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: February 2026

Abstract

We obtain the Information Matrix test for finite Gaussian mixtures exploiting the EM principle, which implies the moments tested are the expectation given the data of the moments one would assess if one knew the latent component each observation belongs to. Simulations indicate analytical expressions for the asymptotic covariance matrix of those moments adjusted for sampling variability in parameter estimators provide reliable finite sample sizes and good power against various alternatives, especially combined with the parametric bootstrap. We apply the test to the univariate distribution of per capita income across countries and its joint distribution with per capita CO_2 emissions.

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

JEL: C46, C52, O44, O47

*We would like to thank Xinyue Bei, Otilia Boldea, Eric Gautier, Paul Johnson, Jan Magnus and Olivier Scaillet, as well as audiences at Duke, Florence, JRC, Liverpool, Montreal, the Barcelona Workshop in Financial Econometrics (ESADE, May 2023), the Workshop in memory of Stefano Fachin (Rome, February 2024), the TSE Financial Econometrics Conference (Toulouse, May 2024), IAAE (Thessaloniki, June 2024), Finance Forum (La Laguna, July 2024), ESEM (Rotterdam, August 2024), LAMES (Montevideo, November 2024), SAEe (Mallorca, December 2024), Encuentro de la SEU (Montevideo, December 2024) and NAWM (San Francisco, January 2025) for useful comments, discussions and suggestions. We are also grateful to Maria Grazia Pittau and Christopher Parmeter for sharing their data with us, and to Tao Wang for excellent research assistance. Feedback from the editor, associate editor and two anonymous referees have also led to substantial improvements. Of course, the usual caveat applies. The first and third authors gratefully acknowledge financial support from grant PID2021-128963NB-I00 funded by MICIU/AEI/1013039/501100011033 and ERDF/EU, 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 (2025) 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 study in detail a specification test for finite Gaussian mixtures that is not a priori targeted to either the number of components or their normality. Specifically, we follow Boldea and Magnus (2009), who discussed the information matrix (IM) test as part of their illustration of the usefulness of the formulas for the score vector and Hessian matrix of the log-likelihood function of finite mixtures of multivariate normals that they obtained using the tools of matrix differential calculus.

As is well known, 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. This result, also known as the second Barlett identity, justifies the calculation of the information matrix as either the (minus) expected value of the Hessian or the variance of the score, so that when it fails because the model is misspecified, it can have important consequences on the reliability of the reported standard errors.

Unlike Boldea and Magnus (2009), though, we rely on the EM principle to show that the moments involved in the IM test are the expectation given the data of the moments the IM test would assess if one knew the identity of the latent component to which each observation belongs.

But given that the influence functions underlying those moment tests for each component of the mixture effectively coincide with the list of all the distinct third- and fourth-order multivariate Hermite polynomials, as shown by Amengual, Fiorentini and Sentana (2024) for observed multivariate Gaussian random vectors, 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 three important advantages. First, it allows us to determine which of those influence functions is either redundant or spanned by the score vector, so that researchers know the correct rank of the relevant covariance matrix, and consequently, the right number of degrees of freedom for the IM test,[1] which in turn avoids the numerical calculation of generalised inverses. Second, it may prove particularly useful for the purposes of indicating in which specific directions modelling efforts to enrich finite mixture models should focus. Finally, 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.

In fact, our approach to deriving the IM test and its interpretation is relevant for 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. Microeconometric examples include the limited dependent variable models that Gouriéroux, Monfort, Renault and Trognon (1987) and Smith (1987) tackled with the same approach, as well as finite mixtures of other distributions and location-scale Gaussian mixtures in which the mixing variable is continuous. In turn, dynamic factor models and Markov switching ones are two important class of time series processes in which our approach yields useful insights too.

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 assess the adequacy of finite Gaussian mixtures in the empirical applications in Pittau, Zelli and Johnson (2010) and Battisti, Delgado and Parmeter (2015) that look at the univariate distribution of per capita income across countries and its joint distribution with per capita carbon dioxide (CO_2) emissions, respectively.

[1] Boldea and Magnus (2024) explicitly acknowledge that they became aware of the singularity that affects the number of degrees of freedom in their Theorem 2 thanks to Proposition 4 below.

The rest of the paper is organised as follows. In Section 2, we formally introduce the IM test 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 in which the latent variables are the mixture component indicators. 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

In this section, we begin by quickly reviewing the IM test and then we obtain a new result that will prove useful for the analysis of Gaussian mixtures in section 3, namely the relationship between the IM test in the complete and incomplete data contexts considered by Dempster, Laird and Rubin (1977).

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.

In what follows, we maintain the regularity conditions White (1982) assumed to prove his Theorem 4.1 on the IM test, which among other things guarantee the consistency of the max-

imum likely estimators (MLE) of the model parameters, $\hat{\boldsymbol{\phi}}_N$, and their asymptotic normality with a full-rank information matrix when we centre them around their true values, $\boldsymbol{\phi}_0$, and suitably scale them by $\sqrt{N}$.

In this context, the IM test directly assesses the IM equality, which states that the sum of the Hessian matrix and the outer product of the score (OPS) vector should be zero in expected value when the estimated model is correctly specified.

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

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

where $\mathbf{D}_p$ is the duplication matrix of order p, which is the unique $p^2 \times p(p+1)/2$ matrix that satisfies $\mathbf{D}_p vech(\mathbf{A}) = vec(\mathbf{A})$ for any $p \times p$ symmetric matrix $\mathbf{A}$, and $\mathbf{D}_p^{+}$ its Moore-Penrose inverse (see Magnus (1988)).

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

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

In his original paper, White (1982) obtained a first-order expansion of (2) around the true parameter values, which effectively required the expected value of the third-order derivatives of $l_i(\boldsymbol{\phi})$. However, Chesher (1983) and Lancaster (1984) realised that in a likelihood context such as this, the generalised information matrix equality implies that the expected value of the Jacobian of (1) with respect to $\boldsymbol{\phi}$ coincides with the (minus) covariance matrix between (1) and $\mathbf{s}_i(\boldsymbol{\phi})$ evaluated at the true values of the parameters, $\boldsymbol{\phi}_0$. Under our *i.i.d.* assumption, this implies that to obtain the asymptotic covariance matrix of (2), 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

$$\begin{bmatrix} \mathcal{R}(\boldsymbol{\phi}_0) & \mathcal{U}(\boldsymbol{\phi}_0) \\ \mathcal{U}'(\boldsymbol{\phi}_0) & \mathcal{I}(\boldsymbol{\phi}_0) \end{bmatrix} = V \left\{ \begin{array}{c} vech[\mathbf{h}_i(\boldsymbol{\phi}_0) + \mathbf{s}_i(\boldsymbol{\phi}_0)\mathbf{s}_i'(\boldsymbol{\phi}_0)] \\ \mathbf{s}_i(\boldsymbol{\phi}_0) \end{array} \right\}. \tag{4}$$

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

$$N\left\{\frac{1}{N}\sum_{i=1}^{N} vech'[\mathbf{h}_i(\hat{\boldsymbol{\phi}}_N)+\mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)\mathbf{s}_i'(\hat{\boldsymbol{\phi}}_N)]\right\}[\mathcal{R}(\boldsymbol{\phi}_0)-\mathcal{U}(\boldsymbol{\phi}_0)\mathcal{I}^{-1}(\boldsymbol{\phi}_0)\mathcal{U}(\boldsymbol{\phi}_0)]^{-}$$
$$\times\left\{\frac{1}{N}\sum_{i=1}^{N} vech[\mathbf{h}_i(\hat{\boldsymbol{\phi}}_N)+\mathbf{s}_i(\hat{\boldsymbol{\phi}}_N)\mathbf{s}_i'(\hat{\boldsymbol{\phi}}_N)]\right\}, \tag{5}$$

where the superscript "$-$" denotes a generalised inverse, which is often necessary 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, as we will see in section 3 below and Appendix C.

On this basis, Chesher (1983) and Lancaster (1984) also realised that a feasible version of the quadratic form (5) could be computed as the sample size 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 OPS 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 (5) 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 given that the asymptotic distribution of the test statistic (5) is chi-square, and therefore pivotal, 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 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 $\mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})$ denote the joint density of $\boldsymbol{\zeta}$ given a vector of parameters $\boldsymbol{\phi}$. We know from basic probability theory that

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

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*

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

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

Proposition 1, which is straightforward application of the law of iterated expectations, 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 $\mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})$ belongs to what they called a "bilinear" exponential family. 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 (5), though, we also need expressions for the different elements that appear in the theoretical expression (3). To obtain the required expressions, we find it convenient to obtain first a general result that applies to any conceivable influence function that one may want to use in a moment test. Let $\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})$ denote a vector of influence functions of the complete data $\boldsymbol{\zeta}$ such that

$$E_{\boldsymbol{\zeta}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})] = \mathbf{0} \tag{8}$$

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

$$\mathbf{m}(\mathbf{y};\boldsymbol{\phi}) = E_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})|\mathbf{y}], \tag{9}$$

where the subscript $\boldsymbol{\zeta}|\mathbf{y}$ denotes an expectation E or variance V taken with respect to the conditional density of the "complete data" $\boldsymbol{\zeta}$ given the "incomplete data" $\mathbf{y}$ defined over R, while the subscripts $\mathbf{y}$ and $\boldsymbol{\zeta}$ represent analogous moments with respect to the marginal distributions of $\mathbf{y}$ and $\boldsymbol{\zeta}$, respectively. 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 general context, we can use the law of iterated covariances to prove the following result, whose first part nests 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 $\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})/\partial\boldsymbol{\phi}$ when $\mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})$ belongs to a "bilinear" exponential family:

Proposition 2 *Let* $\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})$ *denote any vector of influence functions of the complete data* $\boldsymbol{\zeta}$ *that*

satisfy (8) and let $\mathbf{m}(\mathbf{y};\boldsymbol{\phi})$ *be given by (9). Then,*

$$V_{\mathbf{y}}[\mathbf{m}(\mathbf{y};\boldsymbol{\phi})] = V_{\boldsymbol{\zeta}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})] - E_{\mathbf{y}}\{V_{\boldsymbol{\zeta}|\mathbf{y}}[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})|\mathbf{y}]\} \tag{10}$$

and

$$E_{\boldsymbol{y}}\left[\frac{\partial \mathbf{m}(\boldsymbol{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right] = -E_{\boldsymbol{\zeta}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi})\frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right] + E_{\mathbf{y}}\left\{cov_{\boldsymbol{\zeta}|\mathbf{y}}\left[\mathbf{n}(\boldsymbol{\zeta};\boldsymbol{\phi}), \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\middle| \mathbf{y}\right]\right\}. \tag{11}$$

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

$$\left\{vech'\left[\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right], \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right\}', \tag{12}$$

whose elements are the conditional expected values of

$$\left\{vech'\left[\frac{\partial^2 \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right], \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right\}'. \tag{13}$$

Corollary 1 *The application of Proposition 2 to (12) yields*

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

$$\begin{aligned}\mathcal{U}(\boldsymbol{\phi}) &= E_{\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right]\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right\} \\ &= cov_{\boldsymbol{\zeta}}\left\{vech\left[\frac{\partial^2 \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right], \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right\} \\ &-E_{\mathbf{y}}\left[cov_{\boldsymbol{\zeta}|\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right], \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\middle|\mathbf{y}\right\}\right],\end{aligned} \tag{15}$$

and

$$\begin{aligned}\mathcal{R}(\boldsymbol{\phi}) &= V_{\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln f(\mathbf{y};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right]\right\} \\ &= V_{\boldsymbol{\zeta}}\left\{vech\left[\frac{\partial^2 \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right]\right\} \\ &-E_{\mathbf{y}}\left[V_{\boldsymbol{\zeta}|\mathbf{y}}\left\{vech\left[\frac{\partial^2 \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}\partial \boldsymbol{\phi}'} + \frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\frac{\partial \ln \mathfrak{f}(\boldsymbol{\zeta};\boldsymbol{\phi})}{\partial \boldsymbol{\phi}'}\right]\middle|\mathbf{y}\right\}\right].\end{aligned} \tag{16}$$

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

$$\mathbf{y} = \sum_{k=1}^K \xi_k(\boldsymbol{\nu}_k + \boldsymbol{\Gamma}_k^{1/2}\boldsymbol{\varepsilon}) \tag{17}$$

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

$$\boldsymbol{\tau} = E(\mathbf{y}) = \sum_{k=1}^K \lambda_k \boldsymbol{\nu}_k = E_{\boldsymbol{\xi}}[E_{\mathbf{y}|\boldsymbol{\xi}}(\mathbf{y})], \quad \text{and} \tag{18}$$

$$\boldsymbol{\Psi} = V(\mathbf{y}) = \sum_{k=1}^K \lambda_k[(\boldsymbol{\nu}_k\boldsymbol{\nu}_k') + \boldsymbol{\Gamma}_k] - \left(\sum_{k=1}^K \lambda_k \boldsymbol{\nu}_k\right)\left(\sum_{k=1}^K \lambda_k \boldsymbol{\nu}_k'\right) = E_{\boldsymbol{\xi}}[V_{\mathbf{y}|\boldsymbol{\xi}}(\mathbf{y})] + V_{\boldsymbol{\xi}}[E_{\mathbf{y}|\boldsymbol{\xi}}(\mathbf{y}\mathbf{y}')]. \tag{19}$$

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 $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_k, \ldots, \boldsymbol{\gamma}_K)'$, 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

$$\lambda_k = \frac{e^{\pi_k}}{\sum_{l=1}^{K-1} e^{\pi_l} + 1} \ (k = 1, \ldots K-1); \quad \lambda_K = \frac{1}{\sum_{l=1}^{K-1} e^{\pi_l} + 1}, \tag{20}$$

or one could make

$$\lambda_k = \pi_k, \ k = 1, \ldots, K-1 \text{ and } \lambda_K = 1 - \textstyle\sum_{l=1}^{K-1} \pi_l \tag{21}$$

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 first work with the K elements of $\boldsymbol{\lambda}$ as if they were unrestricted, and then use the chain rules for first and second derivatives to obtain the relevant expressions for the underlying $K-1$ parameters that impose the adding up constraint. As a result, the Jacobian matrices $\partial\boldsymbol{\lambda}/\partial\boldsymbol{\pi}'$ for (20) or (21) that appear at the end of Appendix E play an important role in the practical implementation of our IM tests. However, whether we use parametrisation (20) or (21) is inconsequential because Lemma 2 in Appendix D.3 implies that the IM test statistics are numerically invariant.[2] 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.

[2] In addition, the choice of excluded category in both (19) and (20) is arbitrary, so Lemma 2 is also useful to show that the IM test is not afected by it. Consequently, 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

$$l(\mathbf{y};\boldsymbol{\phi}) = \ln\left\{\sum_{k=1}^{K}\lambda_k|\boldsymbol{\Gamma}_k|^{-1/2}\phi_M[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\right\}, \tag{22}$$

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. The identification of finite mixtures of Gaussian distributions from this log-likelihood is proved in Proposition 1 of Teicher (1963) for the univariate case and Proposition 3 of Yakowitz and Spragins (1968) for the multivariate one. Of course, one can always include additional components to a Gaussian mixture that either simply replicate some of the existing ones or are assigned 0 probability, but those components would be redundant. For that reason, identifiability focuses on the mixing distribution. As a result, the true number of components is effectively understood to be the smallest number of components that generate the true distribution, excluding those with 0 probability. Under those circumstances, a finite Gaussian mixture model is a regular one, with a full-rank information matrix and parameter estimators that converge at the usual $N^{1/2}$ rate.

Boldea and Magnus (2009) is the first paper to provide detailed analytical expressions for the score and Hessian of finite mixtures of multivariate Gaussian normal variables when the mixing probabilities are parametrised as in (21). On the basis of the expressions in their Theorem 1, Boldea and Magnus (2009) derive the influence functions for the IM test in their Theorem 2. An alternative way of obtaining the 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 $\boldsymbol{\zeta}_i = (\mathbf{y}_i', \boldsymbol{\xi}_i')$. Not surprisingly, when we follow this alternative route, we end up with the following numerically equivalent expressions to theirs (see Appendix E for a detailed comparison between the two approaches):

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*

$$\boldsymbol{\nu}_k\boldsymbol{\nu}_k' \;:\; w_k(\boldsymbol{\phi})\boldsymbol{\Gamma}_k'^{-1/2}[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]\boldsymbol{\Gamma}_k^{-1/2}, \tag{23}$$

$$\begin{aligned}\boldsymbol{\nu}_k\boldsymbol{\gamma}_k' \;:\; & w_k(\boldsymbol{\phi})\frac{1}{2}\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M \\ & -w_k(\boldsymbol{\phi})[\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M,\end{aligned} \tag{24}$$

$$\boldsymbol{\nu}_k\lambda_k \;:\; w_k(\boldsymbol{\phi})\frac{1}{\lambda_k}\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k), \tag{25}$$

$$
\begin{aligned}
\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k' \quad &: \quad w_k(\boldsymbol{\phi})\frac{1}{4}\mathbf{D}_M'(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M] \\
&\quad \times vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M](\boldsymbol{\Gamma}_k^{-1/2}\otimes\boldsymbol{\Gamma}_k^{-1/2})\mathbf{D}_M \\
&\quad -w_k(\boldsymbol{\phi})\frac{1}{2}\mathbf{D}_M'\{2[(\boldsymbol{\Gamma}_k^{-1}\otimes\boldsymbol{\Gamma}_k'^{-1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2}]-(\boldsymbol{\Gamma}_k^{-1}\otimes\boldsymbol{\Gamma}_k^{-1})\}\mathbf{D}_M, \qquad (26) \\
\boldsymbol{\gamma}_k\lambda_k \quad &: \quad w_k(\boldsymbol{\phi})\frac{1}{2\lambda_k}\mathbf{D}_M'(\boldsymbol{\Gamma}_k'^{-1/2}\otimes\boldsymbol{\Gamma}_k'^{-1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)-\mathbf{I}_M], \qquad (27)
\end{aligned}
$$

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

$$
w_k(\boldsymbol{\phi}) = E(\xi_k|\mathbf{y};\boldsymbol{\phi}) = \Pr(\xi_k=1|\mathbf{y};\boldsymbol{\phi}) = \frac{\lambda_k|\boldsymbol{\Gamma}_k|^{-1/2}\phi_M[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]}{\sum_{l=1}^K \lambda_l|\boldsymbol{\Gamma}_l|^{-1/2}\phi_M[\boldsymbol{\varepsilon}_l^*(\boldsymbol{\theta}_l)]}. \qquad (28)
$$

However, not all those elements can be used as influence functions of the IM test. First, Boldea and Magnus (2009) note that (25) and (27) will be zero at the ML estimators because these vectors are proportional to the score vectors with respect to $\boldsymbol{\nu}_k$ and $\boldsymbol{\gamma}_k$, whose expressions appear in the proof of the proposition. They also indicate that there is no term for $\lambda_k\lambda_k$ or indeed any combination involving parameters from different components. Finally, (23) will also be zero because they are linear combinations of the score vector with respect to $\boldsymbol{\gamma}_k$. Therefore, we are left with (24) and (26), which contain $\frac{1}{2}M^2(M+1)$ and $\frac{1}{8}M(M+1)(M^2+M+2)$ distinct influence functions, respectively, for each component.

Unfortunately, those expressions still include redundant elements, as we explain in detail for the special case $M=K=2$ in Appendix C, which confirms the relevance of generalised inverses for the IM test of this model, as acknowledged by Boldea and Magnus (2024). Nevertheless, the calculation of the strictly necessary influence functions, their 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 (24) and (26) evaluated at the MLEs of the model parameters numerically coincides with a moment test based on the influence functions:*

$$
w_k(\boldsymbol{\phi})\left\{\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \\ \mathbf{H}_4[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)]\end{array}\right\}, \quad k=1,\ldots,K \qquad (29)
$$

evaluated at the same estimators, where

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

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 (29) corrected for the sampling uncertainty in estimating the model parameters under the null is the residual covariance matrix in the multivariate theoretical regression of (29) on*

$$w_k(\boldsymbol{\phi}) \left\{ \begin{array}{c} 1 \\ \mathbf{H}_1[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \\ \mathbf{H}_2[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)] \end{array} \right\}, \quad k = 1, \ldots, K. \tag{30}$$

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*

$$\frac{KM(M+1)(M+2)(M+7)}{24}. \tag{31}$$

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 in which $\mathbf{y}$ is Gaussian ($K = 1$).

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 (14), (15) and (16) to obtain expressions for the three elements of (3). Specifically, we can use expression (10) to write

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

where

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

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

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

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 (17), 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, as shown in Lemma 1 in Amengual, Fiorentini and Sentana (2024). As a result, it must be the case that

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

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 (29) with (30), 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, which is such that $vec(\mathbf{B}') = \mathbf{K}_{mn}vec(\mathbf{B})$ for any $m \times n$ matrix $\mathbf{B}$ (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 (30) 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 (29) and the remaining elements of (30) 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 (20) or (21). 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.

The expressions we have derived in this section also confirm that (4) will be singular when the number of estimated components exceeds the number of true components, which prevents the use of the IM matrix test in a general-to-specific search for K.

4 Monte Carlo simulations

4.1 Size

As stated in Proposition 4, the asymptotic distribution of our proposed IM test is χ^2 with degrees of freedom equal to (31). However, this asymptotic approximation might not be very reliable in finite samples. For that reason, we assess its validity by means of Monte Carlo simulations in a sequence of quadrupling sample sizes, namely $N = 100$, 400, 1,600, 6,400, 25,600 and 102,400. For each of the data generating processes (DGPs) we describe below, we generate $10,000$ samples under the null and compare the version of the IM statistic that relies on a feasible version of the theoretical expression (33) that replaces the true parameter values $\boldsymbol{\phi}_0$ with their MLEs $\hat{\boldsymbol{\phi}}_N$ with the OPS version proposed by Chesher (1983) and Lancaster (1984) and employed by Boldea and Magnus (2024), with the only difference that instead of relying on numerical generalised inverses, we exploit the rank deficiencies we highlight in the discussion that follows Proposition 3. Furthermore, for the empirically realistic sample sizes of $N = 100$, 400 and $1,600$, we also consider parametric bootstrap versions of these two procedures in which we simulate $B = 99$ samples from the mixture model estimated under the null. A formal justification of parametric bootstrap procedures for IM tests in regular models like the ones we consider in our experiments follows from the results in section 2.2 of Horowitz (1994), which in turn rely on earlier results by Beran (1988). Since the number of Monte Carlo replications is 10,000, the 95% asymptotic confidence intervals for the Monte Carlo rejection probabilities under the null are (9.41,10.59), (4.57,5.43) and (.80,1.20) at the 10%, 5% and 1% levels, respectively.

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 $\boldsymbol{\lambda} = K^{-1}\boldsymbol{\iota}_K$, $\boldsymbol{\Gamma}_k = dg[\hat{V}_N(\mathbf{y})]$, 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 (21) 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 Appendix D.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 (5) 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" that we obtain by Gaussian quadrature.

In the univariate case, we consider Gaussian mixtures of two and three components as null hypotheses. For the two-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) found for limited dependent models, the quality of the asymptotic approximation is much better when one uses the theoretical expressions for the weighting matrix instead even in very large samples. 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.

As our second univariate 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 section 5.1 (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. Nevertheless, the quality of the asymptotic approximation to the finite sample distribution of the parameter estimators is lower for the three-component mixture than for the two-component one for any given sample size, which is perhaps not surprising given that the number of estimated parameters is larger.

As for the bivariate case, given that the bootstrap takes considerable more CPU time, 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.

In summary, one can conclude that the chi-square asymptotic distribution in Proposition 4 is adequate for both versions of the test even though the finite sample rejection rates do not necessarily converge monotonically to their limiting values, but the finite sample reliability of our preferred version is notably higher than the NR^2 version used so far even after taking into account the singularity of the asymptotic covariance matrix of the sample means of the influence functions underlying the test evaluated at the MLEs. At the same time, the use of theoretical expressions does not completely eliminate the finite sample distortions of the IM test, and for that reason we recommend using the parametric bootstrap in combination with our proposed version.

4.2 Power

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 $N = 400$ because power is effectively 1 for the larger sample sizes. In view of the results in the previous subsection, 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 to correct the finite sample size distortions, as forcefully argued by Horowitz and Savin (2000).

As the first alternative hypothesis to the two-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). Given that asymmetric Student t's distributions can be understood as location-scale mixtures of normals, we can regard this alternative as an example of unobserved cross-sectional heterogeneity in the means and variances of the components, which is in line with the interpretation of the IM test in Chesher (1984). In addition, we consider a symmetric mixture of three normals that 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. Once again, this three-component mixture can be regarded as an example of a mixture of one normal component with a second component in which the means and variances take two possible values. 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. Interestingly, this alternative can be regarded as an example of a distribution in which there are fewer components than in the estimated model, but those components are not Gaussian. 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 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.

In turn, the first alternative hypothesis we consider to the three-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 two-component mixture as an example of a non-Gaussian distribution that can only be replicated by a Gaussian mixture with a countable number of components. 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 two-component case. The reason is twofold. First, the number of degrees of freedom increases. And second, Gaussian mixtures with an increasing number of components are able to approximate many distributions (see Hamdan (2006) for scale mixtures of normals, Nguyen et al (2020) for general ones, and Norets (2010) for conditional models). We would expect the same conclusions to apply in more complex examples in which the true DGP is a mixture of two non-Gaussian distributions and the estimated model is the mixture of three or more Gaussian components.

Finally, in the bivariate case, 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 that cannot be expressed as a Gaussian mixture with a finite number of components, 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). Once again, this alternative can be regarded as an example of a distribution in which there are fewer components than in the estimated model, but those components are not Gaussian. 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.

In summary, our results indicate that the IM test has power to detect empirically plausible deviations of the null in which the number of components is correctly specified but their distribution is not, the number of components is too low even though their distribution is correctly specified, or the number of components is too large but their distribution is not Gaussian.

5 Empirical applications

In this section we use the IM test to assess the validity of Gaussian mixtures in two published empirical applications that look at the univariate distribution of per capita income across countries and its joint distribution with per capita CO_2 emissions.

5.1 Convergence clubs

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

[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 $N = 100$.

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 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]

5.2 GDP growth and CO_2 emissions

As part of their extensive cross-country analysis of CO_2 emissions and their relationship to economic development, Battisti, Delgado and Parmeter (2015) estimate two-component mixtures of bivariate Gaussian distributions using data on (log) per capita emissions and (log) per capita GDP. Specifically, they consider a balanced panel data set of 84 countries that combines real GDP per capita from the Penn World Table 7.1 with fossil fuel-based CO_2 emissions from Marland, Boden and Andres (2003) at five year intervals for the period 1960-2005. Our estimates are somewhat different from theirs because we do not impose the restriction that the eigenvectors of the covariance matrices of the two underlying components are the same (see Bensmail and Celeux (1996)).[5] Nevertheless, we confirm their main result, which is that there is a low GDP-low emissions component and a high GDP-high emissions component, as can be seen from Figure 4 and Panel A of Table 5. Our results also confirm that the first group of countries is larger but stable, and that there are noticeable changes in means, variances and correlations over the sample period. Specifically, the rate of growth in average GDP per capita is larger for rich countries than for poor ones, while the opposite happens for the rate of growth in average CO_2 per capita. In addition, the dispersion within components is larger for emissions than

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

[5] As we explain in our concluding remarks, the modification required to deal with such parametric restrictions is tedious but otherwise straightforward.

for income, and while the standard deviations of both variables increase over time for the low GDP-low emissions group of countries, they decrease over time for the high GDP-high emissions one. Finally, the correlations between those two variables also follow different trends for the two groups: from medium (.64) to high (.74) for the poor countries and from high (.84) to low (.37) for the rich ones.

In this case, though, the bootstrap version of the IM matrix tests rejects the null of correct specification at the 5% significance level for 1970 and 1980 but not for the other years. To provide further insights into the nature of the rejections, we have carried out two additional tests that separately look at the (co-)skewnes and (co-)kurtosis components of the influence functions (29) for those two years. Interestingly, we find that the source of the rejections is related to the fourth-order moments (bootstrap p-values of 1.64% and 3.37% in 1970 and 1980, respectively) rather than the third-order ones (bootstrap p-values of 25.08% and 22.24%), which suggests that a mixture of elliptical distributions, such as the Student t, might be more appropriate.

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 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 that we have developed 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, like any moment test based on a finite number of influence functions, the IM test is not consistent. As White (1982) indicated, it will have trivial power against misspecifications that do not affect the validity in large samples of standard errors computed from either the OPS or the average Hessian matrix. We conjecture that this low power is likely to be relevant in admittedly contrived alternatives with the right number of components in which the distribution of some of the mixture 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 assess the adequacy of finite Gaussian mixtures to capture the univariate distribution of per capita income across countries and its joint distribution with per capita CO_2 emissions. In this respect, we 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). Our results also suggest that the two-component mixtures of bivariate normals considered by Battisti, Delgado and Parmeter (2015) provide a good fit to the joint distribution of income and emissions except for a couple of years in which a mixture of elliptical distributions would have been more appropriate.

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. Furthermore, we could also consider versions of the IM test based on parameter estimators different from the MLE. In this respect, the theory of moment tests in Newey (1985) and Tauchen (1985) would immediately provide the appropriate weighting matrix for the influence functions (1) evaluated at those alternative estimators.

In contrast, the model we have considered in this paper may be excessively general for some purposes. For example, an empirical researcher might have good reasons to restrict some elements of $\boldsymbol{\nu}_k$ or $\boldsymbol{\gamma}_k$ to be common across regimes. Nevertheless, the chain rules for first and second derivatives would immediately give us the score, Hessian and relevant influence functions and their asymptotic covariance matrix in those restricted models as a function of the corresponding elements in the unrestricted model that we have considered.

At the same time, our proposed IM tests 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. Second, we could allow the probabilities of the different regimes to be a function of some exogenous indicators using a multinomial logit model specification. 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). In Amengual, Fiorentini and Sentana (2025a,b,c), we explore these interesting research avenues.

Other relevant examples with a more microeconometric flavour would be discrete mixtures of beta, gamma or Poisson random variables, in which case the main change would be to replace the Hermite polynomials with Jacobi, Laguerre or Charlier ones (see Schoutens (2000) and Bontemps and Meddahi (2011)). Our approach would also be valid for continuous mixtures of Gaussian random variables, such as the Student t distribution or generalised hyperbolic (GH) distributions more generally, whose parameters are often estimated using the EM algorithm, by exploiting the fact that the posterior distribution of the underlying mixing variable given the observed ones will follow a reciprocal gamma in the case of the Student t or a generalised inverse Gaussian in the GH one (see Mencía and Sentana (2012)).

In fact, the IM tests considered in this paper can be extended to a much wider class of dynamic models with discrete and continuous latent variables that are routinely used in macroeconometric and empirical finance applications, including dynamic factor models (see Ren (2025)) and non-linear state space models with stochastic volatility and non-Gaussian shocks, as long as they can be written in the incomplete data framework of Dempster, Laird and Rubin (1977) after a suitable data augmentation. The main difference would be that numerical techniques, such as Markov chain Monte Carlo or particle filters, would often be required for smoothing purposes.

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. (2025): "Score-type tests for normal mixture models", *Journal of Econometrics* 248, 105717.

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

Amengual, D., Fiorentini, G. and Sentana, E. (2025a): "Information matrix tests for switching regression models", mimeo, CEMFI.

Amengual, D., Fiorentini, G. and Sentana, E. (2025b): "Information matrix tests for finite mixtures with covariate-dependent probabilities", mimeo, CEMFI.

Amengual, D., Fiorentini, G. and Sentana, E. (2025c): "The information matrix test for Markov switching autoregressive models with covariate-dependent transition probabilities", CEMFI Working Paper 2502.

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.

Battisti, M., Delgado, M.S. and Parmeter, C.F. (2015): "Evolution of the global distribution of carbon dioxide: a finite mixture analysis", *Resource and Energy Economics* 42, 31-52.

Bensmail, H. and Celeux, G. (1996): "Regularized Gaussian discriminant analysis through eigenvalue decomposition", *Journal of the American Statistical Association* 91, 1743-1748.

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", *Journal of the American Statistical Association* 119, 2423.

Bontemps, C. and Meddahi, N. (2011): "Testing distributional assumptions: a GMM approach", *Journal of Applied Econometrics* 27, 978-1012.

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.

Hamdan, H. (2006): "Characterizing and approximating infinite scale mixtures of normals", *Communications in Statistics - Theory and Methods* 35, 407-413.

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. (1988): *Linear structures*, Oxford University Press.

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

Marland, G., Boden, T.A. and Andres, R.J. (2003): "Global, regional and national CO_2 emissions", in *Trends: a compendium of data on global change.* Oak Ridge National Laboratory, Carbon Dioxide Information Analysis Center, Oak Ridge, TN.

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.

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

Norets, A. (2010): "Approximation of conditional densities by smooth mixtures of regressions", *Annals of Statistics* 38, 1733–1766.

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.

Ren, J. (2025): "The information matrix test for dynamic factor models", mimeo, CEMFI.

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.

Schoutens, W. (2000): *Stochastic processes and orthogonal polynomials*, Lecture Notes in Statistics 146, Springer.

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.

Teichner, H. (1963): "Identifiability of finite mixtures", *Annals of Mathematical Statistics* 34, 1265-1269.

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

Yakowitz, S.J. and Spragins, J.D. (1968): "On the identifiability of finite mixtures", *Annals of Mathematical Statistics* 39, 209-214.

Appendices

A Proofs

A.1 Proof of Proposition 1

Given that Assumption 1 allows us to interchange integration and differentiation, we can immediately combine expressions (3.1) and (3.2) in Louis (1982) to show that (7) holds. □

A.2 Proof or Proposition 2

First of all, note that (8) and (9) combined with the law of iterated variances immediately implies that

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

whence (10) follows.

In turn,

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

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

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

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

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

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

which coincides with (11). □

A.3 Proof or Proposition 3

First, we use the EM principle to derive expressions for the scores and Hessian. The complete log-likelihood function of a random sample of size N on $\boldsymbol{\zeta} = (\mathbf{y}', \boldsymbol{\xi}')'$ is given by

$$\sum_{i=1}^{N} \ln \mathfrak{f}(\boldsymbol{\zeta}_i;\boldsymbol{\phi}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \xi_{ki} \ln f(\mathbf{y}_i|\xi_{ki}=1;\boldsymbol{\theta}_k) + \sum_{i=1}^{N} \ln f(\boldsymbol{\xi}_i;\boldsymbol{\lambda}), \tag{A1}$$

where

$$\ln f(\mathbf{y}|\xi_k=1;\boldsymbol{\theta}_k) = -\frac{1}{2}\left[M\ln\pi + \ln|\boldsymbol{\Gamma}_k| + \boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*}(\boldsymbol{\theta}_k)\right], \tag{A2}$$

$$\ln f(\boldsymbol{\xi};\boldsymbol{\lambda}) = \sum_{k=1}^{K} \xi_k \ln \lambda_k. \tag{A3}$$

The sequential cut in (A1), (A2) and (A3) considerably simplifies the required expressions. Specifically,

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

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 (28). On this basis, we can use expression (3.1) in Louis (1982) to write

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

Similarly, the only non-zero elements of the first term of expression (3.2) in Louis (1982) will be

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

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

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

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

In contrast, the second term of expressions (3.2) in Louis (1982) is slightly more complex. Specifically, we get

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

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,

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

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

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

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

and

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

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,

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

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,

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

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. □

A.4 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 (31) as long as we interpret K as the effective number of components of the mixture. □

B Multivariate Hermite polynomials

We 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

$$H_{j_1 \ldots j_M}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}), \boldsymbol{\Delta}] \cdot e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\nu})'\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\nu})} = (-1)^j \frac{\partial^j}{(\partial y_1)^{j_1} \ldots (\partial y_M)^{j_M}} \left[e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\nu})'\boldsymbol{\Delta}(\mathbf{y}-\boldsymbol{\nu})} \right], \tag{B4}$$

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 (B4), 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

$$\mathbf{H}_j(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) = \begin{bmatrix} H_{k,0,\cdots,0}(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) \\ H_{k-1,1,\cdots,0}(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) \\ \vdots \\ H_{0,\cdots,0,k}(\boldsymbol{\varepsilon}; \boldsymbol{\Delta}) \end{bmatrix}, \tag{B5}$$

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 $\boldsymbol{\nu}\boldsymbol{\nu}$ term, $\mathbf{H}_3(\boldsymbol{\varepsilon}^*)$ for the $\boldsymbol{\nu}\boldsymbol{\gamma}$ term, and

$\mathbf{H}_4(\boldsymbol{\varepsilon}^*)$ for the $\boldsymbol{\gamma\gamma}$ one.

In Appendix C we will illustrate these results for the univariate and bivariate case.

C Redundant influence functions in two-component examples

In this appendix, we describe in detail the redundant influence functions of the IM test for univariate and bivariate two-component mixtures of Gaussian variables. In both cases, we first derive the influence functions of the complete model, which allows us to clearly identify their redundant elements.

C.1 Univariate case

When $M = 1$ and $K = 2$, expressions (A1), (A2) and (A3) simplify to

$$\sum_{i=1}^{N} \ln \mathfrak{f}(\boldsymbol{\zeta}_i; \boldsymbol{\phi}) = \sum_{i=1}^{N} [\xi_{1i} \ln f(y_i|\xi_{1i}=1; \boldsymbol{\theta}_1) + \xi_{2i} \ln f(y_i|\xi_{2i}=1; \boldsymbol{\theta}_2)] + \sum_{i=1}^{N} \ln f(\boldsymbol{\xi}_i; \boldsymbol{\lambda}), \tag{C6}$$

where

$$\ln f(y|\xi_k=1; \boldsymbol{\theta}_k) = -\frac{1}{2} \left[\ln \pi + \ln \gamma_k + \varepsilon^{*2}(\boldsymbol{\theta}_k)\right], \tag{C7}$$

$$\ln f(\boldsymbol{\xi}; \boldsymbol{\lambda}) = \xi_1 \ln \lambda_1 + \xi_2 \ln \lambda_2, \tag{C8}$$

and $\varepsilon^*(\boldsymbol{\theta}_k) = (y - \nu_k)/\gamma_k$.

This sequential cut leads to

$$\frac{\partial \ln \mathfrak{f}(y, \boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{\partial \nu_k} = \frac{\partial \ln f(y|\boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma})}{\partial \nu_k} = \xi_k \frac{1}{\gamma_k^{1/2}} \varepsilon^*(\boldsymbol{\theta}_k), \tag{C9}$$

$$\frac{\partial \ln \mathfrak{f}(y, \boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{\partial \gamma_k} = \frac{\partial \ln f(y|\boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma})}{\partial \gamma_k} = \xi_k \frac{1}{2\gamma_k} \left[\varepsilon^{*2}(\boldsymbol{\theta}_k) - 1\right], \tag{C10}$$

$$\frac{\partial \ln \mathfrak{f}(y, \boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{\partial \lambda_k} = \frac{\partial \ln f(\boldsymbol{\xi}; \boldsymbol{\lambda})}{\partial \lambda_k} = \xi_k \frac{1}{\lambda_k}. \tag{C11}$$

Similarly,

$$\frac{\partial^2 \ln \mathfrak{f}(y, \boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{(\partial \nu_k)^2} = \frac{\partial^2 \ln f(y|\boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma})}{(\partial \nu_k)^2} = -\xi_k \frac{1}{\gamma_k},$$

$$\frac{\partial^2 \ln \mathfrak{f}(y, \boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{\partial \nu_k \partial \gamma_k} = \frac{\partial^2 \ln f(y|\boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma})}{\partial \nu_k \partial \gamma_k} = -\xi_k \frac{1}{\gamma_k^{3/2}} \varepsilon^*(\boldsymbol{\theta}_k),$$

$$\frac{\partial^2 \ln \mathfrak{f}(y, \boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{(\partial \gamma_k)^2} = \frac{\partial^2 \ln f(y|\boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma})}{(\partial \gamma_k)^2} = -\xi_k \frac{[2\varepsilon^{*2}(\boldsymbol{\theta}_k) - 1]}{2\gamma_k^2}$$

and

$$\frac{\partial^2 \ln \mathfrak{f}(y, \boldsymbol{\xi}; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda})}{(\partial \lambda_k)^2} = \frac{\partial^2 \ln f(\boldsymbol{\xi}; \boldsymbol{\lambda})}{(\partial \lambda_k)^2} = -\xi_k \frac{1}{\lambda_k^2}, \tag{C12}$$

with all other cross-derivatives being 0 because of the sequential cut in (C6). In fact, there will be no IM influence functions associated to the cross-products of the conditional mean and variance parameters across subpopulations because the Hessian is block diagonal and the cross-products of the mean and variance score vectors for different components are identically 0 because $\xi_1\xi_2 = 0$. Therefore, the sum of the Hessian and outer product of the score yields the following relevant terms:

$$\nu_k\nu_k \quad : \quad \xi_k \frac{1}{\gamma_k}[\varepsilon^{*2}(\boldsymbol{\theta}_k) - 1], \tag{C13}$$

$$\nu_k\gamma_k \quad : \quad \xi_k \frac{1}{2\gamma_k^{3/2}}[\varepsilon^{*3}(\boldsymbol{\theta}_k) - 3\varepsilon^{*}(\boldsymbol{\theta}_k)], \tag{C14}$$

$$\nu_k\lambda_k \quad : \quad \xi_k \frac{1}{\lambda_k\gamma_k^{1/2}}\varepsilon^{*}(\boldsymbol{\theta}_k), \tag{C15}$$

$$\gamma_k\gamma_k \quad : \quad \xi_k \frac{1}{4\gamma_k^{2}}[\varepsilon^{*4}(\boldsymbol{\theta}_k) - 6\varepsilon^{*2}(\boldsymbol{\theta}_k) + 3], \tag{C16}$$

$$\gamma_k\lambda_k \quad : \quad \xi_k \frac{1}{2\lambda_k\gamma_k}[\varepsilon^{*2}(\boldsymbol{\theta}_k) - 1], \tag{C17}$$

$$\lambda_k\lambda_k \quad : \quad 0. \tag{C18}$$

where we have exploited the fact that $\xi_k^2 = \xi_k$ for $k = 1, 2$.

Hence, it is immediately clear that (C18) is useless. Intuitively, this simply reflects the fact that both (20) and (21) effectively represent an unrestricted, non-parametric model for the probabilities of the components. In addition, (C17) is proportional to (C13), which in turn is proportional to (C10), while (C15) is proportional to (C9). Consequently, the only terms left after regressing the IM influence functions of the complete model on its score will be

$$\begin{aligned} \xi_k \frac{1}{2\gamma_k^{3/2}}[\varepsilon^{*3}(\boldsymbol{\theta}_k) - 3\varepsilon^{*}(\boldsymbol{\theta}_k)] &= \xi_k \frac{1}{2\gamma_k^{3/2}} H_3[\varepsilon^{*}(\boldsymbol{\theta}_k)], \\ \xi_k \frac{1}{4\gamma_k^{2}}[\varepsilon^{*4}(\boldsymbol{\theta}_k) - 6\varepsilon^{*2}(\boldsymbol{\theta}_k) + 3] &= \xi_k \frac{1}{4\gamma_k^{2}} H_4[\varepsilon^{*}(\boldsymbol{\theta}_k)]. \end{aligned}$$

Therefore, when $M = 1$ and $K = 2$ we end up with only four moment conditions to test instead of fifteen, which is the number of potentially different elements that appear in (1).

If we then combine our Proposition 1 with expression (3.1) in Louis (1982), it is immediately obvious that in the case of the incomplete log-likelihood function the only four influence functions left to test would be

$$w_k(\boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \frac{1}{2\gamma_k^{3/2}} H_3[\varepsilon^{*}(\boldsymbol{\theta}_k)], \tag{C19}$$

$$w_k(\boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) \frac{1}{4\gamma_k^{2}} H_4[\varepsilon^{*}(\boldsymbol{\theta}_k)] \tag{C20}$$

for $k = 1, 2$, where

$$w_k(\boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = P(\xi_k = 1|y; \boldsymbol{\nu}, \boldsymbol{\gamma}, \boldsymbol{\lambda}) = \frac{\frac{\lambda_k}{\gamma_k^{1/2}}\phi[\varepsilon^*(\boldsymbol{\theta}_k)]}{\frac{\lambda_1}{\gamma_1^{1/2}}\phi[\varepsilon^*(\boldsymbol{\theta}_1)] + \frac{\lambda_2}{\gamma_2^{1/2}}\phi[\varepsilon^*(\boldsymbol{\theta}_2)]}$$

i.e., the third and fourth Hermite polynomials of the observed variable y standardised as if it belonged to the k^{th} component of the mixture weighted by the appropriate posterior probability.

C.2 Bivariate case

When $M = 2$ and $K = 2$, expressions (A1), (A2) and (A3) simplify to

$$\sum_{i=1}^{N} \ln \mathfrak{f}(\boldsymbol{\zeta}_i; \boldsymbol{\phi}) = \sum_{i=1}^{N} [\xi_{1i} \ln f(\mathbf{y}_i|\xi_{1i}=1; \boldsymbol{\theta}_1) + \xi_{2i} \ln f(\mathbf{y}_i|\xi_{2i}=1; \boldsymbol{\theta}_2)] + \sum_{i=1}^{N} \ln f(\boldsymbol{\xi}_i; \boldsymbol{\lambda}), \tag{C21}$$

where $\mathbf{y} = (y_1, y_2)'$,

$$\ln f(\mathbf{y}|\xi_k=1; \boldsymbol{\theta}_k) = -\frac{1}{2}\left[M \ln \pi + \ln |\boldsymbol{\Delta}_k| + \boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\right], \tag{C22}$$

$\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k) = (\mathbf{y} - \boldsymbol{\nu}_k)$, with $\boldsymbol{\nu}_k = (\nu_{1k}, \nu_{2k})'$, $\boldsymbol{\Delta}_k = \boldsymbol{\Gamma}_k^{-1}$, with $vech(\boldsymbol{\Delta}_k) = (\delta_{11k}, \delta_{12k}, \delta_{22k})$, and $\ln f(\boldsymbol{\xi}; \boldsymbol{\lambda})$ as in (C8). As a result, (C11), (C12) and (C18) continue to be valid, so, henceforth, we will focus on the parameters in the mean vectors and covariance matrices of the components.

The scores of $\ln f(\mathbf{y}|\xi_k=1; \boldsymbol{\theta}_k)$ with respect to the vector of mean parameters are

$$\mathbf{s}_{\boldsymbol{\nu}_k}(\mathbf{y}; \boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) = \boldsymbol{\Gamma}_k^{-1/2'}\boldsymbol{\varepsilon}^*(\boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) = \boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k) = \begin{pmatrix} \delta_{11k}\varepsilon_1(\boldsymbol{\nu}_k) + \delta_{12}\varepsilon_2(\boldsymbol{\nu}_k) \\ \delta_{12}\varepsilon_1(\boldsymbol{\nu}_k) + \delta_{22}\varepsilon_2(\boldsymbol{\nu}_k) \end{pmatrix}, \tag{C23}$$

which coincide with the $H_{10}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k]$ and $H_{01}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k]$ bivariate Hermite polynomials of $\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)$ in Barndorff-Nielsen and Petersen (1979) defined in Appendix B.

Similarly, the scores with respect to the covariance matrix parameters $\boldsymbol{\gamma}_k = (\gamma_{11k}, \gamma_{12k}, \gamma_{22k})'$ are given by

$$\begin{aligned} \mathbf{s}_{\boldsymbol{\gamma}_k}(\mathbf{y}; \boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) &= \frac{1}{2}\mathbf{D}_2'(\boldsymbol{\Gamma}_k^{-1/2'} \otimes \boldsymbol{\Gamma}_k^{-1/2'})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\nu}_k, \boldsymbol{\gamma}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) - \mathbf{I}_2] \\ &= \frac{1}{2}\mathbf{D}_2'vec[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}^*(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k], \end{aligned}$$

where

$$D_2' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

is the transpose of the duplication matrix and

$$
\begin{aligned}
& vec[\boldsymbol{\Delta}_K \boldsymbol{\varepsilon}^*(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k] \\
= & \left[\begin{array}{c}
\delta_{11k}^2\varepsilon_1^2(\boldsymbol{\nu}_k) + 2\delta_{11k}\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + \delta_{12k}^2\varepsilon_2^2(\boldsymbol{\nu}_k) - \delta_{11k} \\
\delta_{11k}\delta_{12k}\varepsilon_1^2(\boldsymbol{\nu}_k) + (\delta_{12k}^2 + \delta_{11k}\delta_{22k})\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + \delta_{22k}\delta_{12k}\varepsilon_2^2(\boldsymbol{\nu}_k) - \delta_{12k} \\
\delta_{11k}\delta_{12k}\varepsilon_1^2(\boldsymbol{\nu}_k) + (\delta_{12k}^2 + \delta_{11k}\delta_{22k})\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + \delta_{22k}\delta_{12k}\varepsilon_2^2(\boldsymbol{\nu}_k) - \delta_{12k} \\
\delta_{12k}^2\varepsilon_1^2(\boldsymbol{\nu}_k) + 2\delta_{12k}\delta_{22k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + \delta_{22k}^2\varepsilon_2^2(\boldsymbol{\nu}_k) - \delta_{22}
\end{array} \right].
\end{aligned}
$$

Therefore, those scores coincide with the $H_{20}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k]$, $H_{11}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k]$ and $H_{02}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k]$ bivariate Hermite polynomials of $\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)$ in Barndorff-Nielsen and Petersen (1979) presented in Appendix B.

In turn, the Hessian matrix is given by

$$
\begin{aligned}
\mathbf{h}_{\boldsymbol{\nu}_k\boldsymbol{\nu}_k}(\mathbf{y}; \boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) &= -\boldsymbol{\Gamma}_k^{-1} = -\boldsymbol{\Delta}_k, \\
\mathbf{h}_{\boldsymbol{\gamma}_k\boldsymbol{\nu}_k}(\mathbf{y}; \boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) &= -\mathbf{D}_2'[\boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) \otimes \boldsymbol{\Gamma}_k^{-1}] = -\mathbf{D}_2'[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k) \otimes \boldsymbol{\Delta}_k],
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{h}_{\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k}(\mathbf{y}; \boldsymbol{\nu}_k, \boldsymbol{\gamma}_k) &= -\frac{1}{2}\mathbf{D}_2'\{2[(\boldsymbol{\Gamma}_k^{-1} \otimes \boldsymbol{\Gamma}_k^{\prime -1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\nu}_k, \boldsymbol{\gamma}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\nu}_k, \boldsymbol{\gamma}_k)\boldsymbol{\Gamma}_k^{-1/2}] - (\boldsymbol{\Gamma}_k^{-1} \otimes \boldsymbol{\Gamma}_k^{-1})\}\mathbf{D}_2 \\
&= -\frac{1}{2}\mathbf{D}_2'\{2[(\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k] - (\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\}\mathbf{D}_2.
\end{aligned}
$$

For reasons entirely analogous to the ones in the univariate case, the only terms that are going to be relevant are the ones corresponding to the mean and variance parameters within each regime.

Given that the outer product matrix of the score will be

$$
\left\{ \begin{array}{cc}
\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta} & \\
\frac{1}{2}\mathbf{D}_2'vec[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k]\boldsymbol{\varepsilon}'(\boldsymbol{\nu})\boldsymbol{\Delta}_k & \Big\}, \\
\frac{1}{2}\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)vec'[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k]\mathbf{D}_2 & \\
\frac{1}{4}\mathbf{D}_2'vec[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k]vec'[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k]\mathbf{D}_2 & \Big\},
\end{array} \right.
$$

the sum of the outer product of the score and the Hessian yields the following three terms

$$
\begin{aligned}
\boldsymbol{\nu}_k\boldsymbol{\nu}_k &: \boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k \\
\boldsymbol{\gamma}_k\boldsymbol{\nu}_k &: \frac{1}{2}\mathbf{D}_2'\{vec[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k]\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k) - 2[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k) \otimes \boldsymbol{\Delta}_k]\}
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k &: \frac{1}{4}\mathbf{D}_2'vec[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k]vec'[\boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k - \boldsymbol{\Delta}_k]\mathbf{D}_2 \\
&\quad -\frac{1}{2}\mathbf{D}_2'\{2[(\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\boldsymbol{\Delta}_k] - (\boldsymbol{\Delta}_k \otimes \boldsymbol{\Delta}_k)\}\mathbf{D}_2.
\end{aligned}
$$

It is immediately obvious that the $\boldsymbol{\nu}_k\boldsymbol{\nu}_k$ block will be spanned by the second-order bivariate

polynomials that appear in $\mathbf{s}_{\boldsymbol{\gamma}_k}(\mathbf{y};\boldsymbol{\nu}_k,\boldsymbol{\gamma}_k)$.

In turn, the $\boldsymbol{\gamma}_k\boldsymbol{\nu}_k$ block will be given by the 3×2 matrix

$$\left[\begin{array}{c}
[\delta_{11k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)+2\delta_{11k}\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{12k}^2\varepsilon_2^2(\boldsymbol{\nu}_k)-\delta_{11k}]\\
\times[\delta_{11k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{12k}\varepsilon_2(\boldsymbol{\nu}_k)]-2\delta_{11k}[\delta_{11k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{12k}\varepsilon_2(\boldsymbol{\nu}_k)]\\
2[\delta_{11k}\delta_{12k}\varepsilon_1^2(\boldsymbol{\nu}_k)+(\delta_{12k}^2+\delta_{11k}\delta_{22})\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{22k}\delta_{12k}\varepsilon_2^2(\boldsymbol{\nu}_k)-\delta_{12k}]\\
\times[\delta_{11k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{12k}\varepsilon_2(\boldsymbol{\nu}_k)]-2[2\delta_{11k}\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)+(\delta_{12k}^2+\delta_{11k}\delta_{22k})\varepsilon_2(\boldsymbol{\nu}_k)]\\
[\delta_{12k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)+2\delta_{12k}\delta_{22k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{22k}^2\varepsilon_2^2(\boldsymbol{\nu}_k)-\delta_{22k}]\\
\times[\delta_{11k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{12k}\varepsilon_2(\boldsymbol{\nu}_k)]-2\delta_{12k}[\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{22k}\varepsilon_2(\boldsymbol{\nu}_k)]
\end{array}\right.$$

$$\left.\begin{array}{c}
[\delta_{11k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)+2\delta_{11k}\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{12k}^2\varepsilon_{2k}^2(\boldsymbol{\nu}_k)-\delta_{11k}]\\
\times[\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{22k}\varepsilon_2(\boldsymbol{\nu}_k)]-2\delta_{12k}[\delta_{11k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{12k}\varepsilon_2(\boldsymbol{\nu}_k)]\\
2[\delta_{11k}\delta_{12k}\varepsilon_1^2(\boldsymbol{\nu}_k)+(\delta_{12k}^2+\delta_{11k}\delta_{22k})\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{22k}\delta_{12k}\varepsilon_2^2(\boldsymbol{\nu}_k)-\delta_{12k}]\\
\times[\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{22k}\varepsilon_2(\boldsymbol{\nu}_k)]-2[(\delta_{12k}^2+\delta_{11k}\delta_{22k})\varepsilon_1(\boldsymbol{\nu}_k)+2\delta_{22k}\delta_{12k}\varepsilon_2(\boldsymbol{\nu}_k)]\\
[\delta_{12k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)+2\delta_{12k}\delta_{22k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{22k}^2\varepsilon_2^2(\boldsymbol{\nu}_k)-\delta_{22k}]\\
\times[\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{22k}\varepsilon_2(\boldsymbol{\nu}_k)]-2\delta_{22k}[\delta_{12}\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{22k}\varepsilon_2(\boldsymbol{\nu}_k)]
\end{array}\right]$$

It is tedious but trivial to see that the (2,1) and (2,2) elements of this matrix are twice as big as the (1,2) and (3,1) ones, respectively. Therefore, the number of different elements coincides with the number of different third moments, which is $M(M+1)(M+2)/6=4$ in the bivariate case. Those four terms are

$$\begin{array}{c}
\delta_{11k}^3\varepsilon_1^3(\boldsymbol{\nu}_k)+3\delta_{11k}^2\delta_{12k}\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+3\delta_{11k}\delta_{12k}^2\varepsilon_2^2(\boldsymbol{\nu}_k)\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{12k}^3\varepsilon_2^3(\boldsymbol{\nu}_k)\\
-3\delta_{11k}^2\varepsilon_1(\boldsymbol{\nu}_k)-3\delta_{11k}\delta_{12k}\varepsilon_2(\boldsymbol{\nu}_k)=H_{30}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k),\boldsymbol{\Delta}_k],
\end{array}$$

$$\begin{array}{c}
\delta_{11k}^2\delta_{12k}\varepsilon_1^3(\boldsymbol{\nu}_k)+\delta_{11k}\left(\delta_{11k}\delta_{22k}+2\delta_{12k}^2\right)\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{12k}(2\delta_{11k}\delta_{22k}+\delta_{12k}^2)\varepsilon_2^2(\boldsymbol{\nu}_k)\varepsilon_1(\boldsymbol{\nu}_k)\\
+\delta_{22k}\delta_{12k}^2\varepsilon_2^3(\boldsymbol{\nu}_k)-3\delta_{11k}\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)-(\delta_{11k}\delta_{22k}+2\delta_{12k}^2)\varepsilon_2(\boldsymbol{\nu}_k)=H_{21}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k),\boldsymbol{\Delta}_k],
\end{array}$$

$$\begin{array}{c}
\delta_{22k}^2\delta_{12k}\varepsilon_2^3(\boldsymbol{\nu}_k)+\delta_{22k}\left(\delta_{11k}\delta_{22k}+2\delta_{12k}^2\right)\varepsilon_2^2(\boldsymbol{\nu}_k)\varepsilon_1(\boldsymbol{\nu}_k)+\delta_{12k}\left(2\delta_{11k}\delta_{22k}+\delta_{12k}^2\right)\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)\\
+\delta_{11k}\delta_{12k}^2\varepsilon_1^3(\boldsymbol{\nu}_k)-(\delta_{11k}\delta_{22k}+2\delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)-3\delta_{12k}\delta_{22k}\varepsilon_2(\boldsymbol{\nu}_k)=H_{12}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k),\boldsymbol{\Delta}_k],
\end{array}$$

and

$$\begin{array}{c}
\delta_{22k}^3\varepsilon_2^3(\boldsymbol{\nu}_k)+3\delta_{22k}^2\delta_{12k}\varepsilon_2^2(\boldsymbol{\nu}_k)\varepsilon_1(\boldsymbol{\nu}_k)+3\delta_{22k}\delta_{12k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+\delta_{12k}^3\varepsilon_1^3(\boldsymbol{\nu}_k)\\
-3\delta_{22k}\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)-3\delta_{22k}^2\varepsilon_2(\boldsymbol{\nu}_k)=H_{03}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k),\boldsymbol{\Delta}_k],
\end{array}$$

which coincide with the four different bivariate Hermite polynomials of order three in Barndorff-Nielsen and Petersen (1979) presented in Appendix B.

Finally, the $\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k$ block is a symmetric 3×3 matrix with the following six different elements

$$
\begin{gathered}
(\mathbf{1},\mathbf{1}) : \delta_{11k}^4\varepsilon_1^4(\boldsymbol{\nu}_k) + 4\delta_{11k}^3\delta_{12k}\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + 6\delta_{11k}^2\delta_{12k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) - 6\delta_{11k}^3\varepsilon_1^2(\boldsymbol{\nu}_k) \\
+4\delta_{11k}\delta_{12k}^3\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) - 12\delta_{11k}^2\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + \delta_{12k}^4\varepsilon_2^4(\boldsymbol{\nu}_k) - 6\delta_{11k}\delta_{12k}^2\varepsilon_2^2(\boldsymbol{\nu}_k) + 3\delta_{11k}^2, \\
(\mathbf{2},\mathbf{1}) : 2\delta_{11k}^3\delta_{12k}\varepsilon_1^4(\boldsymbol{\nu}_k) + 2\delta_{11k}^2(\delta_{22k}\delta_{11k} + 3\delta_{12k}^2)\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+6\delta_{11k}\delta_{12k}(\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) - 12\delta_{11k}^2\delta_{12k}\varepsilon_1^2(\boldsymbol{\nu}_k) \\
+2\delta_{12k}^2(3\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) - 6\delta_{11k}(\delta_{22k}\delta_{11k} + 3\delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+2\delta_{12k}^3\delta_{22k}\varepsilon_2^4(\boldsymbol{\nu}_k) - 6\delta_{12k}(\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_2^2(\boldsymbol{\nu}_k) + 6\delta_{11k}\delta_{12k}, \\
(\mathbf{3},\mathbf{1}) : \delta_{11k}^2\delta_{12k}^2\varepsilon_1^4(\boldsymbol{\nu}_k) + 2\delta_{11k}\delta_{12k}(\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+(\delta_{11k}^2\delta_{22k}^2 + 4\delta_{11k}\delta_{12k}^2\delta_{22k} + \delta_{12k}^4)\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) - \delta_{11k}(\delta_{11k}\delta_{22k} + 5\delta_{12k}^2)\varepsilon_1^2(\boldsymbol{\nu}_k) \\
+2\delta_{12k}\delta_{22k}(\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) - 4\delta_{12k}(2\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+\delta_{12k}^2\delta_{22k}^2\varepsilon_2^4(\boldsymbol{\nu}_k) - \delta_{22k}(\delta_{11k}\delta_{22k} + 5\delta_{12k}^2)\varepsilon_2^2(\boldsymbol{\nu}_k) + \delta_{11k}\delta_{22k} + 2\delta_{12k}^2,
\end{gathered}
$$

$$
\begin{gathered}
(\mathbf{2},\mathbf{2}) : 4\delta_{11k}^2\delta_{12k}^2\varepsilon_1^4(\boldsymbol{\nu}_k) + 8\delta_{11k}\delta_{12k}(\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+4(\delta_{11k}^2\delta_{22k}^2 + 4\delta_{11k}\delta_{12k}^2\delta_{22k} + \delta_{12k}^4)\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) - 4\delta_{11k}(\delta_{11k}\delta_{22k} + 5\delta_{12k}^2)\varepsilon_1^2(\boldsymbol{\nu}_k) \\
+8\delta_{12k}\delta_{22k}(\delta_{11k}\delta_{22k} + \delta_{12k}^3)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) - 16\delta_{12k}(2\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+4\delta_{12k}^2\delta_{22k}^2\varepsilon_2^4(\boldsymbol{\nu}_k) - 4\delta_{22k}(\delta_{11k}\delta_{22k} + 5\delta_{12k}^2)\varepsilon_2^2(\boldsymbol{\nu}_k) + 4\delta_{11k}\delta_{22k} + 8\delta_{12k}^2, \\
(\mathbf{3},\mathbf{2}) : 2\delta_{11}\delta_{12k}^3\varepsilon_1^4(\boldsymbol{\nu}_k) + 2\delta_{12k}^2(\delta_{12k}^2 + 3\delta_{11k}\delta_{22k})\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+6\delta_{12k}\delta_{22k}(\delta_{12k}^2 + \delta_{11k}\delta_{22k})\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) - 6\delta_{12k}(\delta_{12k}^2 + \delta_{11k}\delta_{22k})\varepsilon_1^2(\boldsymbol{\nu}_k) \\
+2\delta_{22k}^2(3\delta_{12k}^2 + \delta_{11k}\delta_{22k})\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) - 6\delta_{22k}(3\delta_{12k}^2 + \delta_{11k}\delta_{22k})\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+2\delta_{12k}\delta_{22k}^3\varepsilon_2^4(\boldsymbol{\nu}_k) - 12\delta_{12k}\delta_{22k}^2\varepsilon_2^2(\boldsymbol{\nu}_k) + 6\delta_{12k}\delta_{22k},
\end{gathered}
$$

and

$$
\begin{gathered}
(\mathbf{3},\mathbf{3}) : \delta_{12k}^4\varepsilon_1^4(\boldsymbol{\nu}_k) + 4\delta_{12k}^3\delta_{22k}\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + 6\delta_{12k}^2\delta_{22k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) \\
-6\delta_{12k}^2\delta_{22k}\varepsilon_1^2(\boldsymbol{\nu}_k) + 4\delta_{12k}\delta_{22k}^3\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) - 12\delta_{12k}\delta_{22k}^2\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
+\delta_{22k}^4\varepsilon_2^4(\boldsymbol{\nu}_k) - 6\delta_{22k}^3\varepsilon_2^2(\boldsymbol{\nu}_k) + 3\delta_{22k}^2.
\end{gathered}
$$

It is tedious but otherwise straightforward to prove that the (2,2) element is four times the (3,1) one. Therefore, the number of different elements coincides with the number of different fourth moments, which is $M(M+1)(M+2)(M+3)/24 = 5$ in the bivariate case. Those five terms are

$$
\begin{gathered}
\delta_{11k}^4\varepsilon_1^4(\boldsymbol{\nu}_k) + 4\delta_{11k}^3\delta_{12k}\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + 6\delta_{11k}^2\delta_{12k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) + 4\delta_{11k}\delta_{12k}^3\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) \\
+\delta_{12k}^4\varepsilon_2^4(\boldsymbol{\nu}_k) - 6\delta_{11k}^3\varepsilon_1^2(\boldsymbol{\nu}_k) - 12\delta_{11k}^2\delta_{12k}\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) - 6\delta_{11k}\delta_{12k}^2\varepsilon_2^2(\boldsymbol{\nu}_k) + 3\delta_{11k}^2 = H_{40}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k],
\end{gathered}
$$

$$
\begin{gathered}
2\delta_{11k}^3\delta_{12k}\varepsilon_1^4(\boldsymbol{\nu}_k) + 2\delta_{11k}^2(\delta_{11k}\delta_{22k} + 3\delta_{12k}^2)\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) + 6\delta_{11k}\delta_{12k}(\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k) \\
+2\delta_{12k}^2(3\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k) + 2\delta_{12k}^3\delta_{22k}\varepsilon_2^4(\boldsymbol{\nu}_k) - 12\delta_{11k}^2\delta_{12k}\varepsilon_1^2(\boldsymbol{\nu}_k) \\
-6\delta_{11k}(\delta_{11k}\delta_{22k} + 3\delta_{12k}^2)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) - 6\delta_{12k}(\delta_{11k}\delta_{22k} + \delta_{12k}^2)\varepsilon_2^2(\boldsymbol{\nu}_k) + 6\delta_{11k}\delta_{12k} = 2H_{31}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k],
\end{gathered}
$$

$$
\begin{gathered}
\delta_{11k}^2\delta_{12k}^2\varepsilon_1^4(\boldsymbol{\nu}_k) + 2\delta_{11k}\delta_{12k}\left(\delta_{11k}\delta_{22k} + \delta_{12k}^2\right)\varepsilon_2(\boldsymbol{\nu}_k)\varepsilon_1^3(\boldsymbol{\nu}_k) \\
+\left(\delta_{11k}^2\delta_{22k}^2 + 4\delta_{11k}\delta_{12k}^2\delta_{22k} + \delta_{12k}^4\right)\varepsilon_2^2(\boldsymbol{\nu}_k)\varepsilon_1^2(\boldsymbol{\nu}_k) + 2\delta_{12k}\delta_{22k}\left(\delta_{11k}\delta_{22k} + \delta_{12k}^2\right)\varepsilon_2^3(\boldsymbol{\nu}_k)\varepsilon_1(\boldsymbol{\nu}_k) \\
+\delta_{12k}^2\delta_{22k}^2\varepsilon_2^4(\boldsymbol{\nu}_k) - \delta_{11k}\left(\delta_{11k}\delta_{22k} + 5\delta_{12k}^2\right)\varepsilon_1^2(\boldsymbol{\nu}_k) - 4\delta_{12k}\left(2\delta_{11k}\delta_{22k} + \delta_{12k}^2\right)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k) \\
-\delta_{22k}(\delta_{11k}\delta_{22k} + 5\delta_{12k}^2)\varepsilon_2^2(\boldsymbol{\nu}_k) + \left(\delta_{11k}\delta_{22k} + 2\delta_{12k}^2\right) = H_{22}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k), \boldsymbol{\Delta}_k],
\end{gathered}
$$

$$2\delta_{11k}\delta_{12k}^3\varepsilon_1^4(\boldsymbol{\nu}_k)+2\delta_{12k}^2\left(3\delta_{11k}\delta_{22k}+\delta_{12k}^2\right)\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+6\delta_{12k}\delta_{22k}\left(\delta_{11k}\delta_{22k}+\delta_{12k}^2\right)\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k)$$
$$+2\delta_{22k}^2\left(\delta_{11k}\delta_{22k}+3\delta_{12k}^2\right)\varepsilon_2^3(\boldsymbol{\nu}_k)\varepsilon_1(\boldsymbol{\nu}_k)+2\delta_{12k}\delta_{22k}^3\varepsilon_2^4(\boldsymbol{\nu}_k)-6\delta_{12k}\left(\delta_{12k}+\delta_{12k}^2\right)\varepsilon_1^2(\boldsymbol{\nu}_k)$$
$$-6\delta_{22k}\left(\delta_{11k}\delta_{22k}+3\delta_{12k}^2\right)\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)-12\delta_{12k}\delta_{22k}^2\varepsilon_2^2(\boldsymbol{\nu}_k)+6\delta_{12k}\delta_{22k}=2H_{13}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k),\boldsymbol{\Delta}_k]$$

and

$$\delta_{12k}^4\varepsilon_1^4(\boldsymbol{\nu}_k)+4\delta_{12k}^3\delta_{22k}\varepsilon_1^3(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)+6\delta_{12k}^2\delta_{22k}^2\varepsilon_1^2(\boldsymbol{\nu}_k)\varepsilon_2^2(\boldsymbol{\nu}_k)+4\delta_{12k}\delta_{22k}^3\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2^3(\boldsymbol{\nu}_k)+\delta_{22k}^4\varepsilon_2^4(\boldsymbol{\nu}_k)$$
$$-6\delta_{12k}^2\delta_{22k}\varepsilon_1^2(\boldsymbol{\nu}_k)-12\delta_{12k}\delta_{22k}^2\varepsilon_1(\boldsymbol{\nu}_k)\varepsilon_2(\boldsymbol{\nu}_k)-6\delta_{22k}^3\varepsilon_2^2(\boldsymbol{\nu}_k)+3\delta_{22k}^2=H_{04}[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k),\boldsymbol{\Delta}_k],$$

which are (multiples of) the five different bivariate Hermite polynomials of order four in Barndorff-Nielsen and Petersen (1979) presented in Appendix B.

Therefore, in addition to the singularities that we described in the univariate case, we have additional singularities in the multivariate one reflecting the fact that the number of third- and fourth-order bivariate polynomials are four and five respectively, which are smaller than the six potentially different elements in the $\boldsymbol{\gamma}_k\boldsymbol{\nu}_k$ block and the other six elements in the $\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k$ one.

Once again, Proposition 1 immediate implies that the same singularities that arise in the complete log-likelihood arise in the incomplete one. Therefore, the list of influence functions to test will be

$$w_k(\boldsymbol{\phi})\left\{\begin{array}{c}\mathbf{H}_3[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k);\boldsymbol{\Delta}_k]\\ \mathbf{H}_4[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k);\boldsymbol{\Delta}_k]\end{array}\right\}, \tag{C24}$$

where $w_k(\boldsymbol{\phi})$ is given in (28). Finally, the fact that

$$\begin{aligned}
\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k) &= \boldsymbol{\Gamma}_k^{1/2}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\\
vec[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)] &= (\boldsymbol{\Gamma}_k^{1/2}\otimes\boldsymbol{\Gamma}_k^{1/2})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)],\\
vec\{vec[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)]\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)\} &= (\boldsymbol{\Gamma}_k^{1/2}\otimes\boldsymbol{\Gamma}_k^{1/2}\otimes\boldsymbol{\Gamma}_k^{1/2})vec\{vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)]\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\},\\
vec\{vec[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)]vec'[\boldsymbol{\varepsilon}(\boldsymbol{\nu}_k)\boldsymbol{\varepsilon}'(\boldsymbol{\nu}_k)]\} &= (\boldsymbol{\Gamma}_k^{1/2}\otimes\boldsymbol{\Gamma}_k^{1/2}\otimes\boldsymbol{\Gamma}_k^{1/2}\otimes\boldsymbol{\Gamma}_k^{1/2})\\
&\quad\times vec\{vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)]vec'[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)]\}
\end{aligned}$$

implies that the influence functions that appear in Proposition 4 span the influence functions (C24).

D Practical considerations

D.1 Using the EM algorithm for obtaining initial values

To maximise (22) 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:

$$\hat{\boldsymbol{\nu}}_k^{(h)} = \frac{1}{\hat{\lambda}_k^{(h)}} \frac{1}{N} \sum_{i=1}^{N} w_{ki}(\boldsymbol{\nu}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}, \boldsymbol{\lambda}^{(h-1)}) \mathbf{y}_i, \tag{D25a}$$

$$\hat{\boldsymbol{\Gamma}}_k^{(h)} = \frac{1}{\hat{\lambda}_k^{(h)}} \frac{1}{N} \sum_{i=1}^{N} w_{ki}(\boldsymbol{\nu}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}, \boldsymbol{\lambda}^{(h-1)}) \mathbf{y}_i \mathbf{y}_i' - \hat{\boldsymbol{\nu}}_k^{(h)} \hat{\boldsymbol{\nu}}_k^{(h)\prime}, \quad \text{and} \tag{D25b}$$

$$\hat{\lambda}_k^{(h)} = \frac{1}{N} \sum\nolimits_{i=1}^{N} w_{ki}(\boldsymbol{\nu}^{(h-1)}, \boldsymbol{\gamma}^{(h-1)}, \boldsymbol{\lambda}^{(h-1)}), \tag{D25c}$$

Given that (28) 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 $\sum_{k=1}^{K} w_k(\boldsymbol{\phi}) = 1$ for all $\mathbf{y}$ 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 the proof of Proposition 3. 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. In fact, important numerical gains can be achieved by applying the quadratically convergent algorithm to standardised data, especially in the multivariate case. The next three subsections of this appendix provide a formal justification to our claim.

D.2 Numerical invariance to affine transformations of the data

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 1 *The IM test statistics of model (17) and the analogous one for* $\mathbf{x}$ *numerically coincide.*

Proof. 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 (D25a), (D25b) and (D25c) imply that the MLEs of the mean vectors and covariance matrices of the different components will satisfy $\mathbf{c}+\mathbf{D}\hat{\boldsymbol{\nu}}_k$ and $\mathbf{D}\hat{\boldsymbol{\Gamma}}_k\mathbf{D}'$, 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. ■

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 Lemmas 1 and 2 to simplify the calculation of the IM statistic as follows. First, as we explain in Appendix D.4, 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 $\boldsymbol{\tau}$, $\boldsymbol{\aleph}$ 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.

D.3 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 $\boldsymbol{\phi} = r(\boldsymbol{\varphi})$.

The chain rules for first and second derivatives imply that the influence functions underlying the IM test of the reparametrised model will be

$$\left[\frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \otimes \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}}\right] vec\left[\mathbf{h}_i(\boldsymbol{\phi}) + \mathbf{s}_i(\boldsymbol{\phi})\mathbf{s}_i'(\boldsymbol{\phi})\right] + vec\left\{\mathbf{s}_i(\boldsymbol{\phi}) \otimes \mathbf{I}_p \frac{\partial vec[\partial r'(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}]}{\partial \boldsymbol{\varphi}'}\right\}. \tag{D26}$$

Then, we can show that:

Lemma 2 *The infeasible IM test statistic in (5) 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 (D26) written in terms of $\boldsymbol{\varphi}$.*

Proof. The following relationships will prove useful:

$$\begin{aligned} g(\mathbf{y}_i; \boldsymbol{\varphi}) &= f[\mathbf{y}_i; r(\boldsymbol{\varphi})], \\ \frac{\partial \ln g(\mathbf{y}_i; \boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} &= \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \frac{\partial l_i(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}} = \frac{\partial r'(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \mathbf{s}_i(\boldsymbol{\phi}) \end{aligned} \tag{D27}$$

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 (D26).

But

$$vec\left\{\left[\mathbf{s}_i'(\boldsymbol{\phi})\otimes\mathbf{I}_p\right]\frac{\partial vec[\partial r'(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}]}{\partial\boldsymbol{\varphi}'}\right\}=\left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}]}{\partial\boldsymbol{\varphi}}\otimes\mathbf{I}_p\right\}vec\left[\mathbf{s}_i'(\boldsymbol{\phi})\otimes\mathbf{I}_p\right]$$
$$=\left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}]}{\partial\boldsymbol{\varphi}}\otimes\mathbf{I}_p\right\}(\mathbf{I}_p\otimes\mathbf{K}_{p1}\otimes\mathbf{I}_p)\left\{vec\left[\mathbf{s}_i'(\boldsymbol{\phi})\right]\otimes vec(\mathbf{I}_p)\right\}$$
$$=\left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}]}{\partial\boldsymbol{\varphi}}\otimes\mathbf{I}_p\right\}\{\mathbf{s}_i(\boldsymbol{\phi})\otimes vec(\mathbf{I}_p)\}$$
$$=\left\{\frac{\partial vec'[\partial r'(\boldsymbol{\varphi})/\partial\boldsymbol{\varphi}]}{\partial\boldsymbol{\varphi}}\otimes\mathbf{I}_p\right\}[\mathbf{I}_p\otimes vec(\mathbf{I}_p)]\mathbf{s}_i(\boldsymbol{\phi}) \tag{D28}$$

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, (D26) can be written as an (admittedly complex) linear combination of (D26) and $\mathbf{s}_i(\boldsymbol{\phi}_0)$.

In fact, if we ignored the additional term (D28), 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 (D27) 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 (D28), though, does not affect this residual covariance matrix because it is a linear combination of $\mathbf{s}_i(\boldsymbol{\phi}_0)$, and consequently, of (D27) evaluated at $\boldsymbol{\varphi}_0$ too. ■

Intuitively, the sample average of the second summand in (D26) is exactly zero when evaluated at $\hat{\boldsymbol{\varphi}}_N$, so effectively, the influence functions (D26) 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 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.

Lemma 2 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).

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 $\boldsymbol{\phi} = (\mu, \sigma^2)'$,

$$\mathbf{s}(\boldsymbol{\phi}) = \begin{bmatrix} (y-\mu)/\sigma^2 \\ (y-\mu)^2/(2\sigma^4) - 1/\sigma^2 \end{bmatrix}$$

and

$$\mathbf{h}(\boldsymbol{\phi}) = -\begin{bmatrix} 1/\sigma^2 & (y-\mu)/\sigma^4 \\ (y-\mu)/\sigma^4 & (y-\mu)^2/(2\sigma^6) - 1/(2\sigma^4) \end{bmatrix}.$$

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

$$\frac{\partial \ln g(y;\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} = \begin{bmatrix} y/\psi - \tau \\ (y^2 - \tau\psi y + \psi^2)/\psi^3 \end{bmatrix}$$

and

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

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

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

and

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

and using the fact that

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

we can easily verify through straightforward calculations that the result (D26) is indeed correct.

D.4 Standardised multivariate discrete mixtures of normals

Consider the following mixture of two multivariate normals

$$\mathbf{y} \sim \begin{cases} N(\boldsymbol{\nu}_1, \boldsymbol{\Gamma}_1) & \text{with probability } \pi, \\ N(\boldsymbol{\nu}_2, \boldsymbol{\Gamma}_2) & \text{with probability } 1-\pi. \end{cases} \tag{D29}$$

Given (18) and (19), 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

$$\boldsymbol{\Gamma}_{1L}=[\pi\mathbf{I}_M+(1-\pi)\boldsymbol{\aleph}_L\boldsymbol{\aleph}_L']_U^{-1'}\ \text{ and }\ \boldsymbol{\Gamma}_{2L}=\boldsymbol{\Gamma}_{1L}\boldsymbol{\aleph}_L, \tag{D30}$$

where $\boldsymbol{\aleph}_L$ is a lower triangular matrix and $[\pi\mathbf{I}_M+(1-\pi)\boldsymbol{\aleph}_L\boldsymbol{\aleph}_L']_U[\pi\mathbf{I}_M+(1-\pi)\boldsymbol{\aleph}_L\boldsymbol{\aleph}_L']_U'$ is the upper triangular Cholesky decomposition of $[\pi\mathbf{I}_M+(1-\pi)\boldsymbol{\aleph}_L\boldsymbol{\aleph}_L']$, 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 (D30), 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 (D30), 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 $\boldsymbol{\eta} = (\boldsymbol{\delta}', vech'(\boldsymbol{\aleph}_L), \pi)'$.

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

$$\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\Psi}_L \boldsymbol{\varepsilon}^*, \text{ where } \boldsymbol{\varepsilon}^* = \begin{cases} N[\boldsymbol{\nu}_1^*(\boldsymbol{\eta}), \boldsymbol{\Gamma}_1^*(\boldsymbol{\eta})] \text{ with probability } \pi \\ N[\boldsymbol{\nu}_2^*(\boldsymbol{\eta}), \boldsymbol{\Gamma}_2^*(\boldsymbol{\eta})] \text{ with probability } 1-\pi \end{cases}, \tag{D31}$$

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

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

we will have

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

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

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

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

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 $\boldsymbol{\Psi}_L$ and $\boldsymbol{\aleph}_L$ 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 $\boldsymbol{\tau}$, $vech(\boldsymbol{\Psi})$ and $\boldsymbol{\eta}$ from the scores with respect to $\boldsymbol{\nu}_1$, $\boldsymbol{\nu}_2$, $vech(\boldsymbol{\Gamma}_1)$, $vech(\boldsymbol{\Gamma}_2)$ 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 the proof of Proposition 3.

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 (18) that

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

We also know that

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

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

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

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.

E 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

$$\mathbf{C}_k = \begin{Bmatrix} \boldsymbol{\Gamma}_k^{-1} \\ \mathbf{D}_M'[\boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \otimes \boldsymbol{\Gamma}_k^{-1}] \\ [\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k)\boldsymbol{\Gamma}_k^{-1/2} \otimes \boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M \\ \frac{1}{2}\mathbf{D}_M'[\{\boldsymbol{\Gamma}_k^{-1} + 2[\boldsymbol{\Gamma}_k^{-1/2\prime}[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M]\boldsymbol{\Gamma}_k^{-1/2}\} \otimes \boldsymbol{\Gamma}_k^{-1}]\mathbf{D}_M \end{Bmatrix}$$

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

$$\frac{\partial \ln f(\mathbf{y}; \boldsymbol{\phi})}{\partial \boldsymbol{\pi}} = \sum_{k=1}^{K-1} \frac{w_k(\boldsymbol{\phi})}{\pi_k}\mathbf{e}_k - \frac{w_K(\boldsymbol{\phi})}{1 - \sum_{k=1}^{K-1} \pi_k}\boldsymbol{\iota}_{K-1} \tag{E32}$$

and

$$\frac{\partial \ln f(\mathbf{y}; \boldsymbol{\phi})}{\partial \boldsymbol{\theta}_k} = w_k(\boldsymbol{\phi}) \begin{Bmatrix} \boldsymbol{\Gamma}_k^{-1/2\prime}\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k) \\ \frac{1}{2}\mathbf{D}_M'(\boldsymbol{\Gamma}_k^{-1/2\prime} \otimes \boldsymbol{\Gamma}_k^{-1/2\prime})vec[\boldsymbol{\varepsilon}^*(\boldsymbol{\theta}_k)\boldsymbol{\varepsilon}^{*\prime}(\boldsymbol{\theta}_k) - \mathbf{I}_M] \end{Bmatrix}. \tag{E33}$$

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 $k \neq l$.

On this basis, they show that the sum of the Hessian and the outer product of the scores corresponding to a single observation is given by the matrix

$$\mathbf{W} = \begin{pmatrix} \mathbf{0} & \mathbf{A}_1' & \cdots & \mathbf{A}_K' \\ \mathbf{A}_1 & \mathbf{W}_1 & & \mathbf{0} \\ \vdots & & \ddots & \\ \mathbf{A}_K & \mathbf{0} & & \mathbf{W}_K \end{pmatrix}$$

where $\mathbf{A}_k$ is a $\frac{1}{2}M\,(M+3)\times(K-1)$ matrix whose k^{th} column contains the following equations

$$\left[\begin{array}{c} (23) \\ (27) \end{array}\right]$$

while $\mathbf{W}_k$ is a square matrix of order $\frac{1}{2}M\,(M+3)$ consisting of the following ones

$$\left[\begin{array}{cc} (23) & (24) \\ (24)' & (26) \end{array}\right]$$

They also note that $\mathbf{A}_k$ is 0 when evaluated at the MLE. These expressions differ slightly from

ours because they work with $\boldsymbol{\pi}$ rather than $\boldsymbol{\lambda}$. Nevertheless, given that

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_{K-1} \\ \lambda_K \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \\ -1 & \dots & -1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_{K-1} \end{pmatrix} = \mathbf{e}_K + \begin{pmatrix} \mathbf{I}_{K-1} \\ -\boldsymbol{\iota}'_{K-1} \end{pmatrix} \boldsymbol{\pi}, \quad \text{(E34)}$$

so that

$$\frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\pi}'} = \begin{pmatrix} \mathbf{I}_{K-1} \\ -\boldsymbol{\iota}'_{K-1} \end{pmatrix},$$

it is easy to see that

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

coincides with (E32).

Similarly, given that (E34) is affine, so that its second Jacobian is 0, it follows that

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

It is tedious but straightforward to show that analogous calculations applied to the other terms we have derived in the proof of Proposition 3 also coincide with 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 (20), the Jacobian would be instead

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

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
6,400	16.07	9.77	4.06	10.45	5.56	1.33
25,600	11.99	6.63	1.85	9.82	5.10	1.10
102,400	10.62	5.55	1.26	9.99	4.98	1.04

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 3 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 (2024), 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 $\hat{\boldsymbol{\phi}}_N$. 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
6,400	17.55	11.15	4.38	9.91	5.40	1.48
25,600	12.26	7.28	2.32	9.90	5.15	1.15
102,400	10.68	5.47	1.43	9.84	4.93	1.09

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 4 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 (2024), 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 $\hat{\boldsymbol{\phi}}_N$. 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
6,400	21.50	14.12	5.90	10.37	5.47	1.35
25,600	14.19	7.91	2.07	10.68	5.31	1.35
102,400	10.79	5.84	1.36	10.11	5.32	1.10

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 3 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 (2024), 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 $\widehat{\boldsymbol{\phi}}_N$. 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 in %)	
Sample	ν_1	ν_2	ν_3	$\sqrt{\gamma_1}$	$\sqrt{\gamma_2}$	$\sqrt{\gamma_3}$	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	67.56	43.81
1965	2.84	1.01	0.28	1.05	0.39	0.11	0.27	0.40	0.33	62.52	36.88
1970	2.74	0.96	0.27	0.96	0.40	0.10	0.31	0.37	0.33	34.39	13.19
1975	3.08	1.07	0.26	0.65	0.47	0.10	0.24	0.45	0.31	47.10	24.20
1980	2.87	1.08	0.26	0.68	0.40	0.12	0.28	0.38	0.34	74.28	49.12
1985	2.86	0.92	0.20	0.69	0.43	0.07	0.27	0.49	0.24	40.13	20.37
1990	3.12	0.93	0.18	0.56	0.48	0.05	0.24	0.52	0.24	54.74	38.58
1995	3.02	0.89	0.15	0.49	0.48	0.05	0.25	0.50	0.25	70.35	56.49
2000	2.93	0.82	0.15	0.59	0.44	0.05	0.28	0.48	0.24	50.76	34.84

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 $\hat{\boldsymbol{\phi}}_N$. 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.

Table 5: Specification testing for log GDP per capita and log CO2 emissions per capita

	Panel A: Parameter estimates												Panel B: IM test (p-values in %)	
Sample	$\nu_{1,1}$	$\nu_{1,2}$	$\nu_{2,1}$	$\nu_{2,2}$	$\gamma_{1,1}$	$\gamma_{1,2}$	$\gamma_{1,3}$	$\gamma_{2,1}$	$\gamma_{2,2}$	$\gamma_{2,3}$	λ_1	λ_2	Asymptotic	Bootstrap
1960	7.58	-2.23	9.17	0.28	0.78	0.71	1.55	0.16	0.28	0.70	0.73	0.27	7.69	6.67
1965	7.75	-1.95	9.44	0.62	0.86	0.86	1.64	0.10	0.15	0.43	0.75	0.25	43.21	21.60
1970	7.96	-1.47	9.70	0.95	0.98	1.06	1.74	0.04	0.07	0.30	0.79	0.21	0.22	2.72
1975	8.05	-1.34	9.79	0.92	1.02	1.13	1.83	0.03	0.05	0.19	0.77	0.23	13.69	8.93
1980	8.21	-1.16	9.90	1.03	1.13	1.29	2.00	0.03	0.06	0.23	0.79	0.21	2.17	4.69
1985	8.08	-1.33	10.01	0.83	0.89	1.18	2.03	0.05	0.04	0.19	0.73	0.27	42.59	20.91
1990	8.04	-1.33	10.11	0.86	0.86	1.15	2.02	0.06	0.05	0.18	0.70	0.30	41.24	21.30
1995	8.07	-1.25	10.17	0.86	0.96	1.22	2.00	0.06	0.04	0.16	0.68	0.32	71.73	42.48
2000	8.14	-1.23	10.33	0.91	1.00	1.27	1.99	0.06	0.04	0.13	0.68	0.32	56.75	29.85
2005	8.34	-1.03	10.41	0.89	1.22	1.52	2.21	0.03	0.02	0.10	0.72	0.28	41.50	20.96

Notes: The sample includes 84 countries and covers the period 1960–2005. Per capita income from the Penn World Table 7.1 and data fossil fuel-based CO_2 emissions come from Marland et al. (2003). 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 $\hat{\boldsymbol{\phi}}_N$. 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

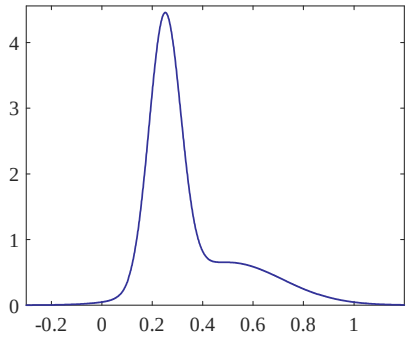


Fig. 1e: Mixture of three normals

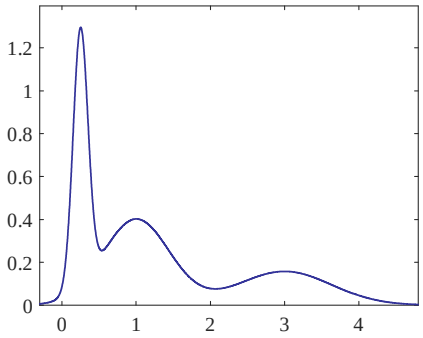


Fig. 1b: Mixture of two asymmetric t's

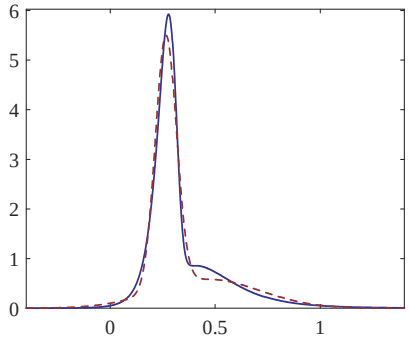


Fig. 1f: Mixture of three asymmetric t's

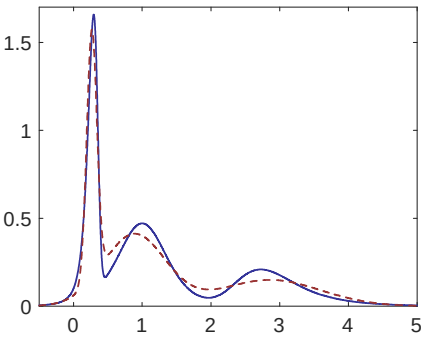


Fig. 1c: Symmetric mixture of three normals

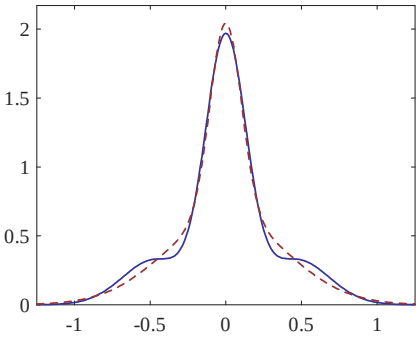


Fig. 1g: Mixture of four normals

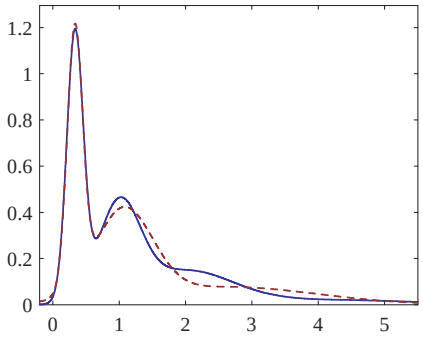


Fig. 1d: Lognormal

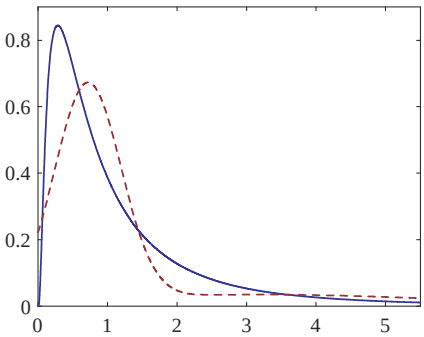
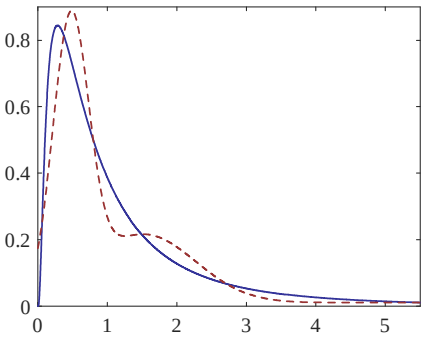


Fig. 1h: Lognormal



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

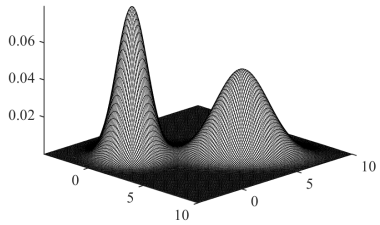


Fig. 2b: Density of a bivariate mixture of two asymmetric t's

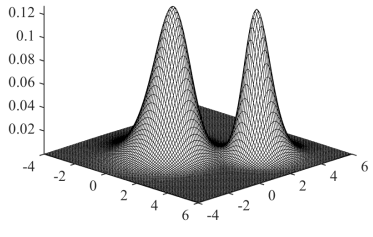


Fig. 2c: Density of a bivariate mixture of three normals

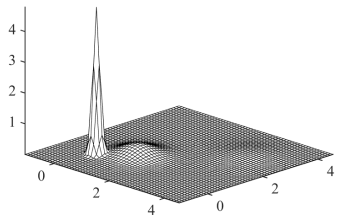


Fig. 2d: Density of a bivariate skew normal

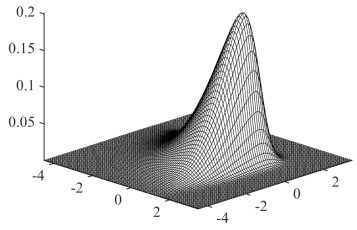


Fig. 2e: Contours of a bivariate mixture of two normals

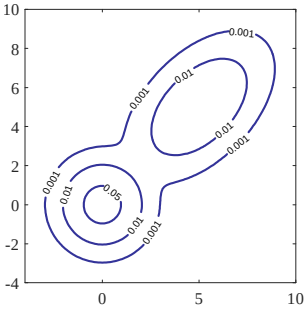


Fig. 2f: Contours of a bivariate mixture of two asymmetric t's

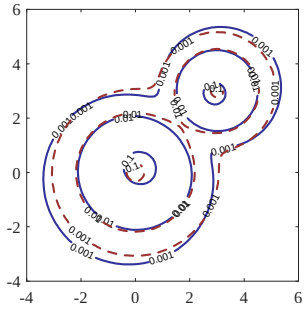


Fig. 2g: Contours of a bivariate mixture of three normals

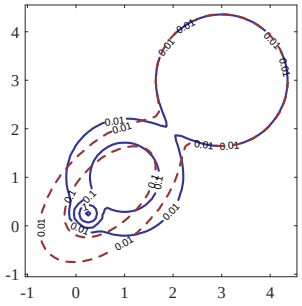
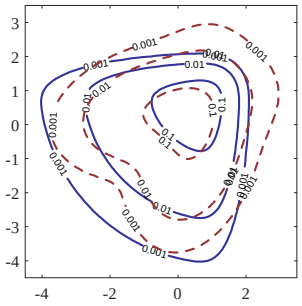


Fig. 2h: Contours of a bivariate skew normal



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

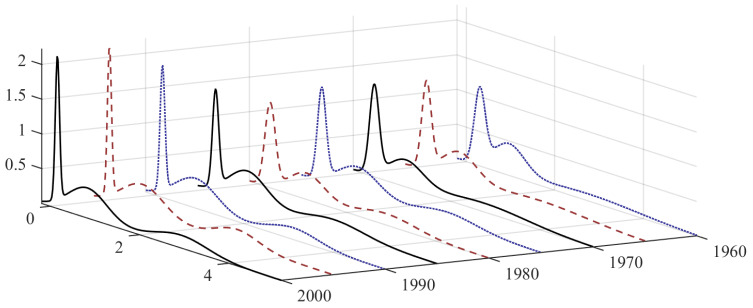


Fig. 3b: 1960, 1965, 1970

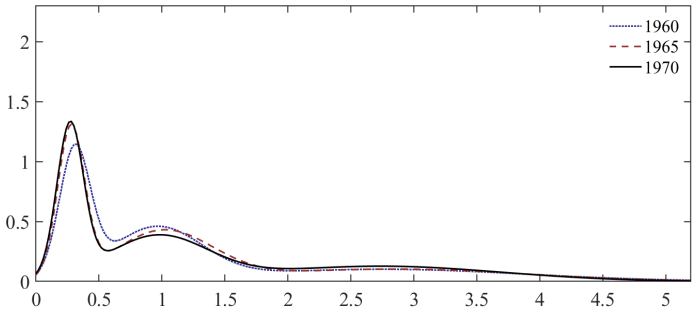


Fig. 3c: 1975, 1980, 1985

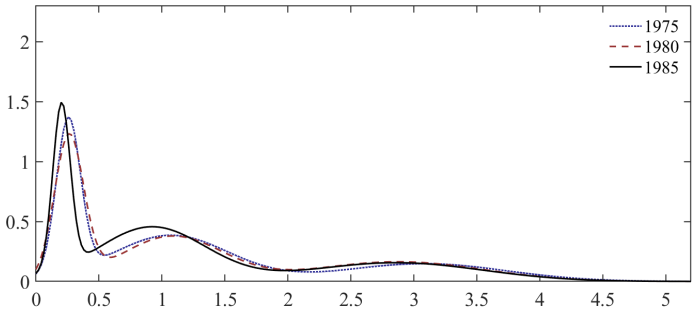
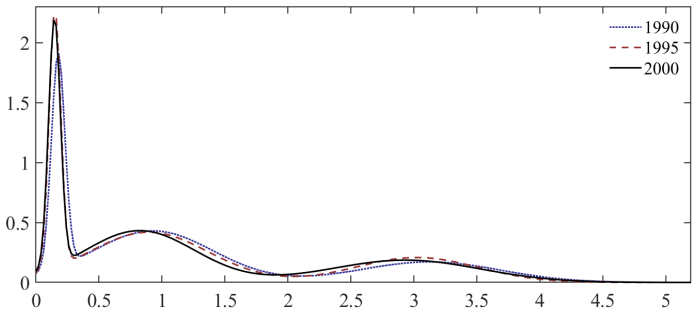
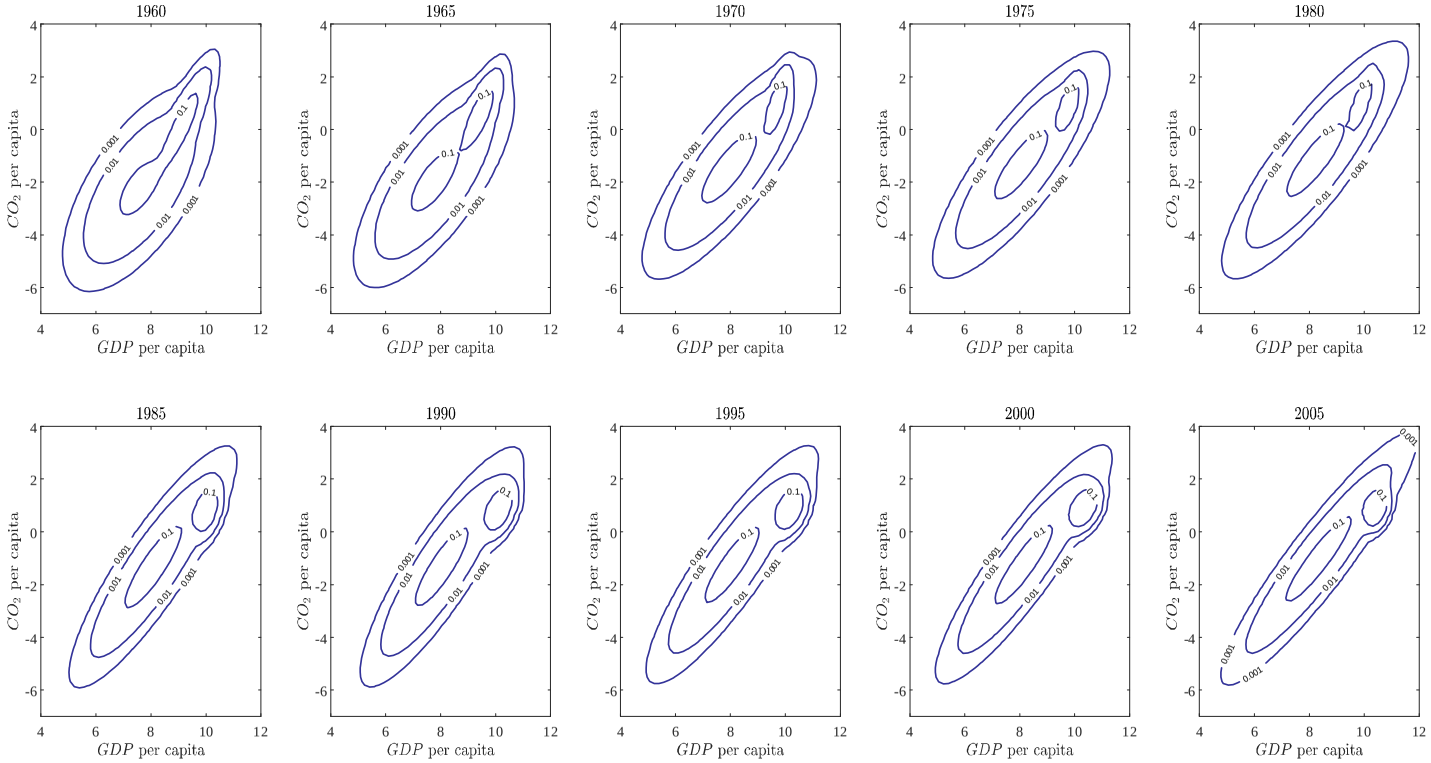


Fig. 3d: 1990, 1995, 2000



Notes: Data: Per capita income from version 6.1 of the Penn World Tables; pdfs of the closest mixture (using MLE estimates) of three univariate normals.

Figure 4: Bivariate analysis for log GDP per capita and log CO_2 emissions per capita



Notes: The sample includes 84 countries and covers the period 1960–2005. Per capita income from the Penn World Table 7.1 and data fossil fuel-based CO_2 emissions come from Marland et al. (2003); contours of the closest mixture (using MLE estimates) of two bivariate normals.