journal of the American Statistical Association* 104, 1539–1549.

Boldea, O. and Magnus, J.R. (2024): Corrigendum to “Maximum likelihood estimation of the multivariate normal mixture model”, mimeo, Tilburg University.

Chesher, A. (1983): “The information matrix test: simplified calculation via a score test interpretation”, *Economics Letters* 13, 45–48.

Chesher, A. (1984): “Testing for neglected heterogeneity”, *Econometrica* 52, 865–872.

Chesher, A., Dhaene, G., Gouriéroux, C. and Scaillet, O. (1999): “Bartlett identities tests”, CREST Discussion Paper 9932.

Cowell, F.A. and Flachaire, E. (2015): “Statistical methods for distributional analysis”, in A.B. Atkinson and F. Bourguignon (eds), *Handbook of Income Distribution* vol. 2A, 359–465, Elsevier

Day, N.E. (1969): “Estimating the components of a mixture of normal distributions”, *Biometrika* 56, 463-474.

Dempster, A., N. Laird, and D. Rubin (1977): “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society Series B* 39, 1–38.

Gouriéroux, C., A. Monfort, E. Renault and A. Trognon (1987): “Generalized residuals”, *Journal of Econometrics* 34, 5–32.

Hamilton, J.D. (1989), "A new approach to the economic analysis of nonstationary time series and the business cycle", *Econometrica* 57, 357–384.

Horowitz, J. (1994): "Bootstrap-based critical values for the information matrix test", *Journal of Econometrics* 61, 395–411.

Horowitz, J. and N.E. Savin (2000): "Empirically relevant critical values for hypothesis tests: a bootstrap approach", *Journal of Econometrics* 95, 375–389.

Johnson, P. and Papageorgiou, C. (2020): "What remains of cross-country convergence?", *Journal of Economic Literature* 58, 129–175

Lancaster, A. (1984): "The covariance matrix of the information matrix test", *Econometrica* 52, 1051–1053.

Louis, T.A. (1982): "Finding the observed information matrix when using the EM algorithm", *Journal of the Royal Statistical Society Series B* 44, 226–233.

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

Mencía, J. and Sentana, E. (2012): "Distributional tests in multivariate dynamic models with Normal and Student t innovations", *Review of Economics and Statistics* 94, 133–152.

Newey, W.K. (1985): "Maximum likelihood specification testing and conditional moment tests", *Econometrica* 53, 1047–70.

Orme, C. (1990): "The small-sample performance of the information-matrix test", *Journal of Econometrics* 46, 309–331.

Pittau, M.G., Zelli R. and Johnson, P.A. (2010): "Mixture models, convergence clubs, and polarization", *Review of Income and Wealth* 56, 102–122.

Rahman, S. (2017): "Wiener–Hermite polynomial expansion for multivariate Gaussian probability measures", *Journal of Mathematical Analysis and Applications* 454, 303–334.

Robertson, C.A. and Fryer, J.G. (1969): "Some descriptive properties of normal mixtures", *Scandinavian Actuarial Journal* 1969, 137–146.

Ruud, P. (2000): *An introduction to classical econometric theory*, Oxford University Press.

Smith, R.J. (1987): "Testing the normality assumption in multivariate simultaneous limited dependent variable models", *Journal of Econometrics* 34, 105–123.

Tauchen, G. (1985): "Diagnostic testing and evaluation of maximum likelihood models", *Journal of Econometrics* 30, 415–443.

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

Appendices

A Proofs

A.1 Proof or Lemma 1

The following relationships will prove useful:

$$\begin{aligned} g(\mathbf{y}_i;\boldsymbol{\varphi}) &= f[\mathbf{y}_i; r(\boldsymbol{\varphi})], \\ \frac{\partial \ln g(\mathbf{y}_i;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} &= \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \frac{\partial l_i(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \mathbf{s}_i(\boldsymbol{\phi}) \end{aligned} \tag{A1}$$

and

$$\begin{aligned} \frac{\partial^2 \ln g(\mathbf{y}_i;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}'} &= \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \frac{\partial^2 l_i(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} \frac{\partial r(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}'} + \left[\frac{\partial l_i(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} \otimes \mathbf{I}_p\right] \frac{\partial vec[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}'} \\ &= \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \mathbf{h}_i(\boldsymbol{\phi}) \frac{\partial r(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}'} + \left[\mathbf{s}_i'(\boldsymbol{\phi}) \otimes \mathbf{I}_p\right] \frac{\partial vec[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}'}. \end{aligned}$$

As a result, the influence functions underlying the IM test of the reparametrised model will be

$$\begin{aligned} &\frac{\partial^2 \ln g(\mathbf{y}_i;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}'} + \frac{\partial \ln g(\mathbf{y}_i;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \frac{\partial \ln g(\mathbf{y}_i;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}'} \\ &= \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \left[\mathbf{h}_i(\boldsymbol{\phi}) + \mathbf{s}_i(\boldsymbol{\phi})\mathbf{s}_i'(\boldsymbol{\phi})\right] \frac{\partial r(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}'} + \left[\mathbf{s}_i'(\boldsymbol{\phi}) \otimes \mathbf{I}_p\right] \frac{\partial vec[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}'}, \end{aligned}$$

which after vectorisation become (5).

But

$$\begin{aligned} vec\left\{\left[\mathbf{s}_i'(\boldsymbol{\phi}) \otimes \mathbf{I}_p\right] \frac{\partial vec[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}'}\right\} &= \left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}} \otimes \mathbf{I}_p\right\} vec\left[\mathbf{s}_i'(\boldsymbol{\phi}) \otimes \mathbf{I}_p\right] \\ = \left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}} \otimes \mathbf{I}_p\right\} &(\mathbf{I}_p \otimes \mathbf{K}_{p1} \otimes \mathbf{I}_p) \left\{vec\left[\mathbf{s}_i'(\boldsymbol{\phi})\right] \otimes vec(\mathbf{I}_p)\right\} \\ &= \left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}} \otimes \mathbf{I}_p\right\} \{\mathbf{s}_i(\boldsymbol{\phi}) \otimes vec(\mathbf{I}_p)\} \\ &= \left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}} \otimes \mathbf{I}_p\right\} [\mathbf{I}_p \otimes vec(\mathbf{I}_p)]\mathbf{s}_i(\boldsymbol{\phi}) \end{aligned} \tag{A2}$$

by virtue of theorem 3.10 in Magnus and Neudecker (2019) and the fact that $\mathbf{s}_i(\boldsymbol{\phi})$ is already a vector, $\mathbf{K}_{p1} = \mathbf{I}_p$, and

$$\{\mathbf{s}_i(\boldsymbol{\phi}) \otimes vec(\mathbf{I}_p)\} = vec\left\{vec(\mathbf{I}_p)\mathbf{s}_i'(\boldsymbol{\phi})\right\} = [\mathbf{I}_p \otimes vec(\mathbf{I}_p)]\mathbf{s}_i(\boldsymbol{\phi}).$$

Therefore, (5) can be written as an (admittedly complex) linear combination of (5) and $\mathbf{s}_i(\boldsymbol{\phi}_0)$.

In fact, if we ignored the additional term (A2), the residual covariance matrix in the least squares projection of

$$\left[\frac{\partial r'(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}} \otimes \frac{\partial r'(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}}\right] vec\left[\mathbf{h}_i(\boldsymbol{\phi}_0) + \mathbf{s}_i(\boldsymbol{\phi}_0)\mathbf{s}'_i(\boldsymbol{\phi}_0)\right]$$

onto the linear span of (A1) evaluated at $\boldsymbol{\varphi}_0$ will be given by

$$\left[\frac{\partial r'(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}} \otimes \frac{\partial r'(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}}\right] [\mathcal{R}(\boldsymbol{\phi}_0) - U(\boldsymbol{\phi}_0)\mathcal{I}^{-1}(\boldsymbol{\phi}_0)U(\boldsymbol{\phi}_0)] \left[\frac{\partial r(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}'} \otimes \frac{\partial r(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}'}\right].$$

The inclusion of the additional term (A2), though, does not affect this residual covariance matrix because it is a linear combination of $\mathbf{s}_i(\boldsymbol{\phi}_0)$, and consequently, of (A1) evaluated at $\boldsymbol{\varphi}_0$ too. □

Lemma 1 is perhaps not entirely surprising given Chesher's (1984) re-interpretation of the IM test as an LM test against neglected parameter heterogeneity, because LM tests computed with either the information matrix or the OPS are numerically invariant to reparametrisation, as explained in section 17.4 of Ruud (2000).

A.2 Proof of Proposition 1

Given that Assumption 1 allows us to interchange integration and differentiation, we can follow Louis (1982) in exploiting (6) to obtain the score of the observed log-likelihood $\ln f(\mathbf{y};\boldsymbol{\phi})$ as:

$$\begin{aligned}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} &= \frac{1}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}} \int_R \frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} d\boldsymbol{\zeta} = \frac{1}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}} \int_R \frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \frac{f(\boldsymbol{\zeta};\boldsymbol{\phi})}{f(\boldsymbol{\zeta};\boldsymbol{\phi})} d\boldsymbol{\zeta} \\ &= \int_R \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \frac{f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}} d\boldsymbol{\zeta} = E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right]. \qquad \text{(A3)}\end{aligned}$$

Louis (1982) also shows that differentiating again the second term of the above chain of equalities we end up with

$$\begin{aligned}\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} &= \frac{1}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}} \int_R \frac{\partial^2 f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} d\boldsymbol{\zeta} \\ &\quad - \frac{1}{\left[\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}\right]^2} \int_R \frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} d\boldsymbol{\zeta} \frac{\partial}{\partial \boldsymbol{\phi}'}\left[\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}\right] \\ &= \frac{1}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}} \int_R \frac{\partial^2 f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} d\boldsymbol{\zeta} \\ &\quad - \frac{1}{\left[\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}\right]} \int_R \frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} d\boldsymbol{\zeta} \cdot \frac{1}{\left[\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}\right]} \int_R \frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} d\boldsymbol{\zeta} \\ &= \frac{1}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}} \int_R \frac{\partial^2 f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} \frac{f(\boldsymbol{\zeta};\boldsymbol{\phi})}{f(\boldsymbol{\zeta};\boldsymbol{\phi})} d\boldsymbol{\zeta} \\ &\quad - \frac{1}{\left[\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}\right]} \int_R \frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \frac{f(\boldsymbol{\zeta};\boldsymbol{\phi})}{f(\boldsymbol{\zeta};\boldsymbol{\phi})} d\boldsymbol{\zeta} \cdot \frac{1}{\left[\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}\right]} \int_R \frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'} \frac{f(\boldsymbol{\zeta};\boldsymbol{\phi})}{f(\boldsymbol{\zeta};\boldsymbol{\phi})} d\boldsymbol{\zeta}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}}\int_R \frac{\partial^2 f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\frac{f(\boldsymbol{\zeta};\boldsymbol{\phi})}{f(\boldsymbol{\zeta};\boldsymbol{\phi})}d\boldsymbol{\zeta} - \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'} \\
&= \int_R \left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} + \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\frac{f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\int_R f(\boldsymbol{\zeta};\boldsymbol{\phi})d\boldsymbol{\zeta}}d\boldsymbol{\zeta} \\
&\quad -\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'} \\
&= E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}\right] + E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right] \\
&\quad -E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right]E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right] \\
&= E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}\right] + V_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right] \qquad \text{(A4)}
\end{aligned}
$$

because

$$
\begin{aligned}
\frac{\partial^2 \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} &= \frac{\partial}{\partial\boldsymbol{\phi}'}\left[\frac{1}{f(\boldsymbol{\zeta};\boldsymbol{\phi})}\frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right] = \frac{1}{f(\boldsymbol{\zeta};\boldsymbol{\phi})}\frac{\partial^2 f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} - \frac{1}{f^2(\boldsymbol{\zeta};\boldsymbol{\phi})}\frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'} \\
&= \frac{1}{f(\boldsymbol{\zeta};\boldsymbol{\phi})}\frac{\partial^2 f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} - \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}.
\end{aligned}
$$

Therefore, expressions (A3) and (A4) imply (7). □

A.3 Proof or Proposition 2

First of all, note that $E_{\boldsymbol{\zeta}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})] = \mathbf{0}$ combined with the law of iterated expectations applied to second moments implies that

$$
\begin{aligned}
V_{\boldsymbol{\zeta}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})] &= E_{\boldsymbol{\zeta}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\mathbf{n}'(\boldsymbol{\zeta};\boldsymbol{\phi})] \\
= E_{\mathbf{y}}\{E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\mathbf{n}'(\boldsymbol{\zeta};\boldsymbol{\phi})]\} &= E_{\mathbf{y}}\{V_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})]\} + E_{\mathbf{y}}\{E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})]E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}'(\boldsymbol{\zeta};\boldsymbol{\phi})]\} \\
&= E_{\mathbf{y}}\{V_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})]\} + V_{\mathbf{y}}[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})],
\end{aligned}
$$

whence (8) follows.

In turn,

$$
\begin{aligned}
E_{\boldsymbol{\zeta}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right] &= E_{\mathbf{y}}\left\{E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\right\} \\
= E_{\mathbf{y}}\Bigg\{E_{\boldsymbol{\zeta}|\mathbf{y}}\Bigg[\{\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi}) - E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})]\}&\left\{\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'} - E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\right\}\Bigg] \\
&+E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})]E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}'}\right]\Bigg\} \\
= E_{\mathbf{y}}\left\{cov_{\boldsymbol{\zeta}|\mathbf{y}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi}), \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right]\right\} &+ cov_{\mathbf{y}}\left\{E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})], E_{\boldsymbol{\zeta}|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right]\right\} \\
= E_{\mathbf{y}}\left\{cov_{\boldsymbol{\zeta}|\mathbf{y}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi}), \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right]\right\} &+ cov_{\mathbf{y}}\left[\mathbf{m}(\mathbf{y};\boldsymbol{\phi}), \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\phi}}\right].
\end{aligned}
$$

But given that both $E_{\mathbf{y}}[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})]$ and $E_{\mathbf{y}}[\partial \ln f(\mathbf{y};\boldsymbol{\phi})/\partial\boldsymbol{\phi}]$ are zero, we can write this last expression as

$$E_{\boldsymbol{\zeta}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] = E_{\mathbf{y}}\left\{cov_{\boldsymbol{\zeta}|\mathbf{y}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi}), \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right]\right\} + E_{\mathbf{y}}\left[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right].$$

We also know from the generalised information matrix equality applied to the log-likelihood functions of the complete and observed data that

$$E_{\boldsymbol{\zeta}}\left[\frac{\partial \mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] + E_{\boldsymbol{\zeta}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] = \mathbf{0}$$

and

$$E_{\mathbf{y}}\left[\frac{\partial \mathbf{m}(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] + E_{\mathbf{y}}\left[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] = \mathbf{0},$$

respectively, where, once gain, Assumption 1 has allowed us to interchange integration and differentiation. Therefore, we can finally write

$$\begin{aligned} E_{\mathbf{y}}\left[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] &= -E_{\mathbf{y}}\left[\frac{\partial \mathbf{m}(\mathbf{y};\boldsymbol{\phi};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] \\ &= E_{\boldsymbol{\zeta}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] - E_{\mathbf{y}}\left\{cov_{\boldsymbol{\zeta}|\mathbf{y}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi}), \frac{\partial \ln f(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right]\right\}, \end{aligned}$$

which coincides with (9). □

A.4 Proof or Proposition 3

The proof is trivial in view of the expressions for the scores and Hessian in Appendix C.□

A.5 Proof or Proposition 4

Given that joint log-likelihood function of the complete data can be written as the sum of the marginal log-likelihood function of the multinomial random vector $\boldsymbol{\xi}$ and a linear combination with weights ξ_k of multivariate Gaussian log-likelihood functions with parameters $\boldsymbol{\nu}_k$ and $\boldsymbol{\gamma}_k$, we can exploit Proposition 1 in Amengual, Fiorentini and Sentana (2024) to express the scores of the complete log-likelihood with respect to λ_k, $\boldsymbol{\nu}_k$ and $\boldsymbol{\gamma}_k$ as linear combinations of 1, $\mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ and $\mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ scaled by ξ_k and the sum of the outer product of those scores and the corresponding Hessian as ξ_k times linear combinations of $\mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ for the $\boldsymbol{\nu}_k\boldsymbol{\nu}_k$ term, $\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ for the $\boldsymbol{\nu}_k\boldsymbol{\gamma}_k$ term, and $\mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ for the $\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k$ one. Therefore, we can avoid generalised inverses by using as influence functions the terms $E\{\xi_k|\mathbf{y}\}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ and $E\{\xi_k|\mathbf{y}\}\mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$, which we can purge from sampling uncertainty resulting from the estimation of the model parameters by regressing on $E(\xi_k|\mathbf{y})$, $E\{\xi_k|\mathbf{y}\}\mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ and $E\{\xi_k|\mathbf{y}\}\mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$, $k = 1, \ldots, K$.

As for the number of degrees of freedom, in principle they correspond to the dimensions of

$\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)$ and $\mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]$ times the number of components, namely

$$K\left[\frac{M(M+1)(M+2)}{6}+\frac{M(M+1)(M+2)(M+3)}{24}\right]=\frac{KM(M+1)(M+2)(M+7)}{24}.$$

However, if the true value of one or more of the $\lambda_k's$ is zero, then $E\{\xi_k|\mathbf{y}\}=0$ for the corresponding elements. Similarly, if two or more underlying components are such that $\boldsymbol{\theta}_k=\boldsymbol{\theta}_l$ at the true values, then $\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]=\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_l)]$ and $\mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]=\mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_l)]$. Nevertheless, in both cases the number of degrees of freedom will continue to be given by (29) as long as we interpret K as the effective number of components of the mixture. □

A.6 Proof or Lemma 2

The proof is entirely analogous to the proof of Lemma 2 in Amengual, Fiorentini and Sentana (2024), but on a component by component basis.

More formally, we have seen that the IM test statistic can be easily computed as a quadratic form in the sample means of the K vectors that contain the distinct third- and fourth-order multivariate Hermite polynomials of the observations standardised with respect to the vector of means and covariance matrix of each of the underlying components multiplied by the posterior probability of those components, with a weighting matrix which is the inverse of the residual covariance matrix in the regression of those influence functions on the K vectors that contains the distinct zero-, first-, and second-order multivariate Hermite polynomials of the same standardised variables multiplied again by the posterior probability of the components.

But the EM recursions (34a), (34b) and (34c) imply that the MLEs of the mean vectors and covariance matrices of the different components will satisfy $\mathbf{c}+\mathbf{D}\hat{\boldsymbol{\nu}}_j$ and $\mathbf{D}\hat{\boldsymbol{\Gamma}}_j\mathbf{D}'$, respectively, while the ML estimators of the mixing probabilities will not be affected. This implies in turn that the observations on $\mathbf{x}$ standardised with respect to the vector of means and covariance matrix of each of the underlying components multiplied by the posterior probability of those components will be numerically identical than the corresponding standardised values of $\mathbf{y}$, and the same will be true of their Hermite polynomials of arbitrary order, whence the result follows. □

B Multivariate Hermite polynomials

Let us follow Barndorff-Nielsen and Petersen (1979) in defining the (centred) multivariate Hermite polynomials of order $j=j_1+\ldots+j_M\geq 0$ associated to the M-dimensional random vector $\mathbf{y}$ as

$$H_{j_1 \dots j_M}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}), \boldsymbol{\Delta}] \cdot e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\nu})'\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\nu})} = (-1)^j \frac{\partial^j}{(\partial y_1)^{j_1} \dots (\partial y_M)^{j_M}} \left[e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\nu})'\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\nu})} \right], \tag{B5}$$

where $\boldsymbol{\varepsilon}(\boldsymbol{\nu}) = (\mathbf{y} - \boldsymbol{\nu})$. As is well known, the mean of any Hermite polynomial of positive degree is zero when $\mathbf{y} \sim N(\boldsymbol{\nu}, \boldsymbol{\Gamma})$, with $\boldsymbol{\Delta} = \boldsymbol{\Gamma}^{-1}$, so they constitute a basis for testing multivariate normality (see e.g. Amengual, Fiorentini and Sentana (2024) and the references therein).

The symmetry of the higher-order partial derivatives in (B5), however, implies that some of the M^j multivariate Hermite polynomials of order k will be replicated several times. Specifically, there are only $\binom{M+j-1}{k}$ different polynomials for a given order, so we can avoid generalised inverse matrices by eliminating the redundant ones. For that reason, we define

$$\mathbf{H}_j(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) = \begin{bmatrix} H_{k,0,\cdots,0}(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) \\ H_{k-1,1,\cdots,0}(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) \\ \vdots \\ H_{0,\cdots,0,k}(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) \end{bmatrix}, \tag{B6}$$

as the $\binom{M+j-1}{k} \times 1$ vector that contains all the non-redundant multivariate Hermite polynomials of order j, which we will simply denote by $\mathbf{H}_j(\boldsymbol{\varepsilon}^*)$ for the special case of $\boldsymbol{\Delta} = \mathbf{I}_M$, so that $\mathbf{H}_1(\boldsymbol{\varepsilon}^*) = \boldsymbol{\varepsilon}^*$ with $V[\mathbf{H}_1(\boldsymbol{\varepsilon}^*)] = \mathbf{I}_M$.

The usefulness of multivariate Hermite polynomials in our context results from Proposition 1 in Amengual, Fiorentini and Sentana (2024), which implies that:

1. The scores with respect to $\boldsymbol{\nu}$ and $\boldsymbol{\gamma} = vech(\boldsymbol{\Gamma})$ of the log-likelihood function associated to the multivariate random vector $\mathbf{x}$ can be written as linear combinations of $\mathbf{H}_1(\boldsymbol{\varepsilon}^*)$ and $\mathbf{H}_2(\boldsymbol{\varepsilon}^*)$, where $\boldsymbol{\varepsilon}^* = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\varepsilon}(\boldsymbol{\nu}) = \boldsymbol{\Gamma}^{-1/2}(\mathbf{y} - \boldsymbol{\nu})$.

2. The sum of the outer product of those scores and the corresponding Hessian matrix can be written as linear combinations of $\mathbf{H}_2(\boldsymbol{\varepsilon}^*)$ for the $\boldsymbol{\nu\nu}$ term, $\mathbf{H}_3(\boldsymbol{\varepsilon}^*)$ for the $\boldsymbol{\nu\gamma}$ term, and $\mathbf{H}_4(\boldsymbol{\varepsilon}^*)$ for the $\boldsymbol{\gamma\gamma}$ one.

C EM expressions for the score vector and Hessian matrix

The complete log-likelihood function of a random sample of size N on $\boldsymbol{\zeta} = (\mathbf{y}', \boldsymbol{\xi}')'$ is given by

$$\sum_{i=1}^{N} \ln f(\boldsymbol{\zeta}_i; \boldsymbol{\phi}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{ki} \ln f(\mathbf{y}_i | \xi_{ki} = 1; \boldsymbol{\theta}_k) + \sum_{i=1}^{N} \ln f(\boldsymbol{\xi}_i; \boldsymbol{\lambda}), \tag{C7}$$

where

$$\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)=-\frac{1}{2}\sum_{k=1}^{K}\left[M\ln\pi+\ln|\boldsymbol{\Gamma}_k|+\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*}(\boldsymbol{\theta}_k)\right], \tag{C8}$$

$$\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})=\sum_{k=1}^{K}\xi_k\ln\lambda_k. \tag{C9}$$

As we shall see, the sequential cut in (C7), (C8) and (C9) considerably simplifies the required expressions. Specifically,

$$\begin{aligned}
\frac{\partial\ln f(\mathbf{y},\boldsymbol{\xi};\boldsymbol{\phi})}{\partial\boldsymbol{\nu}_k} &= \xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k}=\xi_k\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^{*}(\boldsymbol{\theta}_k),\\
\frac{\partial\ln f(\mathbf{y},\boldsymbol{\xi};\boldsymbol{\phi})}{\partial\gamma_k} &= \xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\gamma_k}=-\xi_k\frac{1}{2}\mathbf{D}_M'(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^{*}(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M],\\
\frac{\partial\ln f(\mathbf{y},\boldsymbol{\xi};\boldsymbol{\phi})}{\partial\lambda_k} &= \frac{\partial\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial\lambda_k}=\xi_k\frac{1}{\lambda_k}.
\end{aligned}$$

Hence, the second derivatives will be

$$\begin{aligned}
\frac{\partial^2\ln f(\mathbf{y},\boldsymbol{\xi};\boldsymbol{\phi})}{\partial\boldsymbol{\nu}_k\partial\boldsymbol{\nu}_k'} &= \xi_k\frac{\partial^2\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k\partial\boldsymbol{\nu}_k'}=-\xi_k\boldsymbol{\Gamma}_k^{-1},\\
\frac{\partial^2\ln f(\mathbf{y},\boldsymbol{\xi};\boldsymbol{\phi})}{\partial\boldsymbol{\nu}_k\partial\boldsymbol{\gamma}_k'} &= \xi_k\frac{\partial^2\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k\partial\boldsymbol{\gamma}_k'}=-\xi_k[\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M\\
\frac{\partial^2\ln f(\mathbf{y},\boldsymbol{\xi};\boldsymbol{\phi})}{\partial\boldsymbol{\gamma}_k\partial\boldsymbol{\gamma}_k'} &= \xi_k\frac{\partial^2\ln f(\mathbf{y}_i|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_k\partial\boldsymbol{\gamma}_k'}\\
&= -\xi_k\frac{1}{2}\mathbf{D}_M'\{2[(\boldsymbol{\Gamma}_k^{-1}\otimes\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^{*}(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}]-(\boldsymbol{\Gamma}_k^{-1}\otimes\boldsymbol{\Gamma}_k^{-1})\}\mathbf{D}_M,
\end{aligned}$$

and

$$\frac{\partial^2\ln f(\mathbf{y},\boldsymbol{\xi};\boldsymbol{\phi})}{(\partial\lambda_k)^2}=\frac{\partial^2\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{(\partial\lambda_k)^2}=-\xi_k\frac{1}{\lambda_k^2},$$

with all other cross-derivatives being zero.

The assumption of random sampling implies that the joint distribution of $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_i,\ldots,\boldsymbol{\xi}_N$ given $y_1,\ldots,y_i,\ldots,y_N$ is the product of the N distributions of $\boldsymbol{\xi}_i$ given $\mathbf{y}_i$, which are also categorical but with probabilities $w_{ki}(\boldsymbol{\phi})$ given by (26). On this basis, we can use expression (A3) to write

$$\begin{aligned}
\frac{\partial\ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\nu}_k} &= E\left[\xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k}\right|\mathbf{y}\Bigg]=w_k(\boldsymbol{\phi})\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^{*}(\boldsymbol{\theta}_k),\\
\frac{\partial\ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\gamma_k} &= E\left[\xi_k\frac{\partial\ln f(\mathbf{y}_i|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\gamma_k}\right|\mathbf{y}\Bigg]=-w_k(\boldsymbol{\phi})\frac{1}{2}\mathbf{D}_M'(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})\\
&\quad\times vec[\boldsymbol{\varepsilon}^{*}(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M],\\
\frac{\partial\ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\lambda_k} &= E\left[\frac{\partial\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial\lambda_k}\right|\mathbf{y}\Bigg]=w_k(\boldsymbol{\phi})\frac{1}{\lambda_k}.
\end{aligned}$$

Similarly, the only non-zero elements of the first term in (A4) will be

$$E\left[\xi_k \frac{\partial^2 \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k \partial\boldsymbol{\nu}'_k}\middle|\mathbf{y}\right] = -w_k(\boldsymbol{\phi})\boldsymbol{\Gamma}_k^{-1},$$

$$E\left[\xi_k \frac{\partial^2 \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k \partial\boldsymbol{\gamma}'_k}\middle|\mathbf{y}\right] = -w_k(\boldsymbol{\phi})[\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M,$$

$$E\left[\xi_k \frac{\partial^2 \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_k \partial\boldsymbol{\gamma}'_k}\middle|\mathbf{y}\right]$$
$$= -w_k(\boldsymbol{\phi})\frac{1}{2}\mathbf{D}'_M\{2[(\boldsymbol{\Gamma}_k^{-1} \otimes \boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}] - (\boldsymbol{\Gamma}_k^{-1} \otimes \boldsymbol{\Gamma}_k^{-1})\}\mathbf{D}_M,$$

$$E\left[\frac{\partial^2 \ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{(\partial\lambda_k)^2}\middle|\mathbf{y}\right] = -\frac{1}{\lambda_k^2}w_k(\boldsymbol{\phi}).$$

In contrast, the second term of (A4) is slightly more complex. Specifically, we get

$$V\left[\xi_k \frac{\partial \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k}\middle|\mathbf{y}\right] = w_k(\boldsymbol{\phi})[1-w_k(\boldsymbol{\phi})]\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}$$
$$= w_k(\boldsymbol{\phi})\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2} - w_k^2(\boldsymbol{\phi})\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2},$$

where we have used the fact that ξ_k is a Bernoulli random variable whose variance conditional on $\mathbf{y}$ is $w_k(\boldsymbol{\phi})[1-w_k(\boldsymbol{\phi})]$. In turn,

$$Cov\left[\xi_k \frac{\partial \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k}, \xi_k \frac{\partial \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_k}\middle|\mathbf{y}\right]$$
$$= w_k(\boldsymbol{\phi})[1-w_k(\boldsymbol{\phi})]\frac{1}{2}\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M$$
$$= w_k(\boldsymbol{\phi})\frac{1}{2}\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M$$
$$-w_k^2(\boldsymbol{\phi})\frac{1}{2}\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M,$$

$$Cov\left[\xi_k \frac{\partial \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k}, \frac{\partial \ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial\lambda_k}\middle|\mathbf{y}\right] = w_k(\boldsymbol{\phi})[1-w_k(\boldsymbol{\phi})]\frac{1}{\lambda_k}\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)$$
$$= w_{ki}(\boldsymbol{\phi})\frac{1}{\lambda_k}\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) - w_k^2(\boldsymbol{\phi})\frac{1}{\lambda_k}\boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k),$$

$$V\left[\xi_k \frac{\partial \ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_k}\middle|\mathbf{y}\right]$$
$$= w_k(\boldsymbol{\phi})[1-w_k(\boldsymbol{\phi})]\frac{1}{4}\mathbf{D}'_M(\boldsymbol{\Gamma}_k^{\prime -1/2} \otimes \boldsymbol{\Gamma}_k^{\prime -1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]$$
$$\times vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M$$
$$= w_k(\boldsymbol{\phi})\frac{1}{4}\mathbf{D}'_N(\boldsymbol{\Gamma}_k^{\prime -1/2} \otimes \boldsymbol{\Gamma}_k^{\prime -1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]$$
$$\times vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M$$
$$-w_k^2(\boldsymbol{\phi})\frac{1}{4}\mathbf{D}'_M(\boldsymbol{\Gamma}_k^{\prime -1/2} \otimes \boldsymbol{\Gamma}_k^{\prime -1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]$$
$$\times vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M,$$

$$
\begin{aligned}
&Cov\left[\xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_k},\frac{\partial\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial\lambda_k}\right|\mathbf{y}\right]\\
&=w_k(\boldsymbol{\phi})[1-w_k(\boldsymbol{\phi})]\frac{1}{2\lambda_k}\mathbf{D}'_M(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M]\\
&=w_k(\boldsymbol{\phi})\frac{1}{2\lambda_k}\mathbf{D}'_M(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M]\\
&-w_k^2(\boldsymbol{\phi})\frac{1}{2\lambda_k}\mathbf{D}'_M(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M]
\end{aligned}
$$

and

$$
\begin{aligned}
Cov\left[\frac{\partial\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial\lambda_k}\right|\mathbf{y}\right]&=w_k(\boldsymbol{\phi})[1-w_k(\boldsymbol{\phi})]\frac{1}{\lambda_k^2}\\
&=w_k(\boldsymbol{\phi})\frac{1}{\lambda_k^2}-w_k^2(\boldsymbol{\phi})\frac{1}{\lambda_k^2}.
\end{aligned}
$$

Interestingly, the second terms in the previous expressions are nothing other than the minus products of the corresponding scores.

In addition, we must also compute all the other conditional covariances between the different components of the score. Specifically,

$$
\begin{aligned}
&Cov\left[\xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k},\xi_l\frac{\partial\ln f(\mathbf{y}|\xi_l\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_l}\right|\mathbf{y}\right]\\
=\;\;&-w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})\frac{1}{\gamma_k^2}\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_l)\boldsymbol{\Gamma}_l^{-1/2},
\end{aligned}
$$

where we have used the fact that ξ_k and ξ_l are elements of a multinomial random vector whose covariance conditional on $\mathbf{y}$ is $-w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})$. Similarly,

$$
\begin{aligned}
&Cov\left[\xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k},\xi_l\frac{\partial\ln f(\mathbf{y}|\xi_l=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_l}\right|\mathbf{y}\right]\\
&=-w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})\frac{1}{2}\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\\
&\times vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_l)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_l)-\mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M,\\
&Cov\left[\xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\nu}_k},\frac{\partial\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial\lambda_l}\right|\mathbf{y}\right]=-w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})\frac{1}{\lambda_l}\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k),\\
&Cov\left[\xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_k},\xi_l\frac{\partial\ln f(\mathbf{y}|\xi_l\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_l}\right|\mathbf{y}\right]\\
&=-w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})\frac{1}{4}\mathbf{D}'_M(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M]\\
&\times vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_l)\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_l)-\mathbf{I}_M](\boldsymbol{\Gamma}_l^{-1/2}\otimes\boldsymbol{\Gamma}_l^{-1/2})\mathbf{D}_M,\\
&Cov\left[\xi_k\frac{\partial\ln f(\mathbf{y}|\xi_k\!=\!1;\boldsymbol{\theta}_k)}{\partial\boldsymbol{\gamma}_k},\frac{\partial\ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial\lambda_l}\right|\mathbf{y}\right]\\
&=-w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})\frac{1}{2\lambda_l}\mathbf{D}'_M(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M]
\end{aligned}
$$

and

$$cov_{\zeta|\mathbf{y}}\left[\frac{\partial \ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial \lambda_k}, \frac{\partial \ln f(\boldsymbol{\xi};\boldsymbol{\lambda})}{\partial \lambda_l}\right] = -w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})\frac{1}{\lambda_k \lambda_l},$$

which also coincide with the outer products of the scores involved. Thus, to compute the Hessian we simply need to add to the minus OPS the terms that appear in Proposition 3.

D Scores and Hessian expressions in Boldea and Magnus (2009)

Theorem 1 in Boldea and Magnus (2009) provides analytical expressions for the contribution of a single observation on $\mathbf{y}$ to score and Hessian matrix. As we mentioned before, they reparametrise $\boldsymbol{\lambda}$ so that $\lambda_k = \pi_k$ for $k = 1, \ldots, K-1$, and $\lambda_K = 1 - \sum_{k=1}^{K-1} \pi_k$. Then, they introduce some additional notation. First,

$$\mathbf{a}_k = \pi_k^{-1}\mathbf{e}_k \quad k = 1, \ldots, K-1; \quad \mathbf{a}_K = -(1 - \textstyle\sum_{k=1}^{K-1} \pi_k)^{-1}\boldsymbol{\iota}_{K-1},$$

where $\mathbf{e}_k$ is the k^{th} column of $\mathbf{I}_{K-1}$ and $\boldsymbol{\iota}_{K-1}$ a vector of $K-1$ ones. Next, they define

$$\begin{aligned}
\mathbf{b}_k &= \boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k); \\
\mathbf{B}_k &= -\boldsymbol{\Gamma}_k^{-1/2\prime}[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]\boldsymbol{\Gamma}_k^{-1/2}, \\
\mathbf{c}_k &= \begin{bmatrix} \boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \\ \frac{1}{2}\mathbf{D}_M' vec\{\boldsymbol{\Gamma}_k^{-1/2\prime}[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]\boldsymbol{\Gamma}_k^{-1/2}\} \end{bmatrix}
\end{aligned}$$

and

$$\mathbf{C}_k = \left\{ \begin{array}{c} \boldsymbol{\Gamma}_k^{-1} \\ \mathbf{D}_M'[\boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \otimes \boldsymbol{\Gamma}_k^{-1}] \end{array} \right.$$

$$\left. \begin{array}{c} [\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M \\ \frac{1}{2}\mathbf{D}_M'[\{\boldsymbol{\Gamma}_k^{-1} + 2[\boldsymbol{\Gamma}_k^{-1/2\prime}[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]\boldsymbol{\Gamma}_k^{-1/2}\} \otimes \boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M \end{array} \right\}$$

for $k = 1, \ldots, K$.

In this notation, Theorem 1 in Boldea and Magnus (2009) states that the contribution to the scores of a single observation are given by

$$\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\pi}} = \sum_{k=1}^{K-1} \frac{w_k(\boldsymbol{\phi})}{\pi_k}\mathbf{e}_k - \frac{w_K(\boldsymbol{\phi})}{1 - \sum_{k=1}^{K-1} \pi_k}\boldsymbol{\iota}_{K-1} \tag{D10}$$

and

$$\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\theta}_k} = w_k(\boldsymbol{\phi}) \left\{ \begin{array}{c} \boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \\ \frac{1}{2}\mathbf{D}_M'(\boldsymbol{\Gamma}_k^{-1/2\prime} \otimes \boldsymbol{\Gamma}_k^{-1/2\prime})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M] \end{array} \right\}. \tag{D11}$$

In addition, the same theorem also says that its contribution to the Hessian will be given by the following blocks:

$$
\begin{aligned}
\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\pi}\partial \boldsymbol{\pi}'} &= -\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\pi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\pi}'}, \\
\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\theta}_k\partial \boldsymbol{\pi}'} &= w_k(\boldsymbol{\phi})\left\{\begin{array}{c} \boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \\ \frac{1}{2}\mathbf{D}_M'(\boldsymbol{\Gamma}_k^{-1/2\prime}\otimes\boldsymbol{\Gamma}_k^{-1/2\prime})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M] \end{array}\right\} \\
&\quad\times\left[\frac{1}{\pi_k}\mathbf{e}_k-\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\pi}}\right]', \\
\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\theta}_K\partial \boldsymbol{\pi}'} &= -w_K(\boldsymbol{\phi})\left[\begin{array}{c} \boldsymbol{\Gamma}_K^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_K) \\ \frac{1}{2}\mathbf{D}_M' vec\{\boldsymbol{\Gamma}_K^{-1/2\prime}[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_K)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_K)-\mathbf{I}_M]\boldsymbol{\Gamma}_K^{-1/2}\} \end{array}\right] \\
&\quad\times\left[\frac{1}{1-\sum_{k=1}^{K-1}\pi_k}\boldsymbol{\iota}_{K-1}+\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\pi}}\right]',
\end{aligned}
$$

$$
\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\theta}_k\partial \boldsymbol{\theta}_k'} = -w_k(\boldsymbol{\phi})\left[\begin{array}{c} \mathbf{C}_k \\ -[1-w_k(\boldsymbol{\phi})]\left\{\begin{array}{c} \boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \\ \frac{1}{2}\mathbf{D}_M'(\boldsymbol{\Gamma}_k^{-1/2\prime}\otimes\boldsymbol{\Gamma}_k^{-1/2\prime})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M] \end{array}\right\} \\ \times\left\{\begin{array}{cc} [\boldsymbol{\varepsilon}'(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2} & \frac{1}{2}vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M \end{array}\right\} \end{array}\right]
$$

and

$$
\begin{aligned}
\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\theta}_k\partial \boldsymbol{\theta}_l'} &= -w_k(\boldsymbol{\phi})w_l(\boldsymbol{\phi})\left\{\begin{array}{c} \boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \\ \frac{1}{2}\mathbf{D}_M'(\boldsymbol{\Gamma}_k^{-1/2\prime}\otimes\boldsymbol{\Gamma}_k^{-1/2\prime})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M] \end{array}\right\} \\
&\quad\times\left\{\begin{array}{cc} [\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_l)\boldsymbol{\Gamma}_l^{-1/2} & \frac{1}{2}vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_l)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_l)-\mathbf{I}_M](\boldsymbol{\Gamma}_l^{-1/2}\otimes\boldsymbol{\Gamma}_l^{-1/2})\mathbf{D}_M \end{array}\right\}
\end{aligned}
$$

for $k \neq l$.

These expressions differ from the ones we have obtained in the previous sections because Boldea and Magnus (2009) work with $\boldsymbol{\pi}$ rather than $\boldsymbol{\lambda}$.

Nevertheless, given that

$$
\boldsymbol{\lambda}=\left(\begin{array}{c} \lambda_1 \\ \vdots \\ \lambda_{K-1} \\ \lambda_K \end{array}\right)=\left(\begin{array}{c} 0 \\ \vdots \\ 0 \\ 1 \end{array}\right)+\left(\begin{array}{ccc} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ -1 & \dots & -1 \end{array}\right)\left(\begin{array}{c} \pi_1 \\ \vdots \\ \pi_{K-1} \end{array}\right)=\mathbf{e}_K+\left(\begin{array}{c} \mathbf{I}_{K-1} \\ -\boldsymbol{\iota}_{K-1}' \end{array}\right)\boldsymbol{\pi}, \quad \text{(D12)}
$$

so that

$$
\frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\pi}'}=\left(\begin{array}{c} \mathbf{I}_{K-1} \\ -\boldsymbol{\iota}_{K-1}' \end{array}\right),
$$

it is easy to see that

$$
\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\pi}}=\frac{\partial \boldsymbol{\lambda}'}{\partial \boldsymbol{\pi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\lambda}}=\left(\begin{array}{cc} \mathbf{I}_{K-1} & -\boldsymbol{\iota}_{K-1}' \end{array}\right)\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\lambda}}
$$

coincides with (D10).

Similarly, given that (D12) is affine, so that its second Jacobian is 0, it follows that

$$\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\pi}\partial\boldsymbol{\pi}'} = \left(\begin{array}{cc} \mathbf{I}_{K-1} & -\boldsymbol{\iota}'_{K-1} \end{array} \right) \frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'} \left(\begin{array}{c} \mathbf{I}_{K-1} \\ -\boldsymbol{\iota}'_{K-1} \end{array} \right).$$

It is tedious but straightforward to show that analogous calculations applied to the other terms we have derived in Appendix C yield the results in Theorem 1 in Boldea and Magnus (2009).

Nevertheless, the advantage of deriving the scores and Hessian matrices in terms of $\boldsymbol{\lambda}$ is that they are also useful for alternative reparametrisations of those probabilities. For example, in the multivariate logit case in (18), the Jacobian would be instead

$$\begin{aligned}
\frac{\partial\lambda_k}{\partial\pi_k} &= \frac{e^{\pi_k}}{\sum_{l=1}^{K-1} e^{\pi_l}+1}\left(1-\frac{e^{\pi_k}}{\sum_{k=1}^{K-1} e^{\pi_k}+1}\right) = \lambda_k(1-\lambda_k) \text{ for } k=1,\ldots,K-1, \\
\frac{\partial\lambda_K}{\partial\pi_k} &= -\frac{e^{\pi_k}}{\sum_{l=1}^{K-1} e^{\pi_l}+1}\frac{1}{\sum_{k=1}^{K-1} e^{\pi_k}+1} = -\lambda_k\lambda_K \text{ for } k=1,\ldots,K-1, \\
\frac{\partial\lambda_k}{\partial\pi_l} &= -\frac{e^{\pi_k}}{\sum_{l=1}^{K-1} e^{\pi_l}+1}\frac{e^{\pi_l}}{\sum_{k=1}^{K-1} e^{\pi_k}+1} = -\lambda_k\lambda_l \text{ for } l\neq k,\ k=1,\ldots,K-1.
\end{aligned}$$

E Standardised multivariate discrete mixtures of normals

Consider the following mixture of two multivariate normals

$$\mathbf{y} \sim \left\{ \begin{array}{lll} N(\boldsymbol{\nu}_1,\boldsymbol{\Gamma}_1) & \text{with probability} & \pi, \\ N(\boldsymbol{\nu}_2,\boldsymbol{\Gamma}_2) & \text{with probability} & 1-\pi. \end{array} \right. \tag{E13}$$

Given (16) and (17), this random vector will be standardised if and only if

$$\pi\boldsymbol{\nu}_1 + (1-\pi)\boldsymbol{\nu}_2 = \mathbf{0}$$

and

$$\pi(1-\pi)(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)(\boldsymbol{\nu}_1-\boldsymbol{\nu}_2)' + \pi\boldsymbol{\Gamma}_1 + (1-\pi)\boldsymbol{\Gamma}_2 = \mathbf{I}_M,$$

in which case we will denote it by $\boldsymbol{\varepsilon}^*$.

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

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

Thus, it is not difficult to see that by choosing

$$\boldsymbol{\Gamma}_{1L} = [\pi\mathbf{I}_M + (1-\pi)\boldsymbol{\aleph}_L\boldsymbol{\aleph}'_L]_U^{-1\prime} \text{ and } \boldsymbol{\Gamma}_{2L} = \boldsymbol{\Gamma}_{1L}\boldsymbol{\aleph}_L, \tag{E14}$$

where $\aleph_L$ is a lower triangular matrix and $[\pi\mathbf{I}_M + (1-\pi)\aleph_L\aleph_L']_U[\pi\mathbf{I}_M + (1-\pi)\aleph_L\aleph_L']_U'$ is the upper triangular Cholesky decomposition of $[\pi\mathbf{I}_M + (1-\pi)\aleph_L\aleph_L']$, we can indeed obtain a standardised vector $\boldsymbol{\varepsilon}^*$ because of the relationship between the upper Cholesky decomposition of a matrix and the lower Cholesky decomposition of its inverse.

Now consider the case $\boldsymbol{\delta}\neq \mathbf{0}$, and let

$$\boldsymbol{\Upsilon} = \pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}' + \mathbf{I}_M.$$

Then, it is easy to see that if we call $\boldsymbol{\Upsilon}_U\boldsymbol{\Upsilon}_U'$ the upper triangular Cholesky decomposition of $\boldsymbol{\Upsilon}$, then

$$\boldsymbol{\nu}_1^* = \boldsymbol{\Upsilon}_U^{-1\prime}(1-\pi)\boldsymbol{\delta},\ \ \boldsymbol{\nu}_2^* = -\boldsymbol{\Upsilon}_U^{-1\prime}\pi\boldsymbol{\delta},\ \ \boldsymbol{\Gamma}_1^* = \boldsymbol{\Upsilon}_U^{-1\prime}\boldsymbol{\Gamma}_1\boldsymbol{\Upsilon}_U^{-1},\ \text{ and }\ \boldsymbol{\Gamma}_2^* = \boldsymbol{\Upsilon}_U^{-1\prime}\boldsymbol{\Gamma}_2\boldsymbol{\Upsilon}_U^{-1},$$

with $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$ as in (E14), continue to generate another standardised vector.

In summary, we can generate a standardised, multivariate, two-component Gaussian mixture as

$$\boldsymbol{\varepsilon}^* = \boldsymbol{\Upsilon}_U^{-1\prime}\left\{(\xi-\pi)\boldsymbol{\delta} + [\boldsymbol{\Gamma}_{2L} + \xi(\boldsymbol{\Gamma}_{1L} - \boldsymbol{\Gamma}_{2L})]\boldsymbol{\varepsilon}\right\},$$

where ξ denotes a Bernoulli variable which takes the value 1 with probability π and 0 with probability $1-\pi$, and $\boldsymbol{\varepsilon}|\xi \sim N(\mathbf{0}, \mathbf{I}_2)$. The intuition is as follows. First, note that $(\xi-\pi)\boldsymbol{\delta}$ is a vector version of a shifted and scaled Bernoulli random variable with 0 mean and rank 1 covariance matrix $\pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}'$. But since

$$[\boldsymbol{\Gamma}_{2L} + \xi(\boldsymbol{\Gamma}_{1L} - \boldsymbol{\Gamma}_{2L})]\boldsymbol{\varepsilon},$$

with $\boldsymbol{\Gamma}_{1L}$ and $\boldsymbol{\Gamma}_{2L}$ given by (E14), is a multivariate discrete scale mixture of normals with 0 unconditional mean and unit unconditional covariance matrix that is orthogonal to $(\xi-\pi)\boldsymbol{\delta}$ because of the independence between ξ and $\boldsymbol{\varepsilon}$, the sum of the two random variables will have variance $\mathbf{I}_M + \pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}'$, which explains the $\boldsymbol{\Upsilon}^{-\frac{1}{2}}$ in front of the curly brackets.

Consequently, we can think of an alternative parametrisation with two sets of parameters: the ones that capture the first two unconditional moments of the distribution, namely $\boldsymbol{\tau}$ and $vech(\boldsymbol{\Psi})$, and the ones that characterise the shape of the standardised distribution, which are given by $\boldsymbol{\eta} = (\boldsymbol{\delta}', vech'(\aleph_L), \pi)'$.

Therefore, two equivalent ways of defining and simulating $\mathbf{y}$ with mean $\boldsymbol{\tau}$ and variance $\boldsymbol{\Psi}$ are as follows. First, we can consider

$$\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\Psi}_L\boldsymbol{\varepsilon}^*, \text{ where } \boldsymbol{\varepsilon}^* = \begin{cases} N[\boldsymbol{\nu}_1^*(\boldsymbol{\eta}), \boldsymbol{\Gamma}_1^*(\boldsymbol{\eta})] \text{ with probability } \pi \\ N[\boldsymbol{\nu}_2^*(\boldsymbol{\eta}), \boldsymbol{\Gamma}_2^*(\boldsymbol{\eta})] \text{ with probability } 1-\pi \end{cases}, \tag{E15}$$

where $\boldsymbol{\Psi}_L\boldsymbol{\Psi}_L'$ denotes the lower triangular Cholesky decomposition of $\boldsymbol{\Psi}$,

$$\begin{aligned}
\boldsymbol{\nu}_1^*(\boldsymbol{\eta}) &= [\pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{'} + \mathbf{I}_M]_U^{-1\prime}\boldsymbol{\delta}(1-\pi) \\
\boldsymbol{\nu}_2^*(\boldsymbol{\eta}) &= -[\pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{'} + \mathbf{I}_M]_U^{-1\prime}\boldsymbol{\delta}\pi
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\Gamma}_{1L}^*(\boldsymbol{\eta}) &= [\pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{'} + \mathbf{I}_M]_U^{-1\prime}[\pi\mathbf{I}_M + (1-\pi)\boldsymbol{\aleph}_L\boldsymbol{\aleph}_L']_U^{-1\prime} \\
\boldsymbol{\Gamma}_{2L}^*(\boldsymbol{\eta}) &= [\pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{'} + \mathbf{I}_M]_U^{-1\prime}[\pi\mathbf{I}_M + (1-\pi)\boldsymbol{\aleph}_L\boldsymbol{\aleph}_L']_U^{-1\prime}\boldsymbol{\aleph}_L
\end{aligned}$$

Alternatively, we can use

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

where

$$\boldsymbol{\nu}_k = \boldsymbol{\tau} + \boldsymbol{\Psi}_L\boldsymbol{\nu}_k^*(\boldsymbol{\eta})$$

and

$$\boldsymbol{\Gamma}_{kL} = \boldsymbol{\Psi}_L\boldsymbol{\Gamma}_{kL}^*(\boldsymbol{\eta})$$

for $k = 1, 2$.

To illustrate the procedure in the bivariate case, let

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

so that the vector of shape parameters of $\boldsymbol{\varepsilon}^*$ becomes $\boldsymbol{\eta} = (\delta_1, \delta_2, \varkappa_{11}, \varkappa_{21}, \varkappa_{22}, \pi)'$.

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

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

and

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

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

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

and

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

As for the the lower triangular decompositions of the covariance matrices of the two components, namely

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

we will have

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

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

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

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

and

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

Similar calculations can be applied for general M, although the number of free parameters of $\mathbf{\Psi}_L$ and $\aleph_L$ increase with the square of the cross-sectional dimension. Extensions to mixtures with $K > 2$ components are also feasible by recursively applying the above procedures to the mixture of a spherical Gaussian random vector and a standardised Gaussian mixture with $K-1$ components.

It is of some interest to obtain the scores with respect to $\boldsymbol{\tau}$, $vech(\mathbf{\Psi})$ and $\boldsymbol{\eta}$ from the scores with respect to $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $vech(\mathbf{\Gamma}_1)$, $vech(\mathbf{\Gamma}_2)$ and π. The delta method immediately implies that the former can be written as a linear combination of the latter, whose expressions we have derived in Appendix C.

First of all, note that the fact that $w_2 = 1 - w_1$ for any parameter configuration and data implies that

$$\lambda_1\frac{w_1}{\lambda_1} + \lambda_2\frac{w_2}{\lambda_2} = 1,$$

so there is a linear combination of the scores with respect to the λ's which is identically equal to 1. Note also that the sample average of w_k/λ_k evaluated at the MLE of the model parameters will be identically equal to 1 rather than 0 for all k.

Let us know try to find the score with respect to $\boldsymbol{\tau}$. We know from (16) that

$$\frac{\partial \boldsymbol{\tau}}{\partial \boldsymbol{\nu}'_k} = \lambda_k \mathbf{I}_M,\ \frac{\partial \boldsymbol{\tau}}{\partial \gamma'_k} = \mathbf{0} \text{ and } \frac{\partial \boldsymbol{\tau}}{\partial \lambda_k} = \boldsymbol{\nu}_k.$$

We also know that

$$\begin{aligned} \boldsymbol{\nu}_1 &= \boldsymbol{\tau} + \boldsymbol{\Psi}_L[\pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{'} + \mathbf{I}_M]_U^{-1'}\boldsymbol{\delta}(1-\pi) \\ \boldsymbol{\nu}_2 &= \boldsymbol{\tau} - \boldsymbol{\Psi}_L[\pi(1-\pi)\boldsymbol{\delta}\boldsymbol{\delta}^{'} + \mathbf{I}_M]_U^{-1'}\boldsymbol{\delta}\pi \end{aligned}$$

which means that

$$\frac{\partial \boldsymbol{\nu}_k}{\partial \boldsymbol{\tau}} = \mathbf{I}_M$$

Hence, given that no other parameter of the natural parametrisation depends on $\boldsymbol{\tau}$, the delta method immediately implies that the score with respect to $\boldsymbol{\tau}$ will be given by

$$\frac{\partial l(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\tau}} = \sum_{k=1}^{K} \frac{\partial \boldsymbol{\nu}'_k}{\partial \boldsymbol{\tau}} \frac{\partial l(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\nu}_k} \sum_{k=1}^{K} \lambda_k w_{ki}(\boldsymbol{\phi}) \boldsymbol{\Gamma}_k^{\prime -1/2} \boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) = -\frac{\partial l(\mathbf{y};\boldsymbol{\phi})}{\partial \mathbf{y}}.$$

Similarly, we would expect

$$\frac{\partial l(\mathbf{y};\boldsymbol{\phi})}{\partial vech(\boldsymbol{\Psi})} = vech\left[\mathbf{I}_M - \frac{\partial l(\mathbf{y};\boldsymbol{\phi})}{\partial \mathbf{y}}(\mathbf{y}-\boldsymbol{\tau})'\boldsymbol{\Psi}_L^{-1/2}\right].$$

But we know that the score with respect to $\boldsymbol{\tau}$ evaluated at the sample mean is 0.

Table 1: Finite sample properties of the IM test. Null hypothesis: Mixture of two univariate normals

Panel A: Size properties (asymptotic)

	Test version					
	OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%
100	80.59	76.37	68.60	5.21	2.86	0.95
400	47.35	40.84	30.16	8.55	4.99	1.86
1,600	24.33	17.89	9.86	9.40	5.13	1.60

Panel B: Size properties (bootstrap)

	Test version					
	OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%
100	10.41	4.92	0.92	11.46	6.15	1.31
400	7.20	3.08	0.56	10.65	5.51	1.17
1,600	9.69	4.89	1.01	9.72	4.89	1.04

Panel C: Power properties of the IM test (bootstrap)

	Sample size					
	100			400		
DGP	10%	5%	1%	10%	5%	1%
Non-Gaussian mixture	46.00	36.84	19.24	94.60	90.52	71.08
Mixture of three normals	12.96	5.88	0.80	42.28	23.68	3.88
Lognormal	99.40	97.72	79.88	100.00	100.00	99.84

Notes: Monte Carlo empirical rejection rates based on 10,000 (2,500) replications in Panels A and B (Panel C). OPS refers to the version of the statistic proposed by Chesher (1983) and Lancaster (1984) and employed by Boldea and Magnus (2009), while IM to the feasible version that makes use of the theoretical expression (33) replacing the true parameter values $\boldsymbol{\phi}_0$ by their MLEs $\hat{\boldsymbol{\phi}}_T$. Panel A contains rejection rates based on the asymptotic critical values (see Proposition 4.3) while those in Panels B and C are based on a parametric bootstrap procedure in which we simulate $B = 99$ samples from the mixture model estimated under the null. See section 4 for details about the DGPs.

Table 2: Finite sample properties of the IM test. Null hypothesis: Mixture of three univariate normals

Panel A: Size properties (asymptotic)

	Test version					
	OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%
100	80.81	76.34	68.32	2.66	1.23	0.38
400	53.15	46.82	36.72	6.79	3.79	1.37
1,600	27.12	19.64	10.49	9.74	5.46	2.04

Panel B: Size properties (bootstrap)

	Test version					
	OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%
100	10.74	4.74	0.58	9.21	4.23	0.76
400	9.01	4.34	0.97	10.02	4.89	0.97
1,600	8.51	3.59	0.55	10.35	5.17	1.15

Panel C: Power properties of the IM test (bootstrap)

	Sample size					
	100			400		
DGP	10%	5%	1%	10%	5%	1%
Non-Gaussian mixture	49.52	37.64	16.40	96.48	93.12	70.72
Mixture of four normals	36.72	25.20	8.88	97.64	93.28	60.52
Lognormal	69.20	54.00	22.76	99.76	99.52	94.68

Notes: Monte Carlo empirical rejection rates based on 10,000 (2,500) replications in Panels A and B (Panel C). OPS refers to the version of the statistic proposed by Chesher (1983) and Lancaster (1984) and employed by Boldea and Magnus (2009), while IM to the feasible version that makes use of the theoretical expression (33) replacing the true parameter values $\boldsymbol{\phi}_0$ by their MLEs $\hat{\boldsymbol{\phi}}_T$. Panel A contains rejection rates based on the asymptotic critical values (see Proposition 4.3) while those in Panels B and C are based on a parametric bootstrap procedure in which we simulate $B = 99$ samples from the mixture model estimated under the null. See section 4 for details about the DGPs.

Table 3: Finite sample properties of the IM test. Null hypothesis: Mixture of two bivariate normals

Panel A: Size properties (asymptotic)

	Test version					
	OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%
100	99.66	99.46	98.26	8.03	5.53	3.06
400	80.56	74.50	60.67	10.41	6.76	3.01
1,600	42.72	33.70	18.83	9.96	5.64	1.73

Panel B: Size properties (bootstrap)

	Test version					
	OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%
100	8.44	4.11	0.67	10.39	5.04	0.91
400	9.71	4.66	0.87	9.69	4.96	1.10
1,600	9.84	5.09	1.11	9.52	4.70	0.77

Panel C: Power properties of the IM test (bootstrap)

	Sample size					
	100			400		
DGP	10%	5%	1%	10%	5%	1%
Non-Gaussian mixture	57.96	47.32	24.92	94.12	88.92	62.72
Mixture of three normals	85.00	68.12	23.16	96.84	96.28	83.40
Skew normal	42.16	27.48	8.76	97.12	91.36	63.32

Notes: Monte Carlo empirical rejection rates based on 10,000 (2,500) replications in Panels A and B (Panel C). OPS refers to the version of the statistic proposed by Chesher (1983) and Lancaster (1984) and employed by Boldea and Magnus (2009), while IM to the feasible version that makes use of the theoretical expression (33) replacing the true parameter values $\boldsymbol{\phi}_0$ by their MLEs $\hat{\boldsymbol{\phi}}_T$. Panel A contains rejection rates based on the asymptotic critical values (see Proposition 4.3) while those in Panels B and C are based on a parametric bootstrap procedure in which we simulate $B = 99$ samples from the mixture model estimated under the null. See section 4 for details about the DGPs.

Table 4: Specification testing for convergence clubs in cross-country GDP per capita

	Panel A: Parameter estimates									Panel B: IM test (p-values)	
Sample	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	λ_1	λ_2	λ_3	Asymptotic	Bootstrap
1960	2.74	0.95	0.31	1.14	0.36	0.12	0.29	0.39	0.32	0.68	0.44
1965	2.84	1.01	0.28	1.05	0.39	0.11	0.27	0.40	0.33	0.63	0.37
1970	2.74	0.96	0.27	0.96	0.40	0.10	0.31	0.37	0.33	0.34	0.13
1975	3.08	1.07	0.26	0.65	0.47	0.10	0.24	0.45	0.31	0.47	0.24
1980	2.87	1.08	0.26	0.68	0.40	0.12	0.28	0.38	0.34	0.74	0.49
1985	2.86	0.92	0.20	0.69	0.43	0.07	0.27	0.49	0.24	0.40	0.20
1990	3.12	0.93	0.18	0.56	0.48	0.05	0.24	0.52	0.24	0.55	0.39
1995	3.02	0.89	0.15	0.49	0.48	0.05	0.25	0.50	0.25	0.70	0.56
2000	2.93	0.82	0.15	0.59	0.44	0.05	0.28	0.48	0.24	0.51	0.35

Notes: Data: Per capita income from version 6.1 of the Penn World Tables. IM test refers to the feasible version that makes use of the theoretical expression (33) replacing the true parameter values $\boldsymbol{\phi}_0$ by their MLEs $\hat{\boldsymbol{\phi}}_T$. The first column of Panel B contains p-values based on the asymptotic critical values (see Proposition 4.3) while the second one those based on a parametric bootstrap procedure in which we simulate $B = 9,999$ samples from the mixture model estimated under the null.

Figure 1: Univariate distributions under null hypotheses and different alternatives

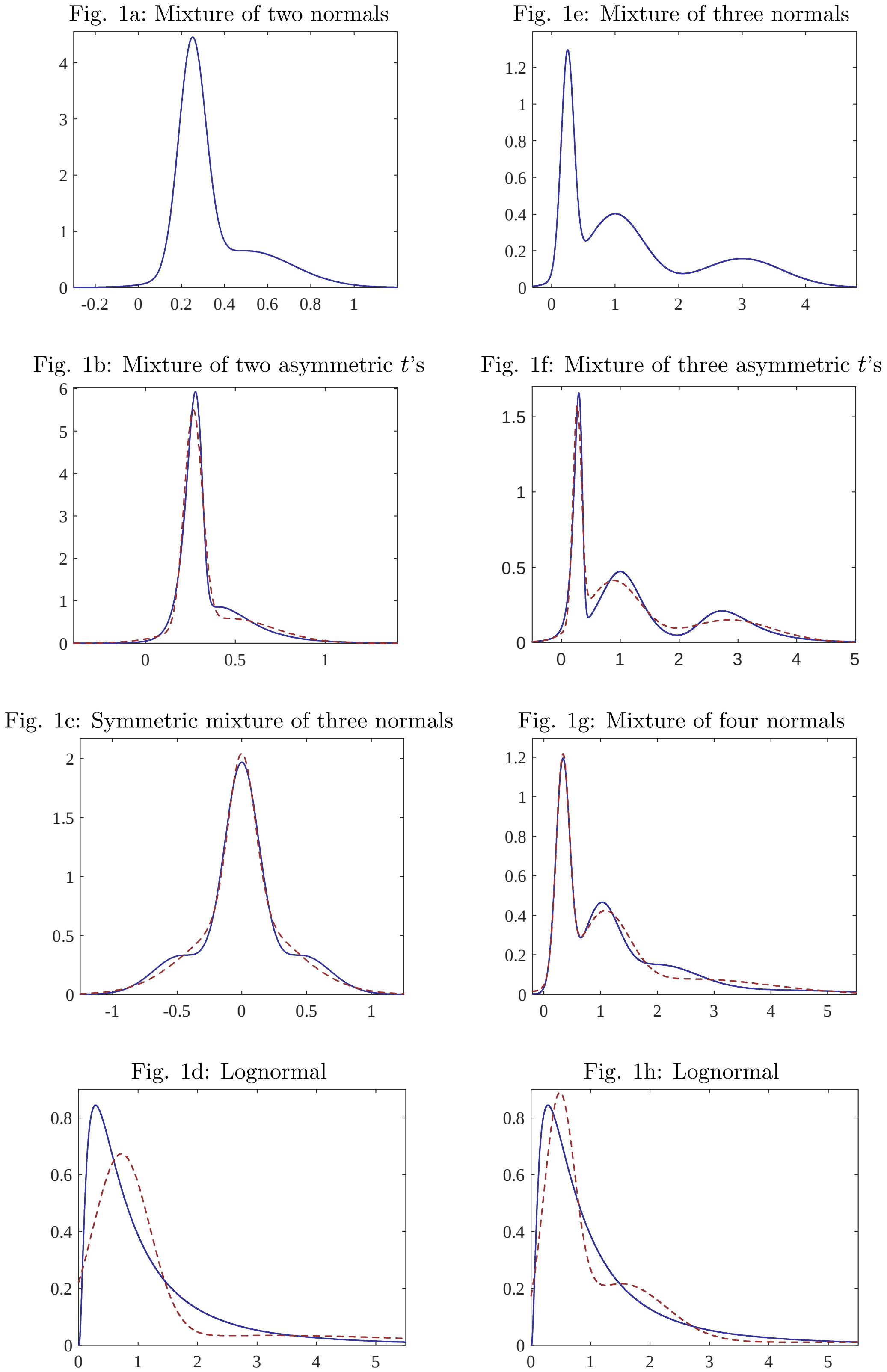

Fig. 1a: Mixture of two normals

Fig. 1e: Mixture of three normals

Fig. 1b: Mixture of two asymmetric t's

Fig. 1f: Mixture of three asymmetric t's

Fig. 1c: Symmetric mixture of three normals

Fig. 1g: Mixture of four normals

Fig. 1d: Lognormal

Fig. 1h: Lognormal

Notes: In figures 1b-d (1f-h) the dashed line represents the pdf of the closest mixture of two (three) normals. See section 4 for details about the DGPs.

Figure 2: Bivariate distributions under the null hypothesis and different alternatives

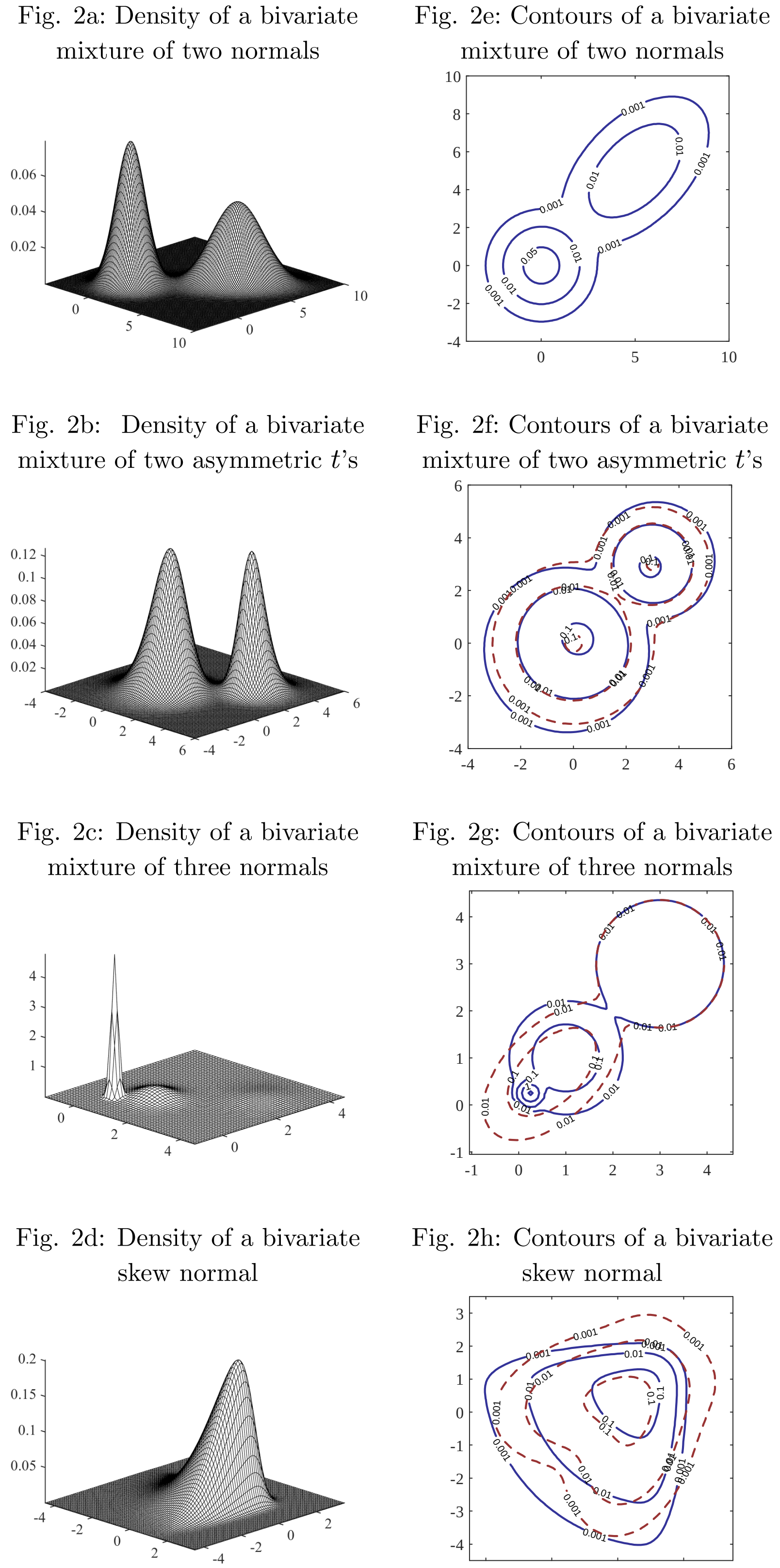

Fig. 2a: Density of a bivariate mixture of two normals

Fig. 2e: Contours of a bivariate mixture of two normals

Fig. 2b: Density of a bivariate mixture of two asymmetric t's

Fig. 2f: Contours of a bivariate mixture of two asymmetric t's

Fig. 2c: Density of a bivariate mixture of three normals

Fig. 2g: Contours of a bivariate mixture of three normals

Fig. 2d: Density of a bivariate skew normal

Fig. 2h: Contours of a bivariate skew normal

Notes: In figures 2f-h the dashed lines represent the contour of the closest mixture of two normals. See section 4 for details about the DGPs.

Figure 3: "Convergence clubs" in cross-country GDP per capita comparisons

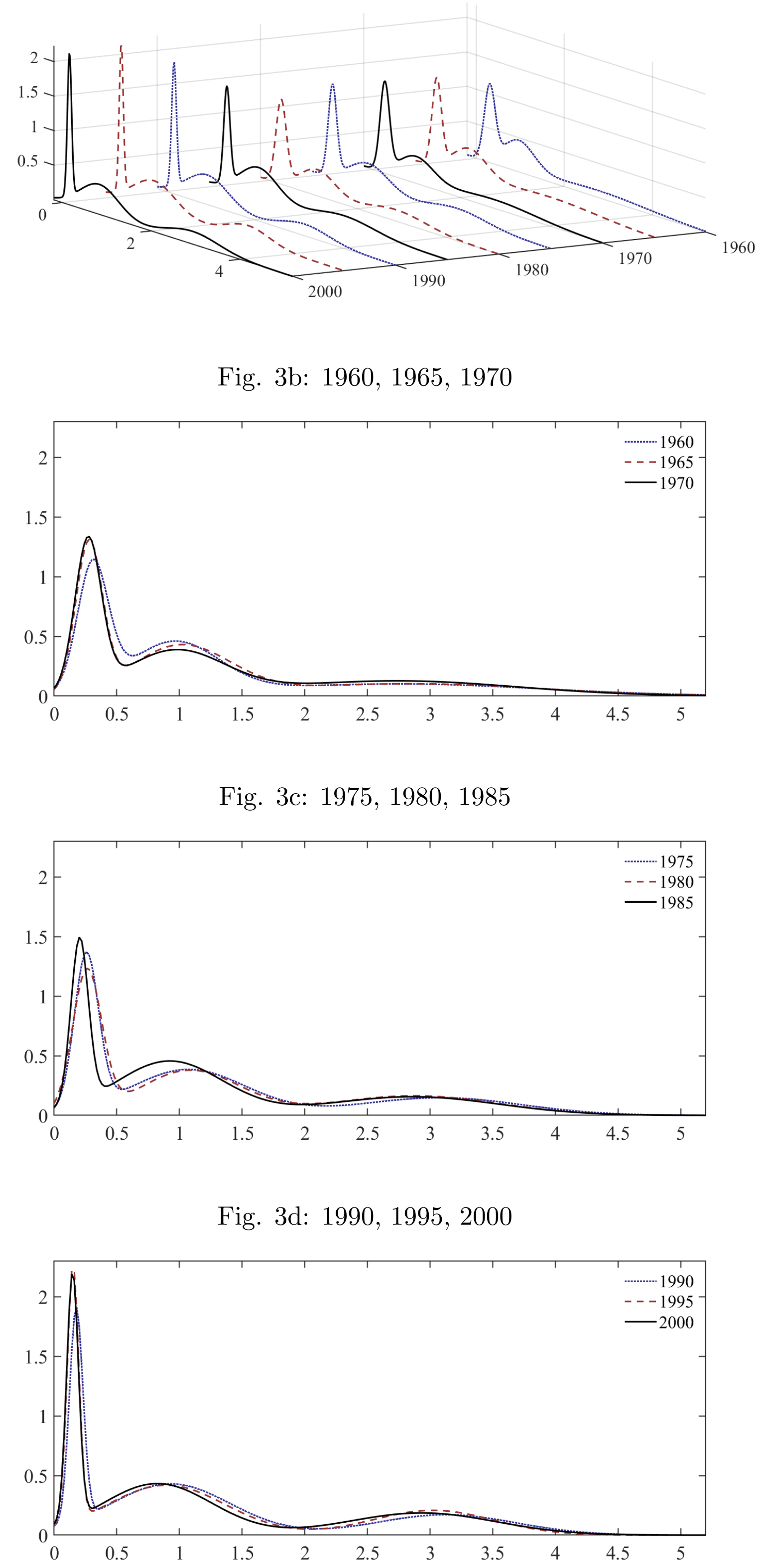

Fig. 3a: All waves

Fig. 3b: 1960, 1965, 1970

Fig. 3c: 1975, 1980, 1985

Fig. 3d: 1990, 1995, 2000

Notes: Data: Per capita income from version 6.1 of the Penn World Tables.