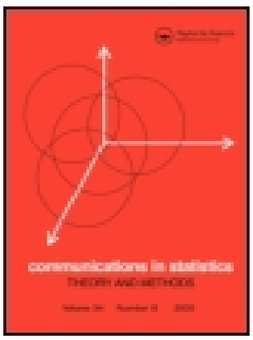

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

The estimation of normal mixtures with latent variables

Gideon Magnus & Jan R. Magnus

To cite this article: Gideon Magnus & Jan R. Magnus (2018): The estimation of normal mixtures with latent variables, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2018.1429625

To link to this article: https://doi.org/10.1080/03610926.2018.1429625

Published online: 08 Feb 2018.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2018, VOL. 0, NO. 0, 1–15
https://doi.org/10.1080/03610926.2018.1429625

Taylor & Francis
Taylor & Francis Group

The estimation of normal mixtures with latent variables

Gideon Magnus[a] and Jan R. Magnus[b]

[a]State Street Bank and Trust Company, New York City, USA; [b]Vrije Universiteit Amsterdam, The Netherlands

ABSTRACT
This paper considers the class of normal latent factor mixture models. It presents a method for estimating the posterior distribution of the parameters, derives analytical expressions for both the first and second derivatives of the posterior kernel (the score and Hessian), and provides posterior approximations that can be computed relatively quickly.

ARTICLE HISTORY
Received 25 February 2017
Accepted 13 January 2018

KEYWORDS
Mixture modelling; Factor analyzers; Hessian matrix.

MATHEMATICS SUBJECT CLASSIFICATION
C11; C38; C51

1. Introduction

For the class of normal latent factor mixture models we present a method for estimating the posterior distribution of the parameters, derive analytical expressions for the first derivative (the score vector) and in particular for the second derivative (the Hessian matrix) of the posterior kernel, and provide two posterior approximations that can be computed relatively quickly. The explicit formulae for the Hessian matrix constitute the main theoretical contribution of this paper.

There is a sizeable literature on mixture models; for overviews see McLachlan and Peel (2000), Frühwirth-Schnatter (2006), and Rossi (2014). Our interest is in a sub-class of the mixture models, namely normal latent factor mixture models. Closely related papers include Ghahramani and Hinton (1996), McLachlan, Peel, and Bean (2003), McLachlan, Ng, and Bean (2006), McLachlan, Bean, and Ben-Tovim Jones (2007), Boldea and Magnus (2009), Baek, McLachlan, and Flack (2010), Andrews and McNicholas (2011), Murray, Browne, and McNicholas (2014), and Magnus (2016).

Our paper is most closely related to Montanari and Viroli (2010), who examine essentially the same model, which they call 'heteroscedastic factor mixture analysis'. We extend their analysis in two important ways.

First, as is common in the mixture modeling literature, Montanari and Viroli focus exclusively on locating the posterior mode, and do not investigate posterior parameter uncertainty. Locating the mode is done through the expectation maximization (EM) algorithm or one of its variants such as alternating expectation conditional maximization (AECM). In contrast, we do not use an EM-type algorithm, but locate the posterior kernel through gradient-based numerical optimization. We study the full posterior distribution using a random walk Metropolis-Hastings sampler, and examine two variance approximations that can be

CONTACT Jan R. Magnus jan@janmagnus.nl Department of Econometrics & OR, Vrije Universiteit Amsterdam, The Netherlands.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta

computed relatively quickly. Second, our model explicitly accounts for missing observations, a feature absent from almost all related models studied in the literature.

We derive analytical expressions for the first and second derivatives (the score and Hessian). The score vector helps us locate the posterior mode through numerical optimization. Since we also know the Hessian matrix, we could utilize this knowledge in the optimization process through Newton-Raphson type methods. In our experience, however, methods based on only the score work as well or better.

The Hessian is, however, important in approximating the posterior variance. The expected outer product of the score vector (the Information Matrix) and the negative of the inverse Hessian matrix provide two approximations of the posterior variance. The first only uses the score, while the second requires the Hessian. Our numerical examples suggest that the Hessian approximation is superior to the Information Matrix approximation.

Although our approach is Bayesian, our analytical results are also relevant for frequentist analysis, because the approximations to the posterior variance are valid approximations to the variance of the maximum likelihood estimator. In our illustration we concentrate on the Bayesian approach, and hence we do not investigate how accurate the approximations are in a frequentist context.

The paper is organized as follows. The model is presented in Section 2. Identification issues are discussed in Section 3. The score and the Hessian are derived in Sections 4 and 5, respectively. Section 6 provides an empirical illustration and Section 7 concludes. There are two mathematical appendices.

2. The model

We are interested in an n-dimensional vector of observables x, and we assume that x is a linear function of a (much) smaller number of latent factors. Specifically, x can be decomposed as

$$x = Bz + \epsilon, \qquad \epsilon \sim \mathrm{N}(0, \Psi). \tag{1}$$

The $m < n$ components of the random vector z are the latent factors, hence unobserved, and the $n \times m$ matrix B contains the factor loadings. We assume that z and ϵ are independent, and that Ψ is diagonal.

The distribution of the factors collected in the m-dimensional vector z is a mixture (weighted average) of g normal densities, so that

$$f(z) = \sum_{i=1}^{g} \pi_i f_i(z; m, \mu_i, V_i), \tag{2}$$

where

$$f_i(z; m, \mu_i, V_i) = (2\pi)^{-m/2} |V_i|^{-1/2} \exp\left\{-\frac{1}{2}(z-\mu_i)' V_i^{-1} (z-\mu_i)\right\} \tag{3}$$

and the π_i are weights satisfying $\pi_i > 0$ and $\sum_i \pi_i = 1$.

Since linear combinations of normals are also normal, it is easy to see that the distribution of x is also a mixture (with the same weights) of g normal densities, so that

$$f(x) = \sum_{i=1}^{g} \pi_i f_i(x; n, m_i, W_i), \tag{4}$$

where

$$m_i = B\mu_i, \qquad W_i = BV_iB' + \Psi. \tag{5}$$

The assumption that $\mathrm{E}\,\epsilon = 0$ and the independence of z and ϵ can be relaxed. Also, the diagonality of Ψ is made for simplicity only and is not essential. Even the normality assumption can be relaxed; see Magnus (2016).

The set of parameters is thus given by

$$\theta = \{B, \Psi, \{\pi_i, \mu_i, V_i\}_{i=1}^g\}. \tag{6}$$

Given a sample $x_1, \ldots, x_T$ of independent and identically distributed (iid) observations from this distribution, we can write the log-likelihood as $\mathcal{L}(\theta) = \sum_{t=1}^T \log f(x_t)$. In fact, however, we may not have access to the complete x_t vectors but only to n_t-dimensional subsets $x_{(t)} = S_t'x_t$, where S_t is an $n \times n_t$ selection matrix. The vectors $x_{(t)}$ are now no longer iid, but they are still independent. Hence the log-likelihood becomes

$$\log \mathcal{L}(\theta) = \sum_{t=1}^T \log \mathcal{L}_t(\theta), \tag{7}$$

where

$$\mathcal{L}_t(\theta) = f(x_{(t)}) = \sum_{i=1}^g \pi_i f_i(x_{(t)}; n_t, S_t'm_i, S_t'W_iS_t) = \sum_{i=1}^g \pi_i e^{-\lambda_{it}(\theta)/2} \tag{8}$$

and

$$\lambda_{it}(\theta) = n_t \log(2\pi) + \log|S_t'W_iS_t| + (x_t - m_i)'S_t(S_t'W_iS_t)^{-1}S_t'(x_t - m_i). \tag{9}$$

We assume an improper prior distribution for all parameters, so that locating the posterior mode corresponds with maximizing the log-likelihood $\log \mathcal{L}(\theta)$. Before we present the derivatives, we perform three transformations of the parameters. First, the mixture probabilities π_i need to be positive and sum to one. This is achieved by writing

$$\pi_i = \frac{\xi_i^2}{\sum_{j=1}^g \xi_j^2} \qquad (i = 1, \ldots, g), \tag{10}$$

where we normalize, without loss of generality, $\xi_1 = 1$. Next, each variance matrix V_i needs to be symmetric and positive definite. We therefore let

$$V_i = \widetilde{V}_i\widetilde{V}_i' + \kappa_1 I_m, \tag{11}$$

where $\widetilde{V}_i$ is lower triangular and κ_1 is a given small positive number. Finally, the diagonal components of Ψ must be strictly positive. We write

$$\Psi = \widetilde{\Psi}^2 + \kappa_2 I_n, \tag{12}$$

where $\widetilde{\Psi}$ is diagonal and κ_2 is a given small positive number. Optimization is performed with respect to $\{\{\xi_i\}_{i=2}^g, \{\mu_i, \widetilde{V}_i\}_{i=1}^g, B, \widetilde{\Psi}\}$, in total

$$N = n(m+1) + g(m+1)(m+2)/2 - 1 \tag{13}$$

parameters.

3. Identification

As in any mixture model, the labeling of the g mixture components can be permuted in $g!$ ways without affecting the data-generation process, i.e. the likelihood function has $g!$ symmetric modes. Put differently, the likelihood function is itself a mixture distribution with $g!$ components.

In Figure 1 we provide an example of what a likelihood for a mixture weight π_i might look like in a model with two mixture components (and thus two possible permutations). The likelihood is shown by the black line. The two mixture weights appear to be around $(1/3,\ 2/3)$ and for each weight there are two modes, one for each permutation. We show the two mixture components in blue and red. Ideally we would like to focus on one of these components, ignore the rest, and end up with a unimodal likelihood surface. There is, however, no straightforward method to achieve this.

We follow an approach commonly taken in the literature, and post-process (draws from) the posterior distribution using an identifiability constraint on some parameter with a mixture component label, in our case the parameters contained in $\{\pi_i, \mu_i, \widetilde{V}_i\}_{i=1}^{g}$. Specifically, we assume that $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_g$, see Titterington, Smith, and Makov (1985, Section 3.1). (The result in our simple example above is given by the green line.) This does not separate out one of the $g!$ mixture components, and the resulting posterior will most likely still have multiple modes (Stephens, 2000). Under such an identification scheme, an approximation to the posterior distribution may be reasonably good around one of the $g!$ modes, but this does not necessarily mean that it is a good approximation when we impose the identifiability constraints. This will be especially problematic when the mixture weights are close to each another but the other mixture parameters are very different.

There are several other identification issues, namely concerning ξ_i, $\widetilde{\Psi}$, $\widetilde{V}_i$, and B, and we shall discuss each in turn.

We don't estimate the π_i directly, but indirectly through the ξ_i^2. The ξ_i are identified up to a sign, and hence the corresponding posterior distribution is symmetric at zero. Of course, the absolute value of the parameter *is* identified, and the posterior for the absolute value is simply twice the posterior (at positive values) of the original parameter. For our approximations of

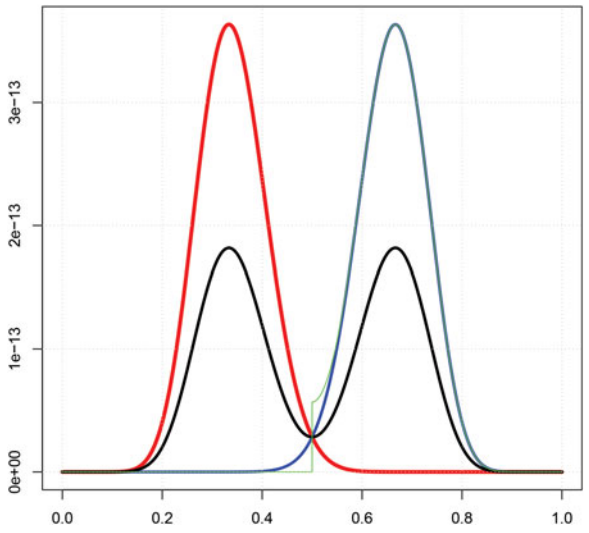

Figure 1. Identification with two mixture components.

the posterior variance (discussed below) this lack of identification does not matter, since the logarithm of the posterior kernel is unaffected by normalizing constants.

Regarding the diagonal elements of $\widetilde{\Psi}$ the same holds as for the ξ_i: they can be positive or negative without affecting the data-generation process, and hence we can examine their absolute values.

The matrices $\widetilde{V}_i$ are lower triangular. For identification there needs to be a one-to-one correspondence between the matrices $\widetilde{V}_i$ and $\widetilde{V}_i\widetilde{V}_i'$. This can be achieved by choosing the diagonal elements of $\widetilde{V}_i$ to be positive, because for every positive definite matrix A there exists a *unique* lower triangular matrix L with positive diagonal elements such that $A = LL'$. The non-singularity of A is essential here. If A is singular and hence only positive semi-definite, then there still exists a lower triangular matrix L with non-negative diagonal elements such that $A = LL'$, but this matrix is not unique. For example, if

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, \qquad L = \begin{pmatrix} 0 & 0 \\ \sin\theta & \cos\theta \end{pmatrix},$$

then $A = LL'$ and the diagonal elements of L are non-negative for every $|\theta| \leq \pi/2$, so that L is not unique.

We do *not* impose positivity on the diagonal elements of $\widetilde{V}_i$ (just as we did not impose positivity for the diagonal elements of $\widetilde{\Psi}$). To find an identified $\widetilde{V}_i$ matrix we transform any 'non-identified' $\widetilde{V}_i$ to an identified one as follows. Define diagonal matrices D_i whose diagonal elements are $+1$ or -1 depending on whether the corresponding diagonal element of $\widetilde{V}_i$ is positive or negative. Then $\widetilde{V}_i\widetilde{V}_i' = (\widetilde{V}_iD_i)(\widetilde{V}_iD_i)'$ and all diagonal elements of $\widetilde{V}_iD_i$ are now positive. Notice that if a diagonal element of $\widetilde{V}_i$ changes sign from negative to positive, then all elements of $\widetilde{V}_i$ in the corresponding column (below the diagonal) change sign as well, and hence it is not correct to adjust only the diagonal elements of $\widetilde{V}_i$. As with $\mathrm{diag}(\widetilde{\Psi})$ and the ξ_i, our posterior approximations are not affected when we examine the transformed $\widetilde{V}_i$ matrices.

The most important identification issue concerns the $n \times m$ matrix B of factor loadings and, associated with B, the matrices $\{\mu_i, V_i\}_{i=1}^g$. This is because for any invertible $m \times m$ matrix M we can transform

$$B^* = BM^{-1}, \qquad \mu_i^* = M\mu_i, \qquad V_i^* = MV_iM',$$

and retain the exact same data-generation process, since $B^*\mu_i^* = B\mu_i$ and $B^*V_i^*B^{*\prime} = BV_iB'$. Jöreskog (1969) was among the first to provide a rule to restrict elements of B in order to identify the matrix, but his rule did not in fact identify the matrix. Sufficient conditions we given by Algina (1980).

One approach follows Geweke and Zhou (1996). We assume that the loadings matrix B takes the form

$$B = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ b_{2,1} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ b_{m,1} & b_{m,2} & b_{m,3} & \dots & 1 \\ b_{m+1,1} & b_{m+1,2} & b_{m+1,3} & \dots & b_{m+1,m} \\ \vdots & \vdots & \vdots & & \vdots \\ b_{n,1} & b_{n,2} & b_{n,3} & \dots & b_{n,m} \end{pmatrix}.$$

This form clearly implies that B has full column-rank and it guarantees invariance under invertible transformations. The specification assumes that the first m rows of B are linearly

independent. This is not completely satisfactory. It is reasonable to assume that the m columns of B are linearly independent, and hence that there exists at least one set of m linearly independent rows (usually many sets), but it is not reasonable to assume that the *first* m rows are linearly independent.

One possible solution to this problem is to first locate a posterior mode without imposing identification on B, then examine if there is an $m \times m$ sub-matrix that is invertible, and then impose identification using this sub-matrix. This is not completely innocuous either, because the fact that this matrix is invertible at the mode does not imply that a posteriori there is no possibility that this matrix is singular.

In a non-identified model with an improper prior, the posterior distribution has a (high-dimensional) ridge that traces out an unbounded set of maximum points, and individual parameters have infinite variance. One may simply ignore this fact, simulate the posterior of a non-identified model, and then investigate the posterior for identified parameters (i.e. $\{m_i, W_i\}_{i=1}^g$). In this paper, however, we examine approximations to the posterior variance, and these approximations are not valid when there are likelihood ridges. We therefore impose identification by assuming the structure of B described above, cognizant of the caveats involved.

4. The score

We wish to obtain the first- and second derivatives of the log-likelihood $\log \mathcal{L}(\theta)$: the score vector and the Hessian matrix. First we need some more notation. For any matrix A, $\mathrm{vec}(A)$ denotes the vector which stacks the columns of A one underneath the other; when A is square, $\mathrm{vech}(A)$ is obtained from $\mathrm{vec}\, A$ by deleting all supra-diagonal elements of A; and $\mathrm{dg}(A)$ is the vector containing only the diagonal elements of A. We let $\widetilde{\psi} = \mathrm{dg}(\widetilde{\Psi})$, so that $\widetilde{\psi}$ contains the n diagonal components of $\widetilde{\Psi}$. Further, e_i denotes the $g \times 1$ vector all whose components are zero except the ith which is one. We let $\xi_* = (\xi_2, \ldots, \xi_g)'$ and $S_* = (0 : I_{g-1})$, and we recall that $\xi_1 = 1$. Then $\xi = e_1 + S_*'\xi_*$. Now let

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_g \end{pmatrix}, \qquad \upsilon = \begin{pmatrix} \mathrm{vech}(\widetilde{V}_1) \\ \vdots \\ \mathrm{vech}(\widetilde{V}_g) \end{pmatrix}, \tag{14}$$

so that the parameter vector can be written as $\theta = (\xi_*, \theta_*)$, where

$$\theta_* = (\mu, \upsilon, \mathrm{vec}\, B, \widetilde{\psi}). \tag{15}$$

It will also prove useful to introduce the weights

$$\bar{\pi}_{it} = \frac{\pi_i e^{-\lambda_{it}(\theta)/2}}{\sum_{j=1}^g \pi_j e^{-\lambda_{jt}(\theta)/2}}, \tag{16}$$

which can be interpreted as the posterior probabilities (after knowing the data at time t) corresponding to the prior probabilities π_i.

Proposition 1 (Score). *We have*

$$d \log \mathcal{L}_t = \sum_{i=1}^g \bar{\pi}_{it} \left(\frac{d\pi_i}{\pi_i} - \frac{d\lambda_{it}}{2} \right),$$

and hence

$$\frac{\partial \log \mathcal{L}_t}{\partial \xi_*} = \sum_{i=1}^{g} \bar{\pi}_{it} \alpha_i^{(1)}, \qquad \frac{\partial \log \mathcal{L}_t}{\partial \mu_i} = \bar{\pi}_{it} \alpha_{it}^{(2)}, \qquad \frac{\partial \log \mathcal{L}_t}{\partial \operatorname{vech} \widetilde{V}_i} = \bar{\pi}_{it} \alpha_{it}^{(3)},$$

$$\frac{\partial \log \mathcal{L}_t}{\partial \operatorname{vec} B} = \sum_{i=1}^{g} \bar{\pi}_{it} \alpha_{it}^{(4)}, \qquad \frac{\partial \log \mathcal{L}_t}{\partial \widetilde{\psi}_j} = \sum_{i=1}^{g} \bar{\pi}_{it} \alpha_{it}^{(5)},$$

where

$$\alpha_i^{(1)} = (2/\xi_i^2) S_* (\xi_i e_i - \pi_i \xi), \qquad \alpha_{it}^{(2)} = B' p_{it}, \qquad \alpha_{it}^{(3)} = -\operatorname{vech}(B' P_{it} B \widetilde{V}_i),$$

$$\alpha_{it}^{(4)} = \operatorname{vec}\left(p_{it} \mu_i' - P_{it} B V_i\right), \qquad \alpha_{it}^{(5)} = -\operatorname{dg}(\widetilde{\Psi} P_{it}),$$

and

$$P_{it} = S_t (S_t' W_i S_t)^{-1} S_t' - p_{it} p_{it}', \qquad p_{it} = S_t (S_t' W_i S_t)^{-1} (x_{(t)} - S_t' m_i).$$

Summing over t gives the required score vector.

Proof. Taking differentials in (8) we obtain

$$d\mathcal{L}_t = \sum_{i=1}^{g} \left((d\pi_i) e^{-\lambda_{it}/2} - \frac{1}{2} \pi_i e^{-\lambda_{it}/2} (d\lambda_{it}) \right) \tag{17}$$

and hence

$$d \log \mathcal{L}_t = \frac{d\mathcal{L}_t}{\mathcal{L}_t} = \sum_{i=1}^{g} \bar{\pi}_{it} \left(\frac{d\pi_i}{\pi_i} - \frac{d\lambda_{it}}{2} \right). \tag{18}$$

The first differentials of π_i and λ_{it} are given in Appendices A and B, respectively, from which we see that

$$d\pi_i/\pi_i = \alpha_i^{(1)'} d\xi_*, \tag{19}$$

and

$$-d\lambda_{it}/2 = \alpha_{it}^{(2)'} d\mu_i + \alpha_{it}^{(3)'} d \operatorname{vech}(\widetilde{V}_i) + \alpha_{it}^{(4)'} d \operatorname{vec} B + \alpha_{it}^{(5)'} d\widetilde{\psi}. \tag{20}$$

Since λ_{it} and $d\lambda_{it}$ do not depend on μ_j and $\operatorname{vech}(\widetilde{V}_j)$ $(j \neq i)$, we obtain

$$-d\lambda_{it}/2 = a_{it}' d\theta_*, \qquad a_{it} = \begin{pmatrix} e_i \otimes \alpha_{it}^{(2)} \\ e_i \otimes \alpha_{it}^{(3)} \\ \alpha_{it}^{(4)} \\ \alpha_{it}^{(5)} \end{pmatrix}. \tag{21}$$

Inserting (19) and (21) into (18), the result follows. □

5. The Hessian

The Hessian matrix is more difficult. Our starting point is the equality

$$d^2 \log \mathcal{L}_t = -\left(\frac{d\mathcal{L}_t}{\mathcal{L}_t} \right)^2 + \frac{d^2 \mathcal{L}_t}{\mathcal{L}_t}. \tag{22}$$

Now, $d\mathcal{L}_t/\mathcal{L}_t$ is given in (18). Differentiating (17) gives

$$\frac{d^2\mathcal{L}_t}{\mathcal{L}_t} = \sum_{i=1}^{g} \bar{\pi}_{it} \left(\frac{d^2\pi_i}{\pi_i} - \frac{(d\pi_i)(d\lambda_{it})}{\pi_i} + \frac{(d\lambda_{it})^2}{4} - \frac{d^2\lambda_{it}}{2} \right). \tag{23}$$

We know $d\pi_i/\pi_i$ and $-d\lambda_{it}/2$ from (19) and (20). Hence we need to find $d^2\pi_i/\pi_i$ and $d^2\lambda_{it}$. This is achieved in Appendices A and B. In fact,

$$d^2\pi_i/\pi_i = -(d\xi_*)' A_i^{(11)} (d\xi_*) \tag{24}$$

and

$$d^2\lambda_{it}/2 = (d\theta_*)' A_{it} (d\theta_*), \tag{25}$$

where $A_i^{(11)}$ is defined in Appendix A and A_{it} in Appendix B. Thus we obtain the following result.

Proposition 2 (Hessian). *The Hessian matrix H is given by $H = \sum_t H_t$, where*

$$-H_t = \begin{pmatrix} \bar{A}_t^{(11)} & -\bar{A}_t^{(1*)} \\ -\bar{A}_t^{(1*)'} & \bar{A}_t^{(**)} \end{pmatrix} + \begin{pmatrix} \bar{a}_t^{(1)} \bar{a}_t^{(1)'} & \bar{a}_t^{(1)} \bar{a}_t^{(*)'} \\ \bar{a}_t^{(*)} \bar{a}_t^{(1)'} & \bar{a}_t^{(*)} \bar{a}_t^{(*)'} \end{pmatrix},$$

with

$$\bar{A}_t^{(11)} = \sum_{i=1}^{g} \bar{\pi}_{it} A_i^{(11)}, \qquad \bar{A}_t^{(1*)} = \sum_{i=1}^{g} \bar{\pi}_{it} \alpha_i^{(1)} a_{it}', \qquad \bar{A}_t^{(**)} = \sum_{i=1}^{g} \bar{\pi}_{it} (A_{it} - a_{it} a_{it}')$$

and

$$\bar{a}_t^{(1)} = \sum_{i=1}^{g} \bar{\pi}_{it} \alpha_i^{(1)}, \qquad \bar{a}_t^{(*)} = \sum_{i=1}^{g} \bar{\pi}_{it} a_{it}.$$

Proof. Given all the ingredients presented above, we have

$$\begin{aligned} \frac{d\mathcal{L}_t}{\mathcal{L}_t} &= \sum_{i=1}^{g} \bar{\pi}_{it} \frac{d\pi_i}{\pi_i} - \sum_{i=1}^{g} \bar{\pi}_{it} \frac{d\lambda_{it}}{2} \\ &= \sum_{i=1}^{g} \bar{\pi}_{it} \alpha_i^{(1)'} d\xi_* + \sum_{i=1}^{g} \bar{\pi}_{it} a_{it}' d\theta_* = \bar{a}_t^{(1)'} d\xi_* + \bar{a}_t^{(*)'} d\theta_* \end{aligned} \tag{26}$$

and

$$\begin{aligned} \frac{d^2\mathcal{L}_t}{\mathcal{L}_t} &= \sum_{i=1}^{g} \bar{\pi}_{it} \left(\frac{d^2\pi_i}{\pi_i} - \frac{(d\pi_i)(d\lambda_{it})}{\pi_i} + \frac{(d\lambda_{it})^2}{4} - \frac{d^2\lambda_{it}}{2} \right) \\ &= -\sum_{i=1}^{g} \bar{\pi}_{it} (d\xi_*)' A_i^{(11)} (d\xi_*) + 2 \sum_{i=1}^{g} \bar{\pi}_{it} (d\xi_*)' \alpha_i^{(1)} a_{it}' (d\theta_*) \\ &\quad + \sum_{i=1}^{g} \bar{\pi}_{it} (d\theta_*)' a_{it} a_{it}' (d\theta_*) - \sum_{i=1}^{g} \bar{\pi}_{it} (d\theta_*)' A_{it} (d\theta_*) \\ &= -(d\xi_*)' \bar{A}_t^{(11)} (d\xi_*) + 2(d\xi_*)' \bar{A}_t^{(1*)} (d\theta_*) - (d\theta_*)' \bar{A}_t^{(**)} (d\theta_*). \end{aligned} \tag{27}$$

Hence,

$$
\begin{aligned}
-d^2 \log \mathcal{L}_t &= (d\mathcal{L}_t/\mathcal{L}_t)^2 - d^2\mathcal{L}_t/\mathcal{L}_t \\
&= (d\xi_*)'\bar{a}_t^{(1)}\bar{a}_t^{(1)'}(d\xi_*) + 2(d\xi_*)\bar{a}_t^{(1)}\bar{a}_t^{(*)'}(d\theta_*) + (d\theta_*)\bar{a}_t^{(*)}\bar{a}_t^{(*)'}(d\theta_*) \\
&\quad + (d\xi_*)'\bar{A}_t^{(11)}(d\xi_*) - 2(d\xi_*)'\bar{A}_t^{(1*)}(d\theta_*) + (d\theta_*)'\bar{A}_t^{(**)}(d\theta_*) \\
&= \begin{pmatrix} d\xi_* \\ d\theta_* \end{pmatrix}' \begin{pmatrix} \bar{A}_t^{(11)} + \bar{a}_t^{(1)}\bar{a}_t^{(1)'} & -\bar{A}_t^{(1*)} + \bar{a}_t^{(1)}\bar{a}_t^{(*)'} \\ -\bar{A}_t^{(1*)'} + \bar{a}_t^{(*)}\bar{a}_t^{(1)'} & \bar{A}_t^{(**)} + \bar{a}_t^{(*)}\bar{a}_t^{(*)'} \end{pmatrix} \begin{pmatrix} d\xi_* \\ d\theta_* \end{pmatrix},
\end{aligned}
\tag{28}
$$

and the result follows. □

6. Empirical illustration

We demonstrate the model and the usefulness of explicit Hessian formulae with two sets of four examples. In each example we use randomly generated parameters and data. We choose $T = 200$ as our sample size and $g = 2$ as the number of mixture components for each model, but the dimensions vary, as shown in Table 1. In Models 1 the ratio of dimension reduction (m/n) is 1/2 for $n = 4, 6, 8, 10$. One could argue, however, that this ratio is typically smaller in the commonly used latent factor mixture models, and hence we also study Models 2 by fixing $m = 5$ and letting $n = 20, 30, 40, 50$, respectively.

We let 15% of the data be randomly missing. We set $\kappa_1 = \kappa_2 = 10^{-10}$, $\xi = (1, 2, \ldots, g)'$, and $\widetilde{\psi}_i = 0.5$ $(i = 1, \ldots, n)$. For the free elements of B we use a truncated standard normal distribution, excluding parameters that are less than 0.1 in absolute value. For the elements of μ_i and $\widetilde{V}_i$ $(i = 1, \ldots, g)$ we use draws from a $\mathrm{N}(0, i^2)$ distribution, with the same truncation as for B. We find the posterior mode through numerical optimization, using the Broyden-Fletcher-Goldfarb-Shanno algorithm as provided by the `optim` function in `R`. The optimization step takes 8.4 seconds (Model 1a) to 7 minutes (Model 2d) to complete.

Next we calculate two approximations to the posterior variance; see Berger (1985, pp. 224–225). The first is the information matrix:

$$IM = \mathrm{E}_{X|\bar{\theta}}\,[q(X|\bar{\theta})\, q'(X|\bar{\theta})],$$

where $q(X|\bar{\theta})$ denotes the score vector evaluated at the mode $\bar{\theta}$ for random data X. We calculate this expectation by simulating random data 20,000 times, which takes about 35 minutes. The second approximation is the negative of the Hessian:

$$-H[(X_{obs}|\bar{\theta})],$$

evaluated at the observed data X_{obs}, also at the mode. Our variance approximations are the inverses of these two matrices. It is also possible to numerically approximate the Hessian matrix. However, except for models of fairly small dimension, the computational time is prohibitive. An analytical Hessian is thus both more accurate and faster to compute.

We estimate the model using 100,000 draws of a random walk Metropolis-Hastings sampler, with proposal variance proportional to the inverse Hessian. At iteration r, the

Table 1. Dimensions of the eight models.

	1a	1b	1c	1d	2a	2b	2c	2d
n	4	6	8	10	20	30	40	50
m	2	3	4	5	5	5	5	5

Table 2. Comparison of posterior variance with two approximations, models 1.

		Model 1a	Model 1b	Model 1c	Model 1d
Standard deviations	$-H$	0.0924	0.0768	0.1615	0.1467
	IM	0.0944	0.0959	0.6505	0.2315
	ratio	0.98	0.80	0.25	0.63
Correlations	$-H$	0.0007	0.0009	0.0023	0.0021
	IM	0.0014	0.0016	0.0030	0.0053
	ratio	0.50	0.56	0.77	0.40

Table 3. Comparison of posterior variance with two approximations, models 2.

		Model 2a	Model 2b	Model 2c	Model 2d
Standard deviations	$-H$	0.1565	0.1313	0.0768	0.1100
	IM	0.3025	0.2255	0.1182	0.1411
	ratio	0.52	0.58	0.65	0.78
Correlations	$-H$	0.0027	0.0011	0.0006	0.0006
	IM	0.0024	0.0013	0.0008	0.0006
	ratio	1.10	0.83	0.76	0.91

proposal parameter vector is drawn according to $\theta^{prop} = \theta^r + \eta$, where $\eta \sim \mathrm{N}(0, -\gamma H^{-1})$. The appropriate value of γ depends on the specific application. In our application we set $\gamma = 0.10$ for Models 1 and $\gamma = 0.01$ for Models 2. The posterior sampler takes between 80 and 269 minutes to complete.

Next we compare the estimated posterior variance matrix with our two approximations. We distinguish between two aspects of the variance matrix: the standard deviations (the square roots of the diagonal elements), and the correlations (the off-diagonal elements, but transformed from covariances to correlations).

Tables 2 and 3 present the results. The standard deviations are scale-dependent, and hence we consider the average absolute log-differences between the approximation and the true posterior. The correlations are scale-independent, and hence we consider the average absolute differences. The ratios between these averaged differences are well below one for all (except one) of the sixteen reported cases, in fact about 0.65 (on average) for the standard deviations and 0.73 for the correlations. This means that the Hessian approximation is about one and a half times as accurate as the Information Matrix approximation.

7. Conclusions

Within the class of normal latent factor mixture models we presented a method for estimating the posterior distribution of the parameters, and we derived analytical expressions for the score vector and the Hessian matrix of the posterior kernel. The latter derivation is the main theoretical contribution of the paper.

We showed that the Hessian is important in approximating the posterior variance. The expected outer product of the score vector and the negative of the inverse Hessian matrix provide two approximations of the posterior variance. The first only uses the score, while the second requires the Hessian. Our numerical examples suggest that the Hessian approximation is about one and a half times as accurate as the Information Matrix approximation.

Acknowledgements

The authors thank the referee for his/her constructive comments, and Eveline de Jong for her loving and critical support.

References

Algina, J. 1980. A note on the identification in the oblique and orthogonal factor analysis models. *Psychometrika* 45:393–396.

Andrews, J. L., and P. D. McNicholas. 2011. Extending mixtures of multivariate *t*-factor analyzers. *Statistics and Computing* 21:361–373.

Baek, J., G. J. McLachlan, and L. K. Flack. 2010. Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualisation of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32:1298–1309.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. New York: Springer.

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

Frühwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer.

Geweke, J. F., and G. Zhou. 1996. Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies* 9:557–587.

Ghahramani, Z., and G. E. Hinton. 1996. The EM algorithm for mixtures of factor analyzers, Technical Report CRG-TR-96-1, University of Toronto.

Jöreskog, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34:183–202.

Magnus, G. 2016. The Student-*t* latent factor mixture model with an application to global asset returns, mimeo.

Magnus, J. R. 1988. *Linear Structures*. Griffin's Statistical Monographs and Courses, No. 42, New York: Oxford University Press.

Magnus, J. R., and H. Neudecker. 1988. *Matrix Differential Calculus with Applications in Statistics and Econometrics*, First revision 1991, Second edition 1999. Chichester/New York: John Wiley.

McLachlan, G. J., R. W. Bean, and L. Ben-Tovim Jones. 2007. Extension of the mixture of factor analyzers model to incorporate the multivariate *t* distribution. *Computational Statistics & Data Analysis* 51:5327–5338.

McLachlan, G. J., S.-K. Ng, and R. W. Bean. 2006. Robust cluster analysis via mixture models. *Australian Journal of Statistics* 35:157–174.

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

McLachlan, G. J., D. Peel, and R. W. Bean. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* 41:379–388.

Montanari, A., and C. Viroli. 2010. Heteroscedastic factor mixture analysis. *Statistical Modelling* 10:441–460.

Murray, P. M., R. P. Browne, and P. D. McNicholas. 2014. Mixtures of skew-*t* factor analyzers. *Computational Statistics & Data Analysis* 77:326–335.

Rossi, P. E. 2014. *Bayesian Non- and Semi-Parametric Methods and Applications*. The Econometric and Tinbergen Institutes Lectures. Princeton, NJ: Princeton University Press.

Stephens, M. 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society* 62:795–809.

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

Appendix A: Derivatives of π_i

Writing (10) as

$$\pi_i = (\xi'\xi)^{-1}(e_i'\xi)^2,$$

we obtain

$$\begin{aligned} d\pi_i &= -(\xi'\xi)^{-2}(d\xi'\xi)(e_i'\xi)^2 + 2(\xi'\xi)^{-1}(e_i'\xi)(e_i'd\xi) \\ &= 2(\xi'\xi)^{-1}(e_i'\xi)(e_i'd\xi) - 2(\xi'\xi)^{-2}(e_i'\xi)^2(\xi'd\xi) \end{aligned}$$

and

$$\begin{aligned} d^2\pi_i &= -8(\xi'\xi)^{-2}(\xi' d\xi)(e_i'\xi)(e_i' d\xi) + 2(\xi'\xi)^{-1}(e_i' d\xi)(e_i' d\xi) \\ &\quad + 8(\xi'\xi)^{-3}(\xi' d\xi)(e_i'\xi)^2(\xi' d\xi) - 2(\xi'\xi)^{-2}(e_i'\xi)^2(d\xi)'(d\xi) \\ &= -4(d\xi)'(\xi'\xi)^{-2}\xi_i(\xi e_i' + e_i\xi')(d\xi) + 2(d\xi)'(\xi'\xi)^{-1}e_i e_i'(d\xi) \\ &\quad + 8(d\xi)'(\xi'\xi)^{-3}\xi_i^2\xi\xi'(d\xi) - 2(d\xi)'(\xi'\xi)^{-2}\xi_i^2(d\xi). \end{aligned}$$

This implies that

$$\frac{d\pi_i}{\pi_i} = \frac{2}{\xi_i^2}\left(\xi_i e_i' - \pi_i\xi'\right)d\xi = \alpha_i^{(1)'}d\xi_*,$$

where $\alpha_i^{(1)}$ is defined in Proposition 1, and

$$\frac{d^2\pi_i}{\pi_i} = -(d\xi_*)'A_i^{(11)}(d\xi_*),$$

where

$$A_i^{(11)} = \frac{2S_*\left[\xi_i^2 I_g + 2\xi_i(\xi e_i' + e_i\xi') - 4\pi_i\xi\xi' - (\xi'\xi)e_i e_i'\right]S_*'}{\xi_i^2(\xi'\xi)}.$$

Appendix B: Derivatives of λ_{it}

With λ_{it} given in (9), using the definitions of P_{it} and p_{it} in Proposition 1, and the matrix differential theory of Magnus and Neudecker (1988), we obtain

$$d\lambda_{it} = \operatorname{tr} P_{it}(dW_i) - 2p_{it}'(dm_i).$$

Also, given (5), we find the differentials of m_i and W_i as

$$dm_i = (dB)\mu_i + B(d\mu_i)$$

and

$$\begin{aligned} dW_i &= (dB)V_iB' + B(dV_i)B' + BV_i(dB)' + d\Psi \\ &= (dB)V_iB' + B(d\widetilde{V}_i)\widetilde{V}_i'B' + B\widetilde{V}_i(d\widetilde{V}_i)'B' + BV_i(dB)' + 2\widetilde{\Psi}(d\widetilde{\Psi}). \end{aligned}$$

Combining terms gives

$$\begin{aligned} -d\lambda_{it}/2 &= p_{it}'B(d\mu_i) - \operatorname{tr}\widetilde{V}_i'B'P_{it}B(d\widetilde{V}_i) \\ &\quad + \operatorname{tr}\left(\mu_i p_{it}' - V_iB'P_{it}\right)(dB) - \operatorname{tr}P_{it}\widetilde{\Psi}(d\widetilde{\Psi}) \\ &= \alpha_{it}^{(2)'}d\mu_i + \alpha_{it}^{(3)'}d\operatorname{vech}(\widetilde{V}_i) + \alpha_{it}^{(4)'}d\operatorname{vec}B + \alpha_{it}^{(5)'}d\widetilde{\psi}, \end{aligned}$$

where $\alpha_{it}^{(2)'}$, $\alpha_{it}^{(3)}$, $\alpha_{it}^{(4)}$, and $\alpha_{it}^{(5)}$ are defined in Proposition 1, and we have used the facts that $\operatorname{tr}A'd\widetilde{V}_i = (\operatorname{vech}A)'(\operatorname{vech}d\widetilde{V}_i)$ and $\operatorname{tr}A'd\widetilde{\Psi} = (\operatorname{dg}A)'(\operatorname{dg}d\widetilde{\Psi})$ for any square matrix A.

The second differential of λ_{it} is more difficult. Let L_m denotes the $\frac{1}{2}m(m+1)\times m^2$ 'elimination' matrix defined by the property that $L_m'\operatorname{vech}(A) = \operatorname{vec}A$ for every lower triangular $m\times m$ matrix A, and Δ_n the $n\times n^2$ matrix defined by the property that $\Delta_n'\operatorname{dg}(A) = \operatorname{vec}A$ for every diagonal $n\times n$ matrix A; see Magnus (1988). We shall prove the following proposition.

Proposition 3. *The symmetric* $(N-g+1)\times(N-g+1)$ *matrix* A_{it} *defined implicitly by*

$$d^2\lambda_{it}/2 = (d\theta_*)'A_{it}(d\theta_*),$$

takes the form

$$A_{it} = \begin{pmatrix} e_i e_i' \otimes A_{it}^{(22)} & e_i e_i' \otimes A_{it}^{(23)} L_m' & e_i \otimes A_{it}^{(24)} & e_i \otimes A_{it}^{(25)} \Delta_n' \\ e_i e_i' \otimes L_m A_{it}^{(32)} & e_i e_i' \otimes L_m A_{it}^{(33)} L_m' & e_i \otimes L_m A_{it}^{(34)} & e_i \otimes L_m A_{it}^{(35)} \Delta_n' \\ e_i' \otimes A_{it}^{(42)} & e_i' \otimes A_{it}^{(43)} L_m' & A_{it}^{(44)} & A_{it}^{(45)} \Delta_n' \\ e_i' \otimes \Delta_n A_{it}^{(52)} & e_i' \otimes \Delta_n A_{it}^{(53)} L_m' & \Delta_n A_{it}^{(54)} & \Delta_n A_{it}^{(55)} \Delta_n' \end{pmatrix},$$

where the relevant blocks are given by

$$\begin{aligned}
A_{it}^{(22)} &= B' Q_{it} B, \\
A_{it}^{(23)} &= p_{it}' B \widetilde{V}_i \otimes B' Q_{it} B + B' Q_{it} B \widetilde{V}_i \otimes p_{it}' B, \\
A_{it}^{(24)} &= (\mu_i + V_i B' p_{it})' \otimes B' Q_{it} - (I_m - B' Q_{it} B V_i) \otimes p_{it}', \\
A_{it}^{(25)} &= 2 p_{it}' \otimes B' Q_{it} \widetilde{\Psi}, \\
A_{it}^{(33)} &= (I_m - \widetilde{V}_i' B' P_{it} B \widetilde{V}_i) \otimes B' Q_{it} B - (I_m - \widetilde{V}_i' B' Q_{it} B \widetilde{V}_i) \otimes B' p_{it} p_{it}' B \\
&\quad + \left(\widetilde{V}_i' B' p_{it} p_{it}' B \otimes B' Q_{it} B \widetilde{V}_i + \widetilde{V}_i' B' Q_{it} B \otimes B' p_{it} p_{it}' B \widetilde{V}_i \right) K_{mm} \\
&\quad - \left(\widetilde{V}_i' B' Q_{it} B \otimes B' Q_{it} B \widetilde{V}_i \right) K_{mm}, \\
A_{it}^{(34)} &= (\widetilde{V}_i' - \widetilde{V}_i' B' Q_{it} B V_i) \otimes B' P_{it} + \widetilde{V}_i' B' p_{it} (\mu_i + V_i B' p_{it})' \otimes B' Q_{it} \\
&\quad + \left(\widetilde{V}_i' B' P_{it} \otimes (I_m - B' Q_{it} B V_i) + \widetilde{V}_i' B' Q_{it} \otimes B' p_{it} (\mu_i + V_i B' p_{it})' \right) K_{nm}, \\
A_{it}^{(35)} &= 2 \widetilde{V}_i' B' Q_{it} \widetilde{\Psi} \otimes B' p_{it} p_{it}' - 2 \widetilde{V}_i' B' P_{it} \otimes B' Q_{it} \widetilde{\Psi}, \\
A_{it}^{(44)} &= (\mu_i + V_i B' p_{it})(\mu_i - V_i B' p_{it})' \otimes Q_{it} + (V_i - V_i B' Q_{it} B V_i) \otimes P_{it} \\
&\quad + \left((\mu_i + V_i B' p_{it}) p_{it}' \otimes Q_{it} B V_i + V_i B' Q_{it} \otimes p_{it} (\mu_i + V_i B' p_{it})' \right) K_{nm} \\
&\quad - (V_i B' Q_{it} \otimes Q_{it} B V_i) K_{nm}, \\
A_{it}^{(45)} &= 2 (\mu_i + V_i B' p_{it}) p_{it}' \otimes Q_{it} \widetilde{\Psi} - 2 V_i B' Q_{it} \otimes P_{it} \widetilde{\Psi},
\end{aligned}$$

and

$$A_{it}^{(55)} = P_{it} \otimes I_n + 2 Q_{it} \otimes \widetilde{\Psi} p_{it} p_{it}' \widetilde{\Psi} - 2 P_{it} \otimes \widetilde{\Psi} Q_{it} \widetilde{\Psi},$$

where $Q_{it} = S_t (S_t' W_i S_t)^{-1} S_t' = P_{it} + p_{it} p_{it}'$ *and* K_{nm} *denotes the* $nm \times nm$ *'commutation' matrix defined by the property that* $K_{nm} \operatorname{vec} A = \operatorname{vec} A'$ *for every* $n \times m$ *matrix* A.

Proof. Since

$$dm_i = (dB)\mu_i + B(d\mu_i)$$

and

$$dW_i = (dB) V_i B' + B(d\widetilde{V}_i) \widetilde{V}_i' B' + B \widetilde{V}_i (d\widetilde{V}_i)' B' + B V_i (dB)' + 2 \widetilde{\Psi} (d\widetilde{\Psi}),$$

we obtain

$$\begin{aligned}
dQ_{it} &= -Q_{it} (dW_i) Q_{it} \\
&= -Q_{it} B (d\widetilde{V}_i) \widetilde{V}_i' B' Q_{it} - Q_{it} B \widetilde{V}_i (d\widetilde{V}_i)' B' Q_{it} - Q_{it} (dB) V_i B' Q_{it} \\
&\quad - Q_{it} B V_i (dB)' Q_{it} - 2 Q_{it} \widetilde{\Psi} (d\widetilde{\Psi}) Q_{it}, \\
dp_{it} &= -Q_{it} (dW_i) p_{it} - Q_{it} (dm_i) \\
&= -Q_{it} B (d\mu_i) - Q_{it} B (d\widetilde{V}_i) \widetilde{V}_i' B' p_{it} - Q_{it} B \widetilde{V}_i (d\widetilde{V}_i)' B' p_{it} \\
&\quad - Q_{it} (dB)(\mu_i + V_i B' p_{it}) - Q_{it} B V_i (dB)' p_{it} - 2 Q_{it} \widetilde{\Psi} (d\widetilde{\Psi}) p_{it},
\end{aligned}$$

and

$$\begin{aligned}
dP_{it} &= dQ_{it} - (dp_{it})p_{it}' - p_{it}(dp_{it})' \\
&= Q_{it}B(d\mu_i)p_{it}' + p_{it}(d\mu_i)'B'Q_{it} - Q_{it}B(d\widetilde{V}_i)\widetilde{V}_i'B'P_{it} \\
&\quad + p_{it}p_{it}'B(d\widetilde{V}_i)\widetilde{V}_i'B'Q_{it} - P_{it}B\widetilde{V}_i(d\widetilde{V}_i)'B'Q_{it} + Q_{it}B\widetilde{V}_i(d\widetilde{V}_i)'B'p_{it}p_{it}' \\
&\quad - P_{it}(dB)V_iB'Q_{it} + Q_{it}(dB)(\mu_i + V_iB'p_{it})p_{it}' - Q_{it}BV_i(dB)'P_{it} \\
&\quad + p_{it}(\mu_i + V_iB'p_{it})'(dB)'Q_{it} - 2Q_{it}\widetilde{\Psi}(d\widetilde{\Psi})P_{it} + 2p_{it}p_{it}'(d\widetilde{\Psi})\widetilde{\Psi}Q_{it}.
\end{aligned}$$

We then prove the proposition in four steps. In step 1 we obtain

$$\begin{aligned}
-d\alpha_{it}^{(2)} &= -d(B'p_{it}) = -(dB)'p_{it} - B'(dp_{it}) \\
&= B'Q_{it}B(d\mu_i) + B'Q_{it}B(d\widetilde{V}_i)\widetilde{V}_i'B'p_{it} + B'Q_{it}B\widetilde{V}_i(d\widetilde{V}_i)'B'p_{it} \\
&\quad + B'Q_{it}(dB)(\mu_i + V_iB'p_{it}) + B'Q_{it}BV_i(dB)'p_{it} - (dB)'p_{it} \\
&\quad + 2B'Q_{it}\widetilde{\Psi}(d\widetilde{\Psi})p_{it} \\
&= A_{it}^{(22)}d\mu_i + A_{it}^{(23)}d\operatorname{vec}\widetilde{V}_i + A_{it}^{(24)}d\operatorname{vec}B + A_{it}^{(25)}d\operatorname{vec}\widetilde{\Psi}.
\end{aligned}$$

In step 2 we note that

$$-d\alpha_{it}^{(3)} = d\operatorname{vech}(B'P_{it}B\widetilde{V}_i) = L_m d\operatorname{vec}B'P_{it}B\widetilde{V}_i.$$

Then, setting $d\mu_i = 0$, we have

$$\begin{aligned}
d(B'P_{it}B\widetilde{V}_i) &= (dB)'P_{it}B\widetilde{V}_i + B'(dP_{it})B\widetilde{V}_i + B'P_{it}(dB)\widetilde{V}_i + B'P_{it}B(d\widetilde{V}_i) \\
&= B'Q_{it}B(d\widetilde{V}_i)(I_m - \widetilde{V}_i'B'P_{it}B\widetilde{V}_i) - B'p_{it}p_{it}'B(d\widetilde{V}_i)(I_m - \widetilde{V}_i'B'Q_{it}B\widetilde{V}_i) \\
&\quad + B'p_{it}p_{it}'B\widetilde{V}_i(d\widetilde{V}_i)'B'Q_{it}B\widetilde{V}_i + B'Q_{it}B\widetilde{V}_i(d\widetilde{V}_i)'B'p_{it}p_{it}'B\widetilde{V}_i \\
&\quad - B'Q_{it}B\widetilde{V}_i(d\widetilde{V}_i)'B'Q_{it}B\widetilde{V}_i \\
&\quad + B'P_{it}(dB)(\widetilde{V}_i - V_iB'Q_{it}B\widetilde{V}_i) + B'Q_{it}(dB)(\mu_i + V_iB'p_{it})p_{it}'B\widetilde{V}_i \\
&\quad + (I_m - B'Q_{it}BV_i)(dB)'P_{it}B\widetilde{V}_i + B'p_{it}(\mu_i + V_iB'p_{it})'(dB)'Q_{it}B\widetilde{V}_i \\
&\quad - 2B'Q_{it}\widetilde{\Psi}(d\widetilde{\Psi})P_{it}B\widetilde{V}_i + 2B'p_{it}p_{it}'(d\widetilde{\Psi})\widetilde{\Psi}Q_{it}B\widetilde{V}_i,
\end{aligned}$$

which implies that

$$d\operatorname{vec}B'P_{it}B\widetilde{V}_i = A_{it}^{(33)}d\operatorname{vec}\widetilde{V}_i + A_{it}^{(34)}d\operatorname{vec}B + A_{it}^{(35)}d\operatorname{vec}\widetilde{\Psi}.$$

The starting point for step 3 is

$$-d\alpha_{it}^{(4)} = -d\operatorname{vec}(p_{it}\mu_i' - P_{it}BV_i).$$

Setting $d\mu_i$ and $d\widetilde{V}_i$ both equal to zero, we have

$$\begin{aligned}
d(p_{it}\mu_i' - P_{it}BV_i) &= d(p_{it})\mu_i' - (dP_{it})BV_i - P_{it}(dB)V_i \\
&= -Q_{it}(dB)(\mu_i + V_iB'p_{it})(\mu_i' - p_{it}'BV_i) - P_{it}(dB)(V_i - V_iB'Q_{it}BV_i) \\
&\quad - Q_{it}BV_i(dB)'p_{it}(\mu_i + V_iB'p_{it})' - p_{it}(\mu_i + V_iB'p_{it})'(dB)'Q_{it}BV_i \\
&\quad + Q_{it}BV_i(dB)'Q_{it}BV_i + 2P_{it}\widetilde{\Psi}(d\widetilde{\Psi})Q_{it}BV_i - 2Q_{it}\widetilde{\Psi}(d\widetilde{\Psi})p_{it}\left(\mu_i + V_iB'p_{it}\right)',
\end{aligned}$$

and hence

$$\begin{aligned}
&-d\operatorname{vec}(p_{it}\mu_i' - P_{it}BV_i) \\
&= \left((\mu_i + V_iB'p_{it})(\mu_i - V_iB'p_{it})' \otimes Q_{it} + (V_i - V_iB'Q_{it}BV_i) \otimes P_{it}\right)d\operatorname{vec}B \\
&\quad + \left((\mu_i + V_iB'p_{it})p_{it}' \otimes Q_{it}BV_i + V_iB'Q_{it} \otimes p_{it}(\mu_i + V_iB'p_{it})'\right)K_{nm}d\operatorname{vec}B \\
&\quad - (V_iB'Q_{it} \otimes Q_{it}BV_i)K_{nm}d\operatorname{vec}B
\end{aligned}$$

$$+2\left((\mu_i + V_i B' p_{it}) p_{it}' \otimes Q_{it}\widetilde{\Psi} - V_i B' Q_{it} \otimes P_{it}\widetilde{\Psi}\right) d\,\mathrm{vec}\,\widetilde{\Psi}$$
$$= A_{it}^{(44)} d\,\mathrm{vec}\,B + A_{it}^{(45)} d\,\mathrm{vec}\,\widetilde{\Psi}.$$

Finally, in step 4, we set $d\mu_i$, $d\widetilde{V}_i$, and dB all equal to zero. Then,

$$-d\alpha_{it}^{(5)} = \mathrm{dg}\, d(\widetilde{\Psi} P_{it})$$

and

$$\begin{aligned} d(\widetilde{\Psi} P_{it}) &= (d\widetilde{\Psi}) P_{it} + \widetilde{\Psi}(dP_{it}) \\ &= (d\widetilde{\Psi}) P_{it} - 2\widetilde{\Psi} Q_{it} \widetilde{\Psi} (d\widetilde{\Psi}) P_{it} + 2\widetilde{\Psi} p_{it} p_{it}' \widetilde{\Psi} (d\widetilde{\Psi}) Q_{it} \end{aligned}$$

so that

$$d\,\mathrm{vec}(\widetilde{\Psi} P_{it}) = A_{it}^{(55)} d\,\mathrm{vec}\,\widetilde{\Psi}.$$

This concludes the proof. □