Maximum Likelihood Estimation of the Multivariate Normal Mixture Model

Otilia BOLDEA and Jan R. MAGNUS

The Hessian of the multivariate normal mixture model is derived, and estimators of the information matrix are obtained, thus enabling consistent estimation of all parameters and their precisions. The usefulness of the new theory is illustrated with two examples and some simulation experiments. The newly proposed estimators appear to be superior to the existing ones.

KEY WORDS: Information matrix; Maximum likelihood; Mixture model.

1. INTRODUCTION

In finite mixture models it is assumed that data are obtained from a finite collection of populations and that the data within each population follow a standard distribution, typically normal, Poisson, or binomial. Such models are particularly useful when the data come from multiple sources, and they find application in such varied fields as criminology, engineering, demography, economics, psychology, marketing, sociology, plant pathology, and epidemiology.

The normal (Gaussian) model has received the most attention. Here we consider an m-dimensional random vector $\mathbf{x}$ whose distribution is a mixture (weighted average) of g normal densities, so that

$$f(\mathbf{x}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{x}), \tag{1}$$

where

$$f_i(\mathbf{x}) = (2\pi)^{-m/2} |\mathbf{V}_i|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

and the π_i are weights satisfying $\pi_i > 0$ and $\sum_i \pi_i = 1$. This is the so-called "multivariate normal mixture model." The parameters of the model are $(\pi_i, \boldsymbol{\mu}_i, \mathbf{V}_i)$ for $i = 1, \ldots, g$ subject to two constraints, namely that the π_i sum to one and that the $\mathbf{V}_i$ are symmetric (in fact, positive definite).

The origin of mixture models is usually attributed to Newcomb (1886) and Pearson (1894), although some 50 years earlier Poisson already used mixtures to analyze conviction rates; see Stigler (1986). But it was only after the introduction of the EM algorithm by Dempster, Laird, and Rubin (1977) that mixture models have gained wide popularity in applied statistics. Since then an extensive literature has developed. Important reviews are given in Titterington, Smith, and Makov (1985), McLachlan and Basford (1988), and McLachlan and Peel (2000).

There are two theoretical problems with mixtures. First, as noted by Day (1969) and Hathaway (1985), the likelihood may be unbounded in which case the maximum likelihood (ML) estimator does not exist. However, we can still determine a sequence of roots of the likelihood equation that is consistent and asymptotically efficient; see McLachlan and Basford (1988, section 1.8). Hence, this is not necessarily a problem in practice. Second, the parameters are not identified unless we impose an additional restriction, such as

$$\pi_1 \geq \pi_2 \geq \cdots \geq \pi_g;$$

see Titterington, Smith, and Makov (1985, section 3.1). This is not a problem in practice either, and we follow Aitken and Rubin (1985) by imposing the restriction but carrying out the ML estimation without it.

The task of estimating the parameters and their precisions, and formulating confidence intervals and test statistics, is difficult and tedious. This is simply because in standard situations with independent and identically distributed observations, the likelihood contains products and therefore the log-likelihood contains sums. But here the likelihood itself is a sum, and therefore the derivatives of the log-likelihood will contain ratios. Taking expectations is therefore typically not feasible. Even the task of obtaining the derivatives of the log-likelihood (score and Hessian matrix) is not trivial.

Currently there are several methods to estimate the variance matrix of the ML estimator in (multivariate) mixture models in terms of the inverse of the observed information matrix, and they differ by the way this inverse is approximated. One method involves using the "complete-data" log-likelihood, that is, the log-likelihood of an augmented data problem, where the assignment of each observation to a mixture component is an unobserved variable coming from a prespecified multinomial distribution. The advantage of using the complete-data log-likelihood instead of the incomplete-data (the original data) log-likelihood lies in its form as a sum of logarithms rather than a logarithm of a sum. The information matrix for the incomplete data can be shown to depend only on the conditional moments of the gradient and curvature of the complete-data log-likelihood function and so can be readily computed; see Louis (1982). Another method, in the context of the original log-likelihood, was proposed by Dietz and Böhning (1996), exploiting the fact that in large samples from regular models, twice the change in log-likelihood on omitting that variable is equal to the square of

Otilia Boldea is Assistant Professor (E-mail: *o.boldea@uvt.nl*) and Jan R. Magnus is Professor of Econometrics (E-mail: *magnus@uvt.nl*), both at the Department of Econometrics & OR, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands. The authors are grateful to Hamparsum Bozdogan for asking a question which led to this paper, to John Einmahl and Evangelos Evangelou for useful comments, to Geoffrey McLachlan for providing his EMMIX FORTRAN code free of charge and for his prompt response to our questions, and to the editor, associate editor, and the referees for helpful comments. The first version of this paper was written during a visit of one of us to the Wang Yanan Institute for Studies in Economics (WISE), Xiamen University, China.

© 2009 American Statistical Association
Journal of the American Statistical Association
December 2009, Vol. 104, No. 488, Theory and Methods
DOI: 10.1198/jasa.2009.tm08273

the t-statistic of that variable; see McLachlan and Peel (2000, p. 68). This method was extended by Liu (1998) to multivariate models. There is also a conditional bootstrap approach described in McLachlan and Peel (2000, p. 67).

In addition, the standard errors of the ML estimator can be computed by at least three bootstrap methods: the parametric bootstrap (Basford et al. 1997; McLachlan and Peel 2000), the nonparametric bootstrap (McLachlan and Peel 2000) which is an extension of Efron (1979), and the weighted bootstrap (Newton and Raftery 1994) which is a version of the nonparametric bootstrap based on scaling the data with weights that are proportional to the number of times an original point occurs in the bootstrap sample. Basford et al. (1997) compare the parametric bootstrap with a method based on the outer product of the scores as a proxy for the observed information matrix, and find simulation evidence that the bootstrap-based standard errors are more reliable in small samples.

In this paper we explicitly derive the score and Hessian matrix for the multivariate normal mixture model, and use the results to estimate the information matrix. This provides a twofold extension of Behboodian (1972) and Ali and Nadarajah (2007), who study the information matrix for the case of a mixture of two (rather then g) univariate (rather than multivariate) normal distributions. Since we work with the original ("incomplete") log-likelihood, we compare our information-based standard errors to the bootstrap-based standard errors which are the natural small-sample counterpart.

We find that in correctly specified models the method based on the observed Hessian-based information matrix is the best in terms of root mean squared error. In misspecified models the method based on the observed "sandwich" matrix is the best.

This paper is organized as follows. In Section 2 we discuss how to take account of the two constraints: symmetry of the variance matrices and the fact that the weights sum to one. Our general result (Theorem 1) is formulated in Section 3, where we also discuss the estimation of the variance of the ML estimator and introduce the misspecification-robust "sandwich" matrix. These results allow us to formally test for misspecification using the information matrix test (Theorem 2), discussed in Section 4. In Section 5 we present the important special case (Theorem 3) where all variance matrices are equal. In Section 6 we study two well-known examples based on the hemophilia dataset and the Iris dataset. These examples demonstrate that our formulae can be implemented without any problems and that the results are credible. But these examples do not yet prove that the information-based estimates of the standard errors are more accurate than the ones currently in use. Therefore we provide Monte Carlo evidence in Section 7. Section 8 concludes. An Appendix contains proofs of the three theorems.

2. SYMMETRY AND WEIGHT CONSTRAINTS

Before we derive the score vector and the Hessian matrix, we need to discuss two constraints that play a role in mixture models: symmetry of the variance matrices and the fact that the weights sum to one. To deal with the symmetry constraint we introduce the half-vec operator vech$(\cdot)$ and the duplication matrix $\mathbf{D}$; see Magnus and Neudecker (1988) and Magnus (1988). Let $\mathbf{V}$ be a symmetric $m \times m$ matrix, and let vech $\mathbf{V}$ denote the $\frac{1}{2}m(m+1) \times 1$ vector that is obtained from vec $\mathbf{V}$ by eliminating all supradiagonal elements of $\mathbf{V}$. Then the elements of vec $\mathbf{V}$ are those of vech $\mathbf{V}$ with some repetitions. Hence, there exists a unique $m^2 \times \frac{1}{2}m(m+1)$ matrix $\mathbf{D}$, such that $\mathbf{D}$ vech $\mathbf{V} =$ vec $\mathbf{V}$. Since the elements of $\mathbf{V}$ are constrained by the symmetry, we must differentiate with respect to vech $\mathbf{V}$ and not with respect to vec $\mathbf{V}$.

The weights π_i must all be positive and they must sum to one. We maximize with respect to $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{g-1})'$ and set $\pi_g = 1 - \pi_1 - \cdots - \pi_{g-1}$. We have

$$d \log \pi_i = \mathbf{a}_i' d\boldsymbol{\pi}, \qquad d^2 \log \pi_i = -(d\boldsymbol{\pi})' \mathbf{a}_i \mathbf{a}_i' (d\boldsymbol{\pi}), \tag{3}$$

where

$$\begin{aligned} \mathbf{a}_i &= (1/\pi_i)\mathbf{e}_i \qquad (i = 1, \ldots, p-1), \\ \mathbf{a}_g &= -(1/\pi_g)\boldsymbol{\iota}, \end{aligned} \tag{4}$$

$\mathbf{e}_i$ denotes the ith column of the identity matrix $\mathbf{I}_{g-1}$, and $\boldsymbol{\iota}$ is the $(g-1)$-dimensional vector of ones. The model parameters are then $\boldsymbol{\pi}$ and, for $i = 1, \ldots, g$, $\boldsymbol{\mu}_i$ and vech $\mathbf{V}_i$. Writing

$$\boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \text{vech } \mathbf{V}_i \end{pmatrix},$$

the complete parameter vector can be expressed as $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_g')'$.

3. SCORE VECTOR, HESSIAN, AND VARIANCE MATRIX

Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of independent and identically distributed random variables from the distribution (1), we write the log-likelihood as

$$L(\boldsymbol{\theta}) = \sum_{t=1}^{n} \log f(\mathbf{x}_t).$$

The score vector is defined by $\mathbf{q}(\boldsymbol{\theta}) = \sum_t \mathbf{q}_t(\boldsymbol{\theta})$, where

$$\mathbf{q}_t(\boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{x}_t)}{\partial \boldsymbol{\theta}} = \text{vec}(\mathbf{q}_t^{\pi}, \mathbf{q}_t^{1}, \ldots, \mathbf{q}_t^{g}),$$

and the Hessian matrix by $\mathbf{Q}(\boldsymbol{\theta}) = \sum_t \mathbf{Q}_t(\boldsymbol{\theta})$, where

$$\mathbf{Q}_t(\boldsymbol{\theta}) = \frac{\partial^2 \log f(\mathbf{x}_t)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = \begin{pmatrix} \mathbf{Q}_t^{\pi\pi} & \mathbf{Q}_t^{\pi 1} & \cdots & \mathbf{Q}_t^{\pi g} \\ \mathbf{Q}_t^{1\pi} & \mathbf{Q}_t^{11} & \cdots & \mathbf{Q}_t^{1g} \\ \vdots & \vdots & & \vdots \\ \mathbf{Q}_t^{g\pi} & \mathbf{Q}_t^{g1} & \cdots & \mathbf{Q}_t^{gg} \end{pmatrix}.$$

Before we can state our main result we need some more notation. We define

$$\phi_{it} = \pi_i f_i(\mathbf{x}_t), \qquad \alpha_{it} = \frac{\phi_{it}}{\sum_j \phi_{jt}}, \tag{5}$$

$$\mathbf{b}_{it} = \mathbf{V}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i), \qquad \mathbf{B}_{it} = \mathbf{V}_i^{-1} - \mathbf{b}_{it}\mathbf{b}_{it}', \tag{6}$$

$$\mathbf{c}_{it} = \begin{pmatrix} \mathbf{b}_{it} \\ -\frac{1}{2}\mathbf{D}' \text{vec } \mathbf{B}_{it} \end{pmatrix}, \tag{7}$$

and

$$\mathbf{C}_{it} = \begin{pmatrix} \mathbf{V}_i^{-1} & (\mathbf{b}_{it}' \otimes \mathbf{V}_i^{-1})\mathbf{D} \\ \mathbf{D}'(\mathbf{b}_{it} \otimes \mathbf{V}_i^{-1}) & \frac{1}{2}\mathbf{D}'((\mathbf{V}_i^{-1} - 2\mathbf{B}_{it}) \otimes \mathbf{V}_i^{-1})\mathbf{D} \end{pmatrix}. \tag{8}$$

We also recall that $\mathbf{a}_i$ is defined in (4) and we let $\bar{\mathbf{a}}_t = \sum_i \alpha_{it}\mathbf{a}_i$. We can now state Theorem 1, which allows direct calculation of the score and Hessian matrix.

Theorem 1. The contribution of the tth observation to the score vector with respect to the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_i$ ($i = 1, \ldots, g$) is given by

$$\mathbf{q}_t^{\pi} = \bar{\mathbf{a}}_t, \qquad \mathbf{q}_t^{i} = \alpha_{it}\mathbf{c}_{it},$$

and the contribution of the tth observation to the Hessian matrix is

$$\mathbf{Q}_t^{\pi\pi} = -\bar{\mathbf{a}}_t\bar{\mathbf{a}}_t', \qquad \mathbf{Q}_t^{\pi i} = \alpha_{it}(\mathbf{a}_i - \bar{\mathbf{a}}_t)\mathbf{c}_{it}',$$

and

$$\mathbf{Q}_t^{ii} = -(\alpha_{it}\mathbf{C}_{it} - \alpha_{it}(1-\alpha_{it})\mathbf{c}_{it}\mathbf{c}_{it}'),$$
$$\mathbf{Q}_t^{ij} = -\alpha_{it}\alpha_{jt}\mathbf{c}_{it}\mathbf{c}_{jt}' \qquad (i \neq j).$$

We note that the expressions for the score in Theorem 1 are the same as in Basford et al. (1997). The expressions for the Hessian are new.

We next discuss the estimation of the variance of $\hat{\boldsymbol{\theta}}$. In maximum likelihood theory the variance is usually obtained from the information matrix. If the model is correctly specified, then the information matrix is defined by

$$\mathcal{I} = -\mathrm{E}(\mathbf{Q}) = \mathrm{E}(\mathbf{q}\mathbf{q}'),$$

where the equality holds because of second-order regularity. In our case we cannot obtain these expectations analytically. Moreover, we cannot be certain that the model is correctly specified. We estimate the information matrix by

$$\mathcal{I}_1 = \sum_{t=1}^{n} \mathbf{q}_t(\hat{\boldsymbol{\theta}})\mathbf{q}_t(\hat{\boldsymbol{\theta}})',$$

based on first-order derivatives, or by

$$\mathcal{I}_2 = -\mathbf{Q}(\hat{\boldsymbol{\theta}}) = -\sum_{t=1}^{n} \mathbf{Q}_t(\hat{\boldsymbol{\theta}}),$$

based on second-order derivatives. The inverses $\mathcal{I}_1^{-1}$ and $\mathcal{I}_2^{-1}$ are consistent estimators of the asymptotic variance of $\hat{\boldsymbol{\theta}}$ if the model is correctly specified. In general, the "sandwich" (or "robust") variance matrix

$$\mathcal{I}_3^{-1} = \widehat{\mathrm{var}}(\hat{\boldsymbol{\theta}}) = \mathcal{I}_2^{-1}\mathcal{I}_1\mathcal{I}_2^{-1} \tag{9}$$

provides a consistent estimator of the variance matrix, whether or not the model is not correctly specified. This was noted by Huber (1967), White (1982), and others, and is based on the realization that the asymptotic normality of $\hat{\boldsymbol{\theta}}$ rests on the facts that the expected value of $(1/n)\mathbf{q}(\boldsymbol{\theta})\mathbf{q}(\boldsymbol{\theta})'$ has a finite positive semidefinite (possibly singular) limit, say $\mathcal{I}_1^{\infty}$, and that $-(1/n)\mathbf{Q}(\boldsymbol{\theta})$ converges in probability to a positive definite matrix, say $\mathcal{I}_2^{\infty}$, and that these two limiting matrices need not be equal; see also Davidson and MacKinnon (2004, pp. 416–417).

We note in passing an important and somewhat counterintuitive property of the sandwich estimator, which is seldom mentioned. If $\mathcal{I}_1 = \mathcal{I}_2$, then $\mathcal{I}_1 = \mathcal{I}_2 = \mathcal{I}_3$. If $\mathcal{I}_1 \neq \mathcal{I}_2$, then one would perhaps expect that $\mathcal{I}_3^{-1}$ lies "in between" $\mathcal{I}_1^{-1}$ and $\mathcal{I}_2^{-1}$, but this is typically not the case, as is easily demonstrated. Let $\boldsymbol{\Psi} = \mathcal{I}_1^{-1} - \mathcal{I}_2^{-1}$. Then,

$$\mathcal{I}_3^{-1} = \mathcal{I}_2^{-1}\mathcal{I}_1\mathcal{I}_2^{-1} = \mathcal{I}_2^{-1}(\mathcal{I}_2^{-1} + \boldsymbol{\Psi})^{-1}\mathcal{I}_2^{-1}$$
$$= (\mathcal{I}_2 + \mathcal{I}_2\boldsymbol{\Psi}\mathcal{I}_2)^{-1}.$$

If $\boldsymbol{\Psi}$ is positive definite ($\mathcal{I}_2^{-1} < \mathcal{I}_1^{-1}$) then $\mathcal{I}_3^{-1} < \mathcal{I}_2^{-1} < \mathcal{I}_1^{-1}$; if $\boldsymbol{\Psi}$ is negative definite ($\mathcal{I}_2^{-1} > \mathcal{I}_1^{-1}$) then $\mathcal{I}_3^{-1} > \mathcal{I}_2^{-1} > \mathcal{I}_1^{-1}$. In practice there is no reason why $\boldsymbol{\Psi}$ should be either positive definite or negative definite. Nevertheless, we should expect an individual variance based on the Hessian to lie in between the variance based on the score and the variance based on the robust estimator, and this expectation is confirmed by the simulation results in Section 7.

4. INFORMATION MATRIX TEST

The information matrix (IM) test, introduced by White (1982), is well known as a general test for misspecification of a parametric likelihood function. Despite the fact that the asymptotic distribution is a poor approximation to the finite-sample distribution of the test statistic, the IM test has established itself in the econometrics profession. Below we obtain the IM test for mixture models. Let us define

$$\mathbf{W}_t(\boldsymbol{\theta}) = \mathbf{Q}_t(\boldsymbol{\theta}) + \mathbf{q}_t(\boldsymbol{\theta})\mathbf{q}_t(\boldsymbol{\theta})'.$$

From Theorem 1 we see that

$$\mathbf{W}_t(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{0} & \mathbf{a}_1(\mathbf{q}_t^1)' & \mathbf{a}_2(\mathbf{q}_t^2)' & \cdots & \mathbf{a}_g(\mathbf{q}_t^g)' \\ \mathbf{q}_t^1\mathbf{a}_1' & \mathbf{W}_t^1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{q}_t^2\mathbf{a}_2' & \mathbf{0} & \mathbf{W}_t^2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathbf{q}_t^g\mathbf{a}_g' & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{W}_t^g \end{pmatrix},$$

where $\mathbf{a}_i$ and $\mathbf{q}_t^i$ have been defined before, and

$$\mathbf{W}_t^i = -\alpha_{it}(\mathbf{C}_{it} - \mathbf{c}_{it}\mathbf{c}_{it}') = -\alpha_{it}\begin{pmatrix} \mathbf{B}_{it} & \boldsymbol{\Gamma}_{it}'\mathbf{D} \\ \mathbf{D}'\boldsymbol{\Gamma}_{it} & \mathbf{D}'\boldsymbol{\Delta}_{it}\mathbf{D} \end{pmatrix}$$

with

$$\boldsymbol{\Gamma}_{it} = \mathbf{b}_{it} \otimes \mathbf{V}_i^{-1} + (1/2)(\mathrm{vec}\,\mathbf{B}_{it})\mathbf{b}_{it}'$$

representing skewness, and

$$\boldsymbol{\Delta}_{it} = (1/2)(\mathbf{V}_i^{-1} \otimes \mathbf{V}_i^{-1}) - \mathbf{B}_{it} \otimes \mathbf{V}_i^{-1} - (1/4)(\mathrm{vec}\,\mathbf{B}_{it})(\mathrm{vec}\,\mathbf{B}_{it})'$$

representing kurtosis. The purpose of the information matrix procedure is to test for the joint significance of the nonredundant elements of the matrix $\mathbf{W}(\hat{\boldsymbol{\theta}}) = \sum_t \mathbf{W}_t(\hat{\boldsymbol{\theta}})$. Now, since $\mathbf{q}(\hat{\boldsymbol{\theta}}) = \sum_t \mathbf{q}_t(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, the IM procedure in our case tests for the joint significance of the nonredundant elements of $\sum_t \mathbf{W}_t^i(\hat{\boldsymbol{\theta}})$ for $i = 1, \ldots, g$.

Following Chesher (1983) and Lancaster (1984) we formulate the White's (1982) IM test as follows.

Theorem 2 (Information matrix test). Define the variance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_t\mathbf{w}_t' - \left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{w}_t\mathbf{q}_t'\right)\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{q}_t\mathbf{q}_t'\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} \mathbf{q}_t\mathbf{w}_t'\right),$$

where $\mathbf{q}_t$ denotes the tth increment to the score, and

$$\mathbf{w}_t = \mathrm{vec}(\mathrm{vech}\,\mathbf{W}_t^1, \mathrm{vech}\,\mathbf{W}_t^2, \ldots, \mathrm{vech}\,\mathbf{W}_t^g).$$

Then, evaluated at $\hat{\boldsymbol{\theta}}$ and under the null hypothesis of correct specification,

$$\mathrm{IM} = n\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{w}_t\right)'\boldsymbol{\Sigma}^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{w}_t\right)$$

asymptotically follows a χ^2-distribution with $gm(m+3)/2$ degrees of freedom.

The above form of the IM test is a variant of the outer-product-of-the-gradient (OPG) regression, often used to calculate Lagrange multiplier tests. Such tests are known to reject true null hypotheses far too often in finite samples, and this is also true for the OPG form of the IM test. We illustrate this fact through some simulations at the end of Section 7. To use the asymptotic critical values is not a good idea. Instead, these values can be bootstrapped; see Horowitz (1994) and Davidson and MacKinnon (2004, section 16.9) for details and references.

5. SPECIAL CASE: EQUAL VARIANCE MATRICES

There are many important special cases of Theorem 1. We may encounter cases where the weights π_i are known or where the means $\boldsymbol{\mu}_i$ are equal across different mixtures. The most important special case, however, is the one where the variances $\mathbf{V}_i$ are equal: $\mathbf{V}_i = \mathbf{V}$. This is the case presented in Theorem 3. Further specialization is of course possible: $\mathbf{V}$ could be diagonal or even proportional to the identity matrix, but we do not exploit these cases here.

When $\mathbf{V}_i = \mathbf{V}$, we write the parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_g', \mathbf{v}')'$, where $\mathbf{v} = \operatorname{vech}\mathbf{V}$. The score is $\mathbf{q}(\boldsymbol{\theta}) = \sum_t \mathbf{q}_t(\boldsymbol{\theta})$ with

$$\mathbf{q}_t(\boldsymbol{\theta}) = \operatorname{vec}(\mathbf{q}_t^{\pi}, \mathbf{q}_t^{1}, \ldots, \mathbf{q}_t^{g}, \mathbf{q}_t^{v}),$$

and the Hessian matrix is $\mathbf{Q}(\boldsymbol{\theta}) = \sum_t \mathbf{Q}_t(\boldsymbol{\theta})$ with

$$\mathbf{Q}_t(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{Q}_t^{\pi\pi} & \mathbf{Q}_t^{\pi 1} & \cdots & \mathbf{Q}_t^{\pi g} & \mathbf{Q}_t^{\pi v} \\ \mathbf{Q}_t^{1\pi} & \mathbf{Q}_t^{11} & \cdots & \mathbf{Q}_t^{1g} & \mathbf{Q}_t^{1v} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{Q}_t^{g\pi} & \mathbf{Q}_t^{g1} & \cdots & \mathbf{Q}_t^{gg} & \mathbf{Q}_t^{gv} \\ \mathbf{Q}_t^{v\pi} & \mathbf{Q}_t^{v1} & \cdots & \mathbf{Q}_t^{vg} & \mathbf{Q}_t^{vv} \end{pmatrix}.$$

Theorem 3 ($\mathbf{V}_i = \mathbf{V}$). The contribution of the tth observation to the score vector with respect to the parameters $\boldsymbol{\pi}$, $\boldsymbol{\mu}_i$ ($i = 1, \ldots, g$), and $\mathbf{v}$ is given by

$$\mathbf{q}_t^{\pi} = \bar{\mathbf{a}}_t, \qquad \mathbf{q}_t^{i} = \alpha_{it}\mathbf{b}_{it}, \qquad \mathbf{q}_t^{v} = -\frac{1}{2}\mathbf{D}'\operatorname{vec}\bar{\mathbf{B}}_t,$$

where

$$\bar{\mathbf{B}}_t = \mathbf{V}^{-1} - \sum_{i=1}^{g}\alpha_{it}\mathbf{b}_{it}\mathbf{b}_{it}',$$

and the contribution of the tth observation to the Hessian matrix is

$$\mathbf{Q}_t^{\pi\pi} = -\bar{\mathbf{a}}_t\bar{\mathbf{a}}_t', \qquad \mathbf{Q}_t^{\pi i} = \alpha_{it}(\mathbf{a}_i - \bar{\mathbf{a}}_t)\mathbf{b}_{it}',$$

$$\mathbf{Q}_t^{\pi v} = -\frac{1}{2}\sum_{i=1}^{g}\alpha_{it}(\mathbf{a}_i - \bar{\mathbf{a}}_t)(\operatorname{vec}\mathbf{B}_{it})'\mathbf{D},$$

$$\mathbf{Q}_t^{ii} = -\alpha_{it}\mathbf{V}^{-1} + \alpha_{it}(1-\alpha_{it})\mathbf{b}_{it}\mathbf{b}_{it}',$$

$$\mathbf{Q}_t^{ij} = -\alpha_{it}\alpha_{jt}\mathbf{b}_{it}\mathbf{b}_{jt}' \qquad (i \neq j),$$

$$\mathbf{Q}_t^{iv} = -\alpha_{it}\left(\mathbf{b}_{it}' \otimes \mathbf{V}^{-1} + \frac{1}{2}\mathbf{b}_{it}(\operatorname{vec}(\mathbf{B}_{it} - \bar{\mathbf{B}}_t))'\right)\mathbf{D},$$

and

$$\mathbf{Q}_t^{vv} = -\mathbf{D}'\left(\left(\sum_{i=1}^{g}\alpha_{it}\mathbf{b}_{it}\mathbf{b}_{it}'\right)\otimes\mathbf{V}^{-1} - \frac{1}{2}\mathbf{V}^{-1}\otimes\mathbf{V}^{-1} - \frac{1}{4}\sum_{i=1}^{g}\alpha_{it}(\operatorname{vec}\mathbf{B}_{it})(\operatorname{vec}\mathbf{B}_{it})' + \frac{1}{4}(\operatorname{vec}\bar{\mathbf{B}}_t)(\operatorname{vec}\bar{\mathbf{B}}_t)'\right)\mathbf{D}.$$

As in Theorem 1 we can use these results to compute $\boldsymbol{\mathcal{I}}_1^{-1}$, $\boldsymbol{\mathcal{I}}_2^{-1}$, and $\boldsymbol{\mathcal{I}}_3^{-1}$.

6. TWO EXAMPLES

To illustrate our theoretical results we present two examples. The maximum likelihood estimates themselves are usually computed via the EM algorithm, which is a derivative-free method, but they can also be computed directly from the likelihood or by setting the score equal to zero or in some other manner. In many cases knowledge of the score (and Hessian) allows an option which will speed up the computations; see Xu and Jordan (1996) for a discussion of gradient-based approaches. The resulting estimates, however, are the same for each method. The purpose of the two examples is to look at the behavior of the information-based standard error estimates in practice and to compare them to other available methods.

Since no explicit formula for the information matrix has been available, researchers typically compute standard errors in multivariate mixture models by means of the bootstrap. The well-known EMMIX software package developed by McLachlan et al. (1999) reports standard errors of the estimates based on four different methods. Methods (A1) and (A2) are parametric and nonparametric bootstrap methods, respectively, tailored to the initial sample. They perform repeated draws from either a multivariate normal mixture with parameters fixed at their estimated values or from the nonparametric estimate of the sampling distribution of the data, then estimate the model for each sample and compute the in-sample bootstrap standard errors of the corresponding parameter estimates. Method (A3) follows Newton and Raftery (1994) and performs the bootstrap on a weighted version of the data. The fourth method computes standard errors from the outer product of the score, and is based on Basford et al. (1997, section 3). This should be the same as our formula for $\boldsymbol{\mathcal{I}}_1^{-1}$, but verification of this fact is not possible because EMMIX does not always provide credible results in this case. This leaves us with three bootstrap methods to consider. Note however that, since we have coded $\boldsymbol{\mathcal{I}}_1$, we can provide comparisons of the Hessian and sandwich estimates of standard errors with both bootstrap-based and outer product-based standard error estimates. Further details about the four methods can be found in McLachlan and Peel (2000, section 2.16).

We compare these three "EM bootstrap" standard errors with the three standard errors computed from our formulae. Method (B1) employs $\boldsymbol{\mathcal{I}}_1^{-1}$ based on the outer product of the score, (B2) uses $\boldsymbol{\mathcal{I}}_2^{-1}$ based on the Hessian matrix, while (B3) uses the robust sandwich matrix $\operatorname{var}\hat{\boldsymbol{\theta}}$ as given in (9).

We consider two popular and much-studied datasets: the hemophilia dataset and the Iris dataset.

The Hemophilia Dataset

Human genes are carried on chromosomes and two of these, labeled X and Y, determine our sex. Females have two X chromosomes, males have an X and a Y chromosome. Hemophilia is a hereditary recessive X-linked blood clotting disorder where an essential clotting factor is either partly or completely missing. While only males have hemophilia, females can carry the affected gene and pass it on to their children. If the mother carries the hemophilia gene and the father does not have hemophilia, then a male child will have a 50:50 chance of having hemophilia (because he will inherit one of his mother's two X chromosomes, one of which is faulty) and a female child will have a 50:50 chance of carrying the gene (for the same reason). If the mother is not a carrier, but the father has hemophilia, then a male child will not be affected (because he inherits his father's normal Y chromosome) but a female child will always be a carrier (because she inherits her father's faulty X chromosome).

The hemophilia data were collected by Habbema, Hermans, and van den Broek (1974), and were extensively analyzed in a number of papers; see *inter alia* McLachlan and Peel (2000, pp. 103–104). The question is how to discriminate between "normal" women and hemophilia A carriers on the basis of measurements on two variables: antihemophilic factor (AHF) activity and AHF-like antigen. We have 30 observations on women who do not carry the hemophilia gene and 45 observations on women who do carry the gene. We thus have $n = 75$ observations on $m = 2$ features from $g = 2$ groups of women.

Our findings are recorded in Table 1, where all estimates and standard errors (except for π_1) have been multiplied by 100 to facilitate presentation. The EM bootstrap results are obtained from 100 samples for each method and the standard errors correspond closely to those reported in the literature. The three EM bootstraps standard errors are roughly of the same order of magnitude. We shall compare our information-based standard errors with the parametric bootstrap (A1), which is the most relevant here given our focus on multivariate normal mixtures.

The standard errors obtained by the explicit score and Hessian formulae are somewhat smaller than the bootstrap standard errors, which confirms the finding in Basford et al. (1997) concerning $\mathcal{I}_1^{-1}$ (outer score). In eight of the eleven cases, the standard errors computed from $\mathcal{I}_2^{-1}$ (Hessian) lie in between the standard error based on the score and the standard error based on the robust estimator, as predicted in Section 3. When this happens, the misspecification-robust standard error (B3) is the smallest of the three. For both groups of women the robust standard error is about 63% of the standard error based on parametric bootstrap (A1).

The Iris Dataset

The Iris flower data were collected by Anderson (1935) with the purpose to quantify the geographic variation of Iris flowers in the Gaspé Peninsula, located on the eastern tip of the province of Québec in Canada. The dataset consists of 50 samples from each of three species of Iris flowers: *Iris setosa* (Arctic iris), *Iris versicolor* (Southern blue flag), and *Iris virginica* (Northern blue flag). Four features were measured from each flower: sepal length, sepal width, petal length, and petal width. Based on the combination of the four features, Sir Ronald Fisher (1936) developed a linear discriminant model to determine which species they are.

The dataset thus consists of $n = 150$ measurements on $m = 4$ features from $g = 3$ Iris species. Table 2 contains parameter estimates and standard errors of the means μ_i and variances v_{ii} (the covariance estimates v_{ij} for $i \neq j$ have been omitted), where all estimates and standard errors (except π_1 and π_2) have again been multiplied by 100. As before, the EM bootstrap results are obtained from 100 samples for each method and the standard errors correspond closely to those reported in the literature.

In contrast to the first example, the standard errors obtained by $\mathcal{I}_1^{-1}$ (outer score) are somewhat larger than the parametric bootstrap standard errors, again in accordance to the finding in Basford et al. (1997). In 18 of the 26 cases, the standard errors computed from $\mathcal{I}_2^{-1}$ (Hessian) lie in between the standard error

Table 1. Estimation results—hemophilia data

		Standard error					
		EM bootstrap			Our method		
Variable	Estimate	(A1)	(A2)	(A3)	(B1)	(B2)	(B3)
		Weight					
π_1	0.51	0.13	0.12	0.14	0.13	0.05	0.03
		Woman does not carry hemophilia					
μ_1	−11.48	3.90	4.16	4.19	3.76	2.36	1.95
μ_2	−2.45	3.22	3.42	2.91	2.30	2.18	2.11
v_{11}	111.48	63.74	71.24	68.43	43.95	37.72	41.25
v_{12}	65.35	45.06	46.90	47.62	29.44	28.98	32.39
v_{22}	123.44	39.89	34.41	34.84	41.78	30.85	24.56
		Woman carries hemophilia					
μ_1	−36.53	4.53	3.99	4.66	4.12	2.75	2.43
μ_2	−4.52	4.73	5.71	7.11	3.23	3.21	3.27
v_{11}	159.56	58.93	53.90	63.85	52.07	44.85	42.25
v_{12}	150.10	67.00	55.41	70.00	57.83	47.94	41.34
v_{22}	322.00	109.11	81.22	204.45	104.51	77.87	63.70

Table 2. Estimation results—Iris data

		Standard error					
		EM bootstrap			Our method		
Variable	Estimate	(A1)	(A2)	(A3)	(B1)	(B2)	(B3)
			Weights				
π_1	0.333	0.037	0.038	0.037	0.039	0.022	0.013
π_2	0.367	0.043	0.044	0.047	0.041	0.023	0.013
			Iris setosa				
μ_1	500.60	5.05	4.90	4.93	5.67	4.93	4.93
μ_2	342.80	5.66	5.10	5.29	5.89	5.31	5.31
μ_3	146.20	2.49	2.90	2.43	2.96	2.43	2.43
μ_4	24.60	1.54	1.70	1.46	2.04	1.48	1.48
v_{11}	12.18	2.31	2.46	1.94	3.04	2.44	2.21
v_{22}	14.08	2.58	3.25	3.03	2.84	2.82	3.30
v_{33}	2.96	0.58	0.74	0.60	0.63	0.59	0.70
v_{44}	1.09	0.20	0.30	0.28	0.25	0.22	0.29
			Iris versicolor				
μ_1	591.50	8.43	7.83	9.20	10.31	7.99	7.97
μ_2	277.78	4.76	5.90	5.89	5.63	4.61	4.67
μ_3	420.16	8.08	8.27	8.51	9.74	6.99	6.80
μ_4	129.70	3.21	3.36	3.35	3.33	2.80	2.78
v_{11}	27.53	6.01	5.36	7.37	8.31	5.88	4.88
v_{22}	9.11	1.96	2.03	2.10	2.56	1.98	1.86
v_{33}	20.06	5.36	4.99	6.60	5.88	4.46	4.39
v_{44}	3.20	0.83	0.78	0.72	1.04	0.72	0.55
			Iris virginica				
μ_1	654.45	9.12	8.85	10.58	10.82	8.57	8.49
μ_2	294.87	4.46	5.49	5.21	4.90	4.53	4.59
μ_3	547.96	8.84	10.08	12.78	10.35	8.10	8.14
μ_4	198.46	4.72	6.07	6.64	4.33	4.23	4.29
v_{11}	38.70	7.76	8.46	6.28	10.32	7.48	7.38
v_{22}	11.03	2.15	2.86	2.37	2.34	2.13	2.37
v_{33}	32.78	6.64	8.82	8.44	11.20	6.53	6.17
v_{44}	8.58	1.91	2.49	2.66	2.83	1.78	1.38

based on the score and the standard error based on the robust estimator, as predicted in Section 3. And again, remarkably, when this happens the misspecification-robust standard error (B3) is the smallest of the three. In this example, contrary to the previous example, the robust standard error is only slightly smaller on average than the standard error based on parametric bootstrap.

Our two examples demonstrate that the implementation of second-order derivative formulae is a practical alternative to the currently used bootstrap. Our program for computing the standard errors of $\mathcal{I}_1^{-1}$ (outer product), $\mathcal{I}_2^{-1}$ (Hessian), and $\mathcal{I}_3^{-1}$ (sandwich) is extremely fast. The resulting standard errors are comparable in size to the bootstrap standard errors, but they are sufficiently different to justify the question which standard errors are the most accurate. This question cannot be answered in estimation exercises. We need a small Monte Carlo experiment where the precision of the estimates is known.

7. SIMULATIONS

We wish to assess the small sample behavior of the information-based estimates and compare it to the behavior of the traditional bootstrap-based methods. We shall assume that the data are generated by an m-variate normal mixture model, determined by the parameters $(\pi_i, \boldsymbol{\mu}_i, \mathbf{V}_i)$ for $i = 1, \ldots, g$, so that we have $g - 1 + gm(m+3)/2$ parameters in total. It is convenient to construct matrices $\mathbf{A}_i$ such that $\mathbf{A}_i\mathbf{A}_i' = \mathbf{V}_i$. We then obtain R samples, each of size n, from this distribution where each sample is generated as follows.

- Draw a sample of size n from the categorical distribution defined by $\Pr(z = i) = \pi_i$. This gives n integer numbers, say $z_1, \ldots, z_n$, such that $1 \le z_j \le g$ for all j.
- Define n_i as the number of times that $z_j = i$. Notice that $\sum_i n_i = n$.
- For $i = 1, \ldots, g$ draw mn_i standard-normal random numbers and assemble these in $m \times 1$ vectors $\boldsymbol{\epsilon}_{i,1}, \ldots, \boldsymbol{\epsilon}_{i,n_i}$. Now define

$$\mathbf{x}_{i,v} = \boldsymbol{\mu}_i + \mathbf{A}_i\boldsymbol{\epsilon}_{i,v} \sim \mathrm{N}(\boldsymbol{\mu}_i, \mathbf{V}_i) \qquad (v = 1, \ldots, n_i).$$

The set $\{\mathbf{x}_{i,v}\}$ then consists of n m-dimensional vectors from the required mixture. Given this sample of size n we estimate the parameters and standard errors, assuming that we know the distribution is a mixture of g normals.

We perform R replications of this procedure. For each $r = 1, \ldots, R$ we obtain an estimate of each of the parameters. The R

estimates together define a distribution for each parameter estimate, and if R is sufficiently large the variance of this distribution is the "true" variance of the estimator. Our question now is how well the information-based standard error approximate this "true" standard error. We perform four experiments. In each case we take $m = g = 2$, $\pi_1 = \pi_2 = 0.5$, and we let $n = 100$ and $n = 500$, respectively.

(a) *Correct specification*. The mixture distributions are both normal. There is no misspecification, so the model is the same as the data-generating process. We let

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 5 \\ 5 \end{pmatrix},$$
$$\mathbf{V}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{V}_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

(b) *Overspecification*. Same as (a), except that

$$\mathbf{V}_1 = \mathbf{V}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

However, we do not know that the variance matrices are the same and hence we estimate them separately.

(c) *Constrained estimation*. Same as (b), except that we now know that the variance matrices are equal and therefore take this constraint into account, using Theorem 3 rather than Theorem 1.

(d) *Misspecification in distribution*. The two mixture distributions are not normal. The true underlying distributions are $F(k_1, k_2)$, but we are ignorant about this and take them to be normal. Instead of sampling from a multivariate F-distribution we draw a sample $\{\eta_h^*\}$ from the univariate $F(k_1, k_2)$-distribution. We then define

$$\eta_h = \sqrt{\frac{k_1(k_2-4)}{2(k_1+k_2-2)}}\left(\frac{k_2-2}{k_2}\eta_h^* - 1\right),$$

so that the $\{\eta_h\}$ are independent and identically distributed with mean zero and variance one, but of course there will be skewness and kurtosis. For $i = 1, \ldots, g$ draw mn_i random numbers η_h in this way, assemble these in $m \times 1$ vectors $\boldsymbol{\epsilon}_{i,1}, \ldots, \boldsymbol{\epsilon}_{i,n_i}$, and obtain $\mathbf{x}_{i,v}$ as before. We let $k_1 = 5$ and $k_2 = 10$, so that the first four moments exist but the fifth and higher moments do not.

Each estimation method provides an algorithm for obtaining estimates and standard errors of the parameters θ_j, which we denote as $\hat{\theta}_j$ and $s_j = \widehat{\text{var}}^{1/2}(\hat{\theta}_j)$, respectively. Based on R replications we approximate the distributions of $\hat{\theta}_j$ and s_j from which we can compute moments of interest. Letting $\hat{\theta}_j^{(r)}$ and $s_j^{(r)}$ denote the estimates in the rth replication, we find the standard error (SE) of $\hat{\theta}_j$ as

$$\text{SE}(\hat{\theta}_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_j^{(r)} - \bar{\theta}_j\right)^2}, \qquad \bar{\theta}_j = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_j^{(r)}.$$

We wish to know whether the reported standard errors are close to the actual standard errors of the estimators, and we evaluate this "closeness" in terms of the root mean squared error (RMSE) of the standard errors of the parameter estimates. We first compute

$$S_{1j} = \frac{1}{R}\sum_{r=1}^{R} s_j^{(r)}, \qquad S_{2j} = \frac{1}{R}\sum_{r=1}^{R}\left(s_j^{(r)}\right)^2,$$

from which we obtain

$$\text{SE}(s_j) = \sqrt{S_{2j} - S_{1j}^2}.$$

In order to find the bias and mean squared error of s_j we need to know the "true" value of s_j. For sufficiently large R, this value is given by $\text{SE}(\hat{\theta}_j)$. We find

$$\text{BIAS}(s_j) = S_{1j} - \text{SE}(\hat{\theta}_j),$$
$$\text{RMSE}(s_j) = \sqrt{\text{SE}^2(s_j) + \text{BIAS}^2(s_j)},$$

and thus we obtain the RMSE, BIAS, and SE of s_j for each j.

In our experiments we use $R = 50{,}000$ replications for computing the "true" standard errors [10,000 in case (d)] and $R = 10{,}000$ replications for computing the estimated standard errors [1000 in case (d)]. The reason we use less replications in case (d) is that we want to avoid draws with badly separated means that could induce label switching. To compute bootstrap-based standard errors, we rely on 100 bootstrap samples (Efron and Tibshirani 1993). We use the EMMIX Fortran code converted to run in R to generate mixture samples, and obtain parameter estimates and bootstrap-based standard errors. We then import the parameter estimates into MATLAB and use them to obtain the information-based standard error estimates.

Notice that in all four cases the means are well separated. This is useful for three reasons: first, label switching problems across simulations are less likely to occur; second, the ML estimates for well-separated means are accurate enough to allow us to focus on standard error analysis rather than inaccuracies in parameter estimates; and third, we expect the bootstrap-based standard errors to work particularly well when accurate parameter estimates are used for bootstrap samples. Thus, to bring out possible advantages of the information-based method, we consider cases where the bootstrap-based methods should work particularly well.

Let us now discuss the simulation results, where we confine our discussion to the standard errors of the ML estimates, because the ML estimates themselves are the same for each method. In Table 3 we report the RMSE of the estimated standard errors for $n = 500$ in the correctly specified case (a). We see that method (B2) based on $\mathcal{I}_2^{-1}$ (the Hessian) outperforms the EM parametric bootstrap method (A1), which in turn is slightly better than methods (B3) (sandwich) and (B1) (outer score). The observed information matrix $\mathcal{I}_1^{-1}$ based on the outer product of the scores typically performs worst of the three information-based estimates and is therefore not recommended. The poor performance of the outer score matrix confirms results in previous studies; see, for example, Basford et al. (1997). In correctly specified cases we would expect that the parametric bootstrap and the Hessian-based observed information matrix perform well relative to other methods, and this is indeed the case. Our general conclusion for correctly specified cases is that method (B2) based on $\mathcal{I}_2^{-1}$ performs best, followed by the parametric bootstrap method (A1). In contrast to the claim of Day

Table 3. Simulation results, case (a), $n = 500$

		Root mean square error of SE					
		EM bootstrap			Our method		
Variable	Value	(A1)	(A2)	(A3)	(B1)	(B2)	(B3)
				Weight			
π_1	0.5	0.0008	0.0016	0.0016	0.0001	0.0067	0.0114
				Group 1			
μ_1	0	0.0061	0.0059	0.0059	0.0036	0.0034	0.0036
μ_2	0	0.0050	0.0059	0.0059	0.0036	0.0035	0.0036
v_{11}	1	0.0115	0.0139	0.0138	0.0114	0.0092	0.0120
v_{12}	0	0.0060	0.0085	0.0083	0.0066	0.0052	0.0069
v_{22}	1	0.0107	0.0138	0.0138	0.0114	0.0093	0.0121
				Group 2			
μ_1	5	0.0066	0.0088	0.0088	0.0056	0.0055	0.0057
μ_2	5	0.0069	0.0089	0.0088	0.0056	0.0056	0.0058
v_{11}	2	0.0262	0.0305	0.0305	0.0258	0.0217	0.0265
v_{12}	1	0.0193	0.0243	0.0243	0.0206	0.0178	0.0210
v_{22}	2	0.0237	0.0305	0.0309	0.0254	0.0221	0.0269

(1969) and McLachlan and Peel (2000, p. 68) that one needs very large sample sizes before the observed information matrix gives accurate results, we find that very good accuracy can be obtained for $n = 500$ and even for $n = 100$.

The mean squared error of the standard error is the sum of the variance and the square of the bias. The contribution of the bias is small. In the case reported in Table 3, the ratio of the absolute bias to the RMSE is 9% for method (B2) when we average over all 11 parameters. The bias is typically negative for all methods. As McLachlan and Peel (2000, p. 67) point out, delta methods such as the "supplemented" EM method or the conditional bootstrap often underestimate the standard errors, and the same occurs here. Since the bias is small in all correctly specified models, this is not a serious problem.

We notice that the RMSE of the standard error of the mixing proportion $\hat{\pi}_1$ is relatively high for methods (B2) and (B3), both of which employ the Hessian matrix. The situation is somewhat different here than for the other parameters, because the standard error of $\hat{\pi}_1$ is estimated very precisely but with a relatively large negative bias. Of course, the bias decreases when n increases, but in small samples the standard error of $\hat{\pi}_1$ is systematically underestimated. This seems to be a general phenomenon when estimating mixing proportions with information-based methods, and it can possibly be repaired through a bias-correction factor. We do not, however, pursue this problem here. Even with the relatively large RMSE of the mixing proportion, method (B2) performs best, and this underlines the fact that this method estimates the standard errors of the means μ_i and the variance components v_{ij} very precisely.

In Table 4 we provide a general overview of the RMSE results of all four cases considered, for $n = 100$ and $n = 500$. In cases (b) and (c) we illustrate the special case where $\mathbf{V}_1 = \mathbf{V}_2$. In case (b) we are ignorant of this fact and hence the model is overspecified but not misspecified. In case (c) we take the constraint into account and this leads to more precision of the standard errors. The RMSE is reduced by about 50% when $n = 100$ and by about 35% when $n = 500$. Again, the Hessian-based estimate $\mathcal{I}_2^{-1}$ is the most accurate of the six variance matrix esti-

Table 4. Overview of the four simulation experiments

	Root mean square error of SE					
	EM bootstrap			Our method		
Experiment	(A1)	(A2)	(A3)	(B1)	(B2)	(B3)
			Correctly specified			
100	0.0674	0.0920	0.0869	0.0793	0.0647	0.0827
500	0.0137	0.0169	0.0169	0.0139	0.0121	0.0149
			Overspecified			
100	0.0307	0.0373	0.0378	0.0430	0.0295	0.0372
500	0.0072	0.0089	0.0090	0.0075	0.0061	0.0081
			Constrained			
100	0.0155	0.0201	0.0206	0.0207	0.0150	0.0204
500	0.0052	0.0055	0.0055	0.0037	0.0036	0.0054
			Misspecified, $F(5, 10)$			
100	1.5500	1.4433	1.5085	–	1.5143	1.3605
500	0.9799	0.9767	1.1627	1.1524	1.0960	0.9241

Table 5. Size of IM test, simulation results

	Critical values				
n	9.34	12.55	15.99	18.31	23.21
100	1.0000	0.9999	0.9999	0.9996	0.9984
500	0.9975	0.9843	0.9500	0.9180	0.8186
1000	0.9898	0.9564	0.8868	0.8228	0.6571
∞	0.5000	0.2500	0.1000	0.0500	0.0100

mates considered. In case (d) we consider misspecified models where both skewness and kurtosis are present in the underlying distributions, but ignored in the estimation. One would expect that the nonparametric bootstrap estimates (A2) and (A3) and our proposed sandwich estimate (B3) would perform well in misspecified models, and this is usually, but not always, the case. Our sandwich estimate $\mathcal{I}_3^{-1}$ has the lowest RMSE in all cases. The outer score estimate (B1) fails to produce credible outcomes when $n = 100$. If we repeat the experiment based on other F-distributions we obtain similar results.

Finally we consider the information matrix test presented in Section 4. The IM test has limitations in practice because the asymptotic χ^2-distribution is typically a poor approximation to the finite sample distribution of the test statistic. We briefly investigate the finite sample properties of our version of the IM test via simulations to give some idea of just how useful it can be. Let us consider the correctly specified model (a) with $m = g = 2$ so that the IM test of Theorem 2 should be asymptotically χ^2-distributed with $gm(m+3)/2 = 10$ degrees of freedom. In Table 5 we compute the sizes for $n = 100$, 500, and 1000, based on 10,000 replications and using the critical values that are valid in the asymptotic distribution. As expected, the results are not encouraging, thus confirming findings by many authors; see Davidson and MacKinnon (2004, section 16.9). There is, however, a viable alternative based on the same IM statistic, proposed by Horowitz (1994) (see also Davidson and MacKinnon 2004, pp. 663–665), namely to bootstrap the critical values of the IM test for each particular application. This is what we recommend.

8. CONCLUSIONS

Despite McLachlan and Krishnan's (1997, p. 111) claim that analytical derivation of the Hessian matrix of the log-likelihood for multivariate mixtures seems to be difficult or at least tedious, we show that it pays to have these formulae available for normal mixtures. In correctly specified models the method based on the observed Hessian-based information matrix $\mathcal{I}_2^{-1}$ is the best in terms of RMSE. In misspecified models the method based on the sandwich matrix $\mathcal{I}_3^{-1}$ is the best, even if the standard errors of the observed information matrix based on the outer product of the scores are large, as is sometimes the case. In general, the bias of the two methods is either the smallest in their category (correctly specified or misspecified) or if not, it becomes the smallest as the sample size increases to $n = 500$. Our MATLAB code for computing the standard errors runs in virtually no time unless both m and g are very large, and it is even faster than the bootstrap.

There are at least two additional advantages in using information-based methods. First, the Hessian we computed can be useful to detect instances where the EM algorithm has not converged to the ML solution. Second, if the sample size is not too large relative to the number of parameters to estimate, the methods based on $\mathcal{I}_2^{-1}$ and $\mathcal{I}_3^{-1}$ can be readily used to compute asymptotically valid confidence intervals, while nonparametric bootstrap confidence intervals are often difficult to compute.

APPENDIX: PROOFS

Proof of Theorem 1

Let ϕ_{it} and α_{it} be defined as in (5). Then, since $f(\mathbf{x}_t) = \sum_i \phi_{it}$, we obtain

$$d\log f(\mathbf{x}_t) = \frac{df(\mathbf{x}_t)}{f(\mathbf{x}_t)} = \sum_{i=1}^{g} \frac{d\phi_{it}}{\sum_j \phi_{jt}} = \sum_{i=1}^{g} \alpha_{it}\, d\log\phi_{it} \tag{A.1}$$

and

$$\begin{aligned} d^2\log f(\mathbf{x}_t) &= \left(\frac{d^2 f(\mathbf{x}_t)}{f(\mathbf{x}_t)} - \left(\frac{df(\mathbf{x}_t)}{f(\mathbf{x}_t)}\right)^2\right) \\ &= \left(\frac{\sum_i d^2\phi_{it}}{\sum_j \phi_{jt}} - \left(\frac{\sum_i d\phi_{it}}{\sum_j \phi_{jt}}\right)^2\right) \\ &= \left(\sum_{i=1}^{g} \alpha_{it}(d^2\log\phi_{it} + (d\log\phi_{it})^2)\right. \\ &\quad \left. - \left(\sum_{i=1}^{g} \alpha_{it}\, d\log\phi_{it}\right)^2\right). \end{aligned} \tag{A.2}$$

To evaluate these expressions, we need the first-order and second-order derivatives of $\log\phi_{it}$. Since, using (2),

$$\log f_i(\mathbf{x}) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i),$$

we find

$$\begin{aligned} d\log f_i(\mathbf{x}) &= -\frac{1}{2}d\log|\mathbf{V}_i| + (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}\,d\boldsymbol{\mu}_i \\ &\quad - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'d(\mathbf{V}_i^{-1})(\mathbf{x}-\boldsymbol{\mu}_i) \\ &= -\frac{1}{2}\operatorname{tr}(\mathbf{V}_i^{-1}d\mathbf{V}_i) + (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}\,d\boldsymbol{\mu}_i \\ &\quad + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \end{aligned}$$

and

$$\begin{aligned} d^2\log f_i(\mathbf{x}) &= -\frac{1}{2}\operatorname{tr}((d\mathbf{V}_i^{-1})\,d\mathbf{V}_i) - (d\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad + (\mathbf{x}-\boldsymbol{\mu}_i)'(d\mathbf{V}_i^{-1})\,d\boldsymbol{\mu}_i - (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}\,d\boldsymbol{\mu}_i \\ &\quad - (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) \\ &= \frac{1}{2}\operatorname{tr}\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}\,d\mathbf{V}_i - (d\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - 2(\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}\,d\boldsymbol{\mu}_i \\ &\quad - (\mathbf{x}-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i), \end{aligned}$$

and hence, using (3) and the definitions (6)–(8),

$$\begin{aligned} d\log\phi_{it} &= d\log\pi_i + (\mathbf{x}_t-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}\,d\boldsymbol{\mu}_i - \frac{1}{2}\operatorname{tr}\mathbf{V}_i^{-1}\,d\mathbf{V}_i \\ &\quad + \frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_i) \end{aligned}$$

$$= \mathbf{a}_i' d\boldsymbol{\pi} + \mathbf{b}_{it}' d\boldsymbol{\mu}_i - \frac{1}{2}\operatorname{tr}(\mathbf{B}_{it}\, d\mathbf{V}_i)$$

$$= \mathbf{a}_i' d\boldsymbol{\pi} + \mathbf{b}_{it}' d\boldsymbol{\mu}_i - \frac{1}{2}(\operatorname{vec}\mathbf{B}_{it})'\mathbf{D}\, d\operatorname{vech}\mathbf{V}_i$$

$$= \mathbf{a}_i' d\boldsymbol{\pi} + \mathbf{c}_{it}' d\boldsymbol{\theta}_i \tag{A.3}$$

and

$$\begin{aligned} d^2 \log\phi_{it} &= d^2\log\pi_i - (d\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - 2(\mathbf{x}_t - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - (\mathbf{x}_t - \boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_i) \\ &\quad + \frac{1}{2}\operatorname{tr}\mathbf{V}_i^{-1}(d\mathbf{V}_i)\mathbf{V}_i^{-1}(d\mathbf{V}_i) \\ &= -(d\boldsymbol{\pi})'\mathbf{a}_i\mathbf{a}_i'(d\boldsymbol{\pi}) - (d\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - 2\mathbf{b}_{it}'(d\mathbf{V}_i)\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - \frac{1}{2}\operatorname{tr}(\mathbf{V}_i^{-1} - 2\mathbf{B}_{it})(d\mathbf{V}_i)\mathbf{V}_i^{-1}(d\mathbf{V}_i) \\ &= -(d\boldsymbol{\pi})'\mathbf{a}_i\mathbf{a}_i'(d\boldsymbol{\pi}) - (d\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - 2(d\operatorname{vec}\mathbf{V}_i)'(\mathbf{b}_{it}\otimes\mathbf{V}_i^{-1})(d\boldsymbol{\mu}_i) \\ &\quad - \frac{1}{2}(d\operatorname{vec}\mathbf{V}_i)'((\mathbf{V}_i^{-1} - 2\mathbf{B}_{it})\otimes\mathbf{V}_i^{-1})(d\operatorname{vec}\mathbf{V}_i) \\ &= -(d\boldsymbol{\pi})'\mathbf{a}_i\mathbf{a}_i'(d\boldsymbol{\pi}) - (d\boldsymbol{\mu}_i)'\mathbf{V}_i^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - 2(d\operatorname{vech}\mathbf{V}_i)'\mathbf{D}'(\mathbf{b}_{it}\otimes\mathbf{V}_i^{-1})(d\boldsymbol{\mu}_i) \\ &\quad - \frac{1}{2}(d\operatorname{vech}\mathbf{V}_i)'\mathbf{D}'((\mathbf{V}_i^{-1} - 2\mathbf{B}_{it})\otimes\mathbf{V}_i^{-1})\mathbf{D}(d\operatorname{vech}\mathbf{V}_i) \\ &= -\begin{pmatrix} d\boldsymbol{\pi} \\ d\boldsymbol{\theta}_i \end{pmatrix}' \begin{pmatrix} \mathbf{a}_i\mathbf{a}_i' & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{it} \end{pmatrix} \begin{pmatrix} d\boldsymbol{\pi} \\ d\boldsymbol{\theta}_i \end{pmatrix}. \end{aligned} \tag{A.4}$$

Inserting (A.3) in (A.1), and (A.3) and (A.4) in (A.2) completes the proof.

Proof of Theorem 2

This follows from the expression of $\mathbf{W}_t(\boldsymbol{\theta})$ and the development in Lancaster (1984).

Proof of Theorem 3

From (A.3) we see that

$$\begin{aligned} d\log\phi_{it} &= \mathbf{a}_i' d\boldsymbol{\pi} + \mathbf{c}_{it}' d\boldsymbol{\theta}_i \\ &= \mathbf{a}_i' d\boldsymbol{\pi} + \mathbf{b}_{it}' d\boldsymbol{\mu}_i - \frac{1}{2}(\operatorname{vec}\mathbf{B}_{it})'\mathbf{D}\, d\mathbf{v}, \end{aligned}$$

and from (A.4) that

$$\begin{aligned} d^2\log\phi_{it} &= -(d\boldsymbol{\pi})'\mathbf{a}_i\mathbf{a}_i'(d\boldsymbol{\pi}) - (d\boldsymbol{\theta}_i)'\mathbf{C}_{it}(d\boldsymbol{\theta}_i) \\ &= -(d\boldsymbol{\pi})'\mathbf{a}_i\mathbf{a}_i'(d\boldsymbol{\pi}) - (d\boldsymbol{\mu}_i)'\mathbf{V}^{-1}(d\boldsymbol{\mu}_i) \\ &\quad - 2(d\boldsymbol{\mu}_i)'(\mathbf{b}_{it}'\otimes\mathbf{V}^{-1})\mathbf{D}(d\mathbf{v}) \\ &\quad - \frac{1}{2}(d\mathbf{v})'\mathbf{D}'((2\mathbf{b}_{it}\mathbf{b}_{it}' - \mathbf{V}^{-1})\otimes\mathbf{V}^{-1})\mathbf{D}(d\mathbf{v}). \end{aligned}$$

The results then follow—after some tedious but straightforward algebra—from (A.1) and (A.2).

[Received May 2008. Revised May 2009.]

REFERENCES

Aitken, M., and Rubin, D. B. (1985), "Estimation and Hypothesis Testing in Finite Mixture Models," *Journal of the Royal Statistical Society, Ser. B*, 47, 67–75.

Ali, M. M., and Nadarajah, S. (2007), "Information Matrices for Normal and Laplace Mixtures," *Information Sciences*, 177, 947–955.

Anderson, E. (1935), "The Irises of the Gaspé Peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

Basford, K. E., Greenway, D. R., McLachlan, G. J., and Peel, D. (1997), "Standard Errors of Fitted Means Under Normal Mixture Models," *Computational Statistics*, 12, 1–17.

Behboodian, J. (1972), "Information Matrix for a Mixture of Two Normal Distributions," *Journal of Statistical Computation and Simulation*, 1, 1–16.

Chesher, A. D. (1983), "The Information Matrix Test: Simplified Calculation via a Score Test Interpretation," *Economics Letters*, 13, 15–48.

Davidson, R., and MacKinnon, J. G. (2004), *Econometric Theory and Methods*, New York: Oxford University Press.

Day, N. E. (1969), "Estimating the Components of a Mixture of Normal Distributions," *Biometrika*, 56, 463–474.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38.

Dietz, E., and Böhning, D. (1996), "Statistical Inference Based on a General Model of Unobserved Heterogeneity," in *Advances in GLIM and Statistical Modeling. Lecture Notes in Statistics*, eds. L. Fahrmeir, F. Francis, R. Gilchrist, and G. Tutz, Berlin: Springer, pp. 75–82.

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, London: Chapman & Hall.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Habbema, J. D. F., Hermans, J., and van den Broek, K. (1974), "A Step-Wise Discriminant Analysis Program Using Density Estimation," in *Proceedings in Computational Statistics, Compstat 1974*, Wien: Physica Verlag, pp. 101–110.

Hathaway, R. J. (1985), "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions," *The Annals of Statistics*, 13, 795–800.

Horowitz, J. L. (1994), "Bootstrap-Based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61, 395–411.

Huber, P. J. (1967), "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, eds. L. M. LeCam and J. Neyman, Berkeley: University of California Press, pp. 221–233.

Lancaster, A. (1984), "The Covariance Matrix of the Information Matrix Test," *Econometrica*, 52, 1051–1053.

Liu, C. (1998), "Information Matrix Computation From Conditional Information via Normal Approximation," *Biometrika*, 85, 973–979.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 44, 226–233.

Magnus, J. R. (1988), *Linear Structures. Griffin's Statistical Monographs and Courses*, Vol. 42, London: Edward Arnold and New York: Oxford University Press.

Magnus, J. R., and Neudecker, H. (1988), *Matrix Differential Calculus With Applications in Statistics and Econometrics* (2nd ed.), Chichester/New York: Wiley.

McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley.

McLachlan, G. J., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

McLachlan, G. J., Peel, D., Basford, K. E., and Adams, P. (1999), "Fitting of Mixtures of Normal and *t*-Components," *Journal of Statistical Software*, 4 (2), available at *www.maths.uq.edu.au/~gjm/emmix/emmix.html*.

Newcomb, S. (1886), "A Generalized Theory of the Combination of Observations so as to Obtain the Best Result," *American Journal of Mathematics*, 8, 343–366.

Newton, M. A., and Raftery, A. E. (1994), "Approximate Bayesian Inference With the Weighted Likelihood Bootstrap" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 56, 3–48.

Pearson, K. (1894), "Contribution to the Mathematical Theory of Evolution," *Philosophical Transactions of the Royal Society, Ser. A*, 185, 71–110.

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge, MA: Belknap.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–26.

Xu, L., and Jordan, M. I. (1996), "On Convergence Properties of the EM Algorithm for Gaussian Mixtures," *Neural Computation*, 8, 129–151.