Biostatistics (2010), **11**, 2, *pp.* 317–336
doi:10.1093/biostatistics/kxp062
Advance Access publication on January 27, 2010

Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-*t* distributions

SYLVIA FRÜHWIRTH-SCHNATTER*

Department of Applied Statistics and Econometrics,
Johannes Kepler Universität Linz, A-1040 Linz, Austria
sylvia.fruehwirth-schnatter@jku.at

SAUMYADIPTA PYNE

Department of Medical Oncology, Dana-Farber Cancer Institute,
Harvard Medical School, Boston, MA 02115, USA

SUMMARY

Skew-normal and skew-*t* distributions have proved to be useful for capturing skewness and kurtosis in data directly without transformation. Recently, finite mixtures of such distributions have been considered as a more general tool for handling heterogeneous data involving asymmetric behaviors across subpopulations. We consider such mixture models for both univariate as well as multivariate data. This allows robust modeling of high-dimensional multimodal and asymmetric data generated by popular biotechnological platforms such as flow cytometry.

We develop Bayesian inference based on data augmentation and Markov chain Monte Carlo (MCMC) sampling. In addition to the latent allocations, data augmentation is based on a stochastic representation of the skew-normal distribution in terms of a random-effects model with truncated normal random effects. For finite mixtures of skew normals, this leads to a Gibbs sampling scheme that draws from standard densities only. This MCMC scheme is extended to mixtures of skew-*t* distributions based on representing the skew-*t* distribution as a scale mixture of skew normals.

As an important application of our new method, we demonstrate how it provides a new computational framework for automated analysis of high-dimensional flow cytometric data. Using multivariate skew-normal and skew-*t* mixture models, we could model non-Gaussian cell populations rigorously and directly without transformation or projection to lower dimensions.

Keywords: Flow cytometry; Gibbs sampling; Kurtosis; Markov chain Monte Carlo; Skewness; Stochastic representation.

1. INTRODUCTION

When modeling empirical univariate or multivariate data $\mathbf{y}_1, \dots, \mathbf{y}_N$ that exhibit multimodality, skewness, or excess kurtosis, it is often assumed that the data are independent realizations of a random variable $\mathbf{Y}$

*To whom correspondence should be addressed.

from a finite-mixture distribution. This leads to the standard finite-mixture model considered, for example, in McLachlan and Peel (2000) and Frühwirth-Schnatter (2006). An important special case of such a model is a mixture of normal distributions which allows an arbitrarily close modeling of any distribution by increasing the number of components. The flexibility, however, causes problems when such a model is used in a clustering context because several normal distributions may be necessary to capture skewness and kurtosis of a single cluster, thus leading to wrong inference about the number of clusters in the data (Jasra *and others*, 2006). Similarly, in the context of supervised learning, groups of observations represented by asymmetrically distributed data can lead to wrong classification.

For illustration, we show in Figure 1, the histogram of the global cognition scores of 451 patients suffering from Alzheimer's disease (AD). These data will be analyzed in detail in Section 4.1. The left-hand side of Figure 1 shows the result of fitting a 3-component mixture of normal distributions which correspond to the optimal number of components as will be demonstrated in Section 4.1. Interestingly, the bimodality of the fitted mixture indicates the presence of 2 clusters, however, the normal mixture needs 2 components to fit the skewness present in the second cluster.

To address such practical issues formally, attention has shifted recently toward finite-mixture models where the component densities themselves capture skewness and excess kurtosis. Applications and case studies for modeling with skew distributions now include research areas, such as economics, finance, climatology, environmetrics, engineering, and biomedical sciences (Genton, 2004). On the other hand, for robustness against outliers in multimodal data, mixtures of Student-t distributions have been applied by Peel and McLachlan (2000) and Lin *and others* (2004) which allow for heavy tails of each component. Very recently, application of finite-mixture models have been to the univariate skew-normal (Lin, Lee, and Yen, 2007) and skew-t distribution (Jasra *and others*, 2006; Lin, Lee, and Hseih, 2007), to the univariate skew Student-t-normal distribution (Cabral *and others*, 2008), and to the multivariate skew-normal (Lin, 2009) and skew-t distribution (Lin, 2010; Pyne *and others*, 2009).

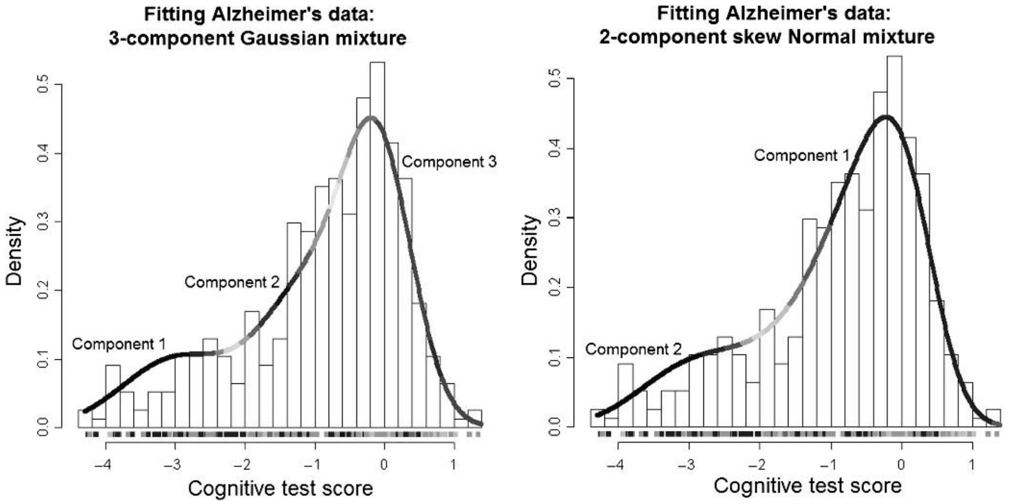

Fig. 1. Gaussian and skew-normal mixture modeling of AD data set. The histogram, common to both plots, shows the univariate cognition test scores of subjects in the data set. The "rugplot" common to both plots (it appears just below the x-axis in either plot) shows each subject's genotype. A darker point in the rug indicates more e4 alleles in a subject's genotype implying higher risk factor for AD. In the left-hand side plot, fitting of a 3-component Gaussian mixture is shown. In the right-hand side plot, fitting of a 2-component skew-normal mixture is shown.

Following this important work, we consider univariate as well as multivariate skew-normal and skew-*t* distributions as defined by Azzalini (1985, 1986), Azzalini and Dalla Valle (1996), and Azzalini and Capitanio (2003) as building blocks for a finite-mixture model. We apply our methodology to the (univariate) clinical data from AD introduced above and will show that the optimal mixture of skew-normal distributions needs only 2 components to fit the observed distribution, see the right-hand side of Figure 1. In addition, we consider clustering multivariate flow cytometric data from Graft versus Host Disease (GvHD). Flow cytometry is a biotechnological platform commonly used in immunology, cancer biology, and molecular biology. It is used to investigate expression of proteins on the surface and within every cell in a given sample with fluorophore-conjugated antibodies (or markers). Currently, up to 17 markers can be measured for each of the tens to hundreds of thousands of cells per sample (Perfetto *and others*, 2004), thus producing high-throughput high-dimensional data. In addition, flow cytometric data are often multimodal, skewed, and noisy. At present, the analysis of flow cytometric data analysis, which involves identification of cell populations, is done manually by projecting the data in 2D. Our Bayesian mixture modeling with multivariate skew distributions can allow automatic high-dimensional clustering to substitute the current slow and subjective manual approach to flow cytometric data analysis. As noted above, our model also allows the asymmetry in data to be modeled directly without the need for any transformation which might lead to imprecise inference about the number of clusters in a data set.

Although the extension from a standard to a skew finite-mixture model appears quite natural, the actual estimation results in a complex computational problem. Subsequently, we pursue a Bayesian approach using data augmentation and Markov chain Monte Carlo (MCMC). Toward this, we use a representation of the skew-normal and the skew-*t* distribution that combines the standard hierarchical representation of a finite-mixture model introduced in Diebolt and Robert (1994) with a stochastic representation of the skew-normal and the skew-*t* distribution in terms of a random-effects model with truncated normal random effects (Azzalini, 1986; Henze, 1986). After applying a suitable transformation of the component-specific parameters, this leads to a straightforward MCMC sampling scheme that involves a 2-block Gibbs sampler for finite mixtures both of univariate and multivariate skew-normal distributions. For finite mixtures of univariate and multivariate skew-*t* distributions, a third block has to be added that involves a Metropolis–Hastings step for the degrees of freedom and a Gibbs step for the latent scaling factors in the infinite-mixture representation of the skew-*t* distribution.

The rest of the paper is organized as follows. Section 2 reviews skew-normal and skew-*t* distributions. Section 3 introduces finite mixtures of such distributions and discusses Bayesian estimation using MCMC. Section 4 provides applications to clustering univariate clinical data from AD and multivariate cytometric data from GvHD.

2. Skew-normal and skew-*t* distributions

2.1 *The scalar skew-normal distribution*

A univariate random variable X follows a standard skew-normal distribution with skewness parameter α, $X \sim \mathcal{SN}(\alpha)$, if the density takes the form $p(x|\alpha) = 2\phi(x)\Phi(\alpha x)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the probability density function (pdf) and the cumulative distribution function (cdf) of the standard normal distribution. Evidently, for $\alpha = 0$, the standard normal $\mathcal{N}(0, 1)$ results. Choosing $\alpha \neq 0$ leads to a density with a skewness coefficient in $[-0.9953, 0.9953]$. The first systematic treatment of this density has been given by Azzalini (1985, 1986).

In our subsequent Bayesian analysis, we use the following stochastic representations of the skew-normal distribution (Azzalini, 1986; Henze, 1986). Let $Z \sim \mathcal{TN}_{[0,\infty)}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 1)$, independently, and let $\delta \in (-1, 1)$. The random variable X defined by

$$X = \delta Z + \sqrt{1 - \delta^2}\varepsilon \tag{2.1}$$

follows the standard skew-normal $\mathcal{SN}(\alpha)$ distribution with skewness parameter $\alpha = \delta/\sqrt{1-\delta^2}$. Thus, the skew-normal distribution may be seen as the superposition of a normal random variable with a latent truncated standard normal random effect.

The expectation and the variance of X are given by $E(X) = \sqrt{\frac{2}{\pi}}\delta$ and $V(X) = 1 - \frac{2}{\pi}\delta^2$. To adjust for an arbitrary location and scale, a location parameter $\xi \in \Re$ and a scale parameter $\omega \in \Re^+$ are introduced. The random variable $Y = \xi + \omega X$, where $X \sim \mathcal{SN}(\alpha)$, is said to follow the skew-normal distribution $\mathcal{SN}(\xi, \omega^2, \alpha)$. The density of this distribution reads:

$$f_{\mathcal{SN}}(y; \xi, \omega^2, \alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi(\alpha\omega^{-1}(y-\xi)). \tag{2.2}$$

A stochastic representation of the $\mathcal{SN}(\xi, \omega^2, \alpha)$ distribution is obtained by applying the affine transformation $Y = \xi + \omega X$ to (2.1):

$$Y = \xi + \omega\delta Z + \omega\sqrt{1-\delta^2}\varepsilon, \tag{2.3}$$

where $Z \sim \mathcal{TN}_{[0,\infty)}(0, 1)$ and $\varepsilon \sim \mathcal{N}(0, 1)$, independently and $\delta = \alpha/(\sqrt{1+\alpha^2})$.

2.2 *The multivariate skew-normal distribution*

A multivariate version of the skew-normal distribution has been defined in Azzalini and Dalla Valle (1996) by generalizing the stochastic representation (2.1). The r components of a multivariate random variable $\mathbf{X} = (X_1, \dots, X_r)' \in \Re^r$ are defined for $j = 1, \dots, r$ as $X_j = \delta_j Z + \sqrt{1-\delta_j^2}\varepsilon_j$, where $\delta_j \in (-1, 1)$, $Z \sim \mathcal{TN}_{[0,\infty)}(0, 1)$ as before, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_r)' \sim \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}})$ is independent of Z and multivariate normal with an arbitrary correlation matrix $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}$. Applying the affine transformation $\mathbf{Y} = \boldsymbol{\xi} + \boldsymbol{\omega}\mathbf{X}$ with location parameter $\boldsymbol{\xi} = (\xi_1, \dots, \xi_r)' \in \Re^r$ and diagonal scale matrix $\boldsymbol{\omega} = \text{Diag}(\omega_1, \dots, \omega_r)$ with $\omega_j > 0$ immediately leads to the stochastic representation

$$Y_j = \xi_j + \omega_j\delta_j Z + \omega_j\sqrt{1-\delta_j^2}\varepsilon_j. \tag{2.4}$$

The resulting distribution is called the basic multivariate skew-normal distribution, denoted by $\mathcal{SN}_r$ $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$, with density

$$f_{\mathcal{SN}}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = 2\phi_r(\mathbf{y}-\boldsymbol{\xi}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\mathbf{y}-\boldsymbol{\xi})), \tag{2.5}$$

where $\phi_r(\mathbf{x}; \boldsymbol{\Omega})$ is the pdf of the multivariate zero mean $\mathcal{N}_r(\mathbf{0}, \boldsymbol{\Omega})$ distribution and $\Phi(\cdot)$ is the cdf of the univariate $\mathcal{N}(0, 1)$ distribution. The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\Omega}$ are related to the parameters $\boldsymbol{\delta} = (\delta_1, \dots, \delta_r)'$, $\boldsymbol{\omega}$ and $\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}$ in the stochastic representation (2.4) through

$$\boldsymbol{\Omega} = \boldsymbol{\omega}\overline{\boldsymbol{\Omega}}\boldsymbol{\omega}, \quad \boldsymbol{\alpha} = \frac{1}{\sqrt{1-\boldsymbol{\delta}'\boldsymbol{\delta}}}\overline{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}, \tag{2.6}$$

where $\overline{\boldsymbol{\Omega}} = \boldsymbol{\Delta}\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}}\boldsymbol{\Delta} + \boldsymbol{\delta}\boldsymbol{\delta}'$ and $\boldsymbol{\Delta} = \text{Diag}\left(\sqrt{1-\delta_1^2}, \dots, \sqrt{1-\delta_r^2}\right)$. The matrix $\overline{\boldsymbol{\Omega}}$ is a correlation matrix because $\overline{\boldsymbol{\Omega}}_{jj} = (1-\delta_j^2)(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}})_{jj} + \delta_j^2 = 1$, thus $\boldsymbol{\Omega}_{jj} = \omega_j^2$.

Given the parameter $(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ of a $\mathcal{SN}_r(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ distribution, the parameters $(\boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}})$ in the stochastic representation (2.4) are obtained from

$$\boldsymbol{\delta} = \frac{1}{\sqrt{1+\boldsymbol{\alpha}'\overline{\boldsymbol{\Omega}}\boldsymbol{\alpha}}}\overline{\boldsymbol{\Omega}}\boldsymbol{\alpha}, \quad \boldsymbol{\Omega}_{\boldsymbol{\varepsilon}} = \boldsymbol{\Delta}^{-1}\overline{\boldsymbol{\Omega}}\boldsymbol{\Delta}^{-1} - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}', \tag{2.7}$$

where $\overline{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$ with $\boldsymbol{\omega} = \text{Diag}(\boldsymbol{\Omega})^{1/2}$ being a diagonal matrix obtained from the diagonal elements of $\boldsymbol{\Omega}$, $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_r)'$ with $\tilde{\alpha}_j = \delta_j/\sqrt{1-\delta_j^2}$ and $\boldsymbol{\Delta}$ is the same as above. The marginal distribution of Y_j is equal to the scalar skew normal $\mathcal{SN}(\xi_j, \omega_j^2, \tilde{\alpha}_j)$, hence,

$$E(\mathbf{Y}) = \boldsymbol{\xi} + \boldsymbol{\omega}\boldsymbol{\delta}\sqrt{\frac{2}{\pi}}. \tag{2.8}$$

For alternative ways of constructing multivariate skew-normal distributions, see Arellano-Valle and Azzalini (2006).

2.3 *Skew-t distributions*

The kurtosis coefficient of a skew-normal distribution is restricted to the interval [3, 3.8692]. To achieve a higher degree of excess kurtosis, skew-*t* distributions have been introduced by Azzalini and Capitanio (2003). A univariate random variable Y follows the scalar skew-*t* distribution, $Y \sim \mathcal{ST}(\xi, \omega^2, \alpha, \nu)$, if it has the following stochastic representation:

$$Y = \xi + \omega\frac{X}{\sqrt{W}}, \tag{2.9}$$

where $X \sim \mathcal{SN}(\alpha)$ and $W \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, independently. The Gamma distribution $\mathcal{G}(a, b)$ is defined with density $p(y|a, b) = b^a y^{a-1} \mathrm{e}^{-by}/\,\Gamma(a)$. The pdf of Y reads:

$$f_{\mathcal{ST}}(y; \xi, \omega^2, \alpha, \nu) = \frac{2}{\omega} t_\nu(x_y) T_{\nu+1}\left(\alpha x_y \sqrt{\frac{\nu+1}{\nu+x_y^2}}\right), \tag{2.10}$$

where $x_y = (y-\xi)/\omega$ and t_ν and T_ν denote, respectively, the pdf and the cdf of a standard Student-*t* distribution with ν degrees of freedom. A random variable $\mathbf{Y}$ taking values in $\Re^r$ follows the multivariate skew-*t* distribution, $\mathbf{Y} \sim \mathcal{ST}_r(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu)$, if it has the following stochastic representation:

$$\mathbf{Y} = \boldsymbol{\xi} + \frac{1}{\sqrt{W}}\mathbf{X}, \tag{2.11}$$

where $\mathbf{X} \sim \mathcal{SN}_r(\mathbf{0}, \boldsymbol{\Omega}, \boldsymbol{\alpha})$ and $W \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$, independently. The pdf of $\mathbf{Y}$ reads:

$$f_{\mathcal{ST}}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) = 2\, f_{t_r}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) T_{\nu+r}\left(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\mathbf{y}-\boldsymbol{\xi})\sqrt{\frac{\nu+r}{\nu+Q_y}}\right), \tag{2.12}$$

where $\boldsymbol{\omega} = \text{Diag}(\boldsymbol{\Omega})^{1/2}$, $Q_y = (\mathbf{y}-\boldsymbol{\xi})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\xi})$, $f_{t_r}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu)$ denotes the pdf of the multivariate Student-*t* distribution $t_r(\boldsymbol{\xi}, \boldsymbol{\Omega}, \nu)$, and T_ν denotes the cdf of the scalar standard Student-*t* distribution as above. The skew-*t* distribution converges to the skew-normal distribution as $\nu \to \infty$. For any $r \geqslant 1$, the expectation of the skew-*t* distribution, provided that $\nu > 1$, is given by

$$E(\mathbf{Y}) = \boldsymbol{\xi} + \boldsymbol{\omega}\boldsymbol{\mu}_X, \quad \boldsymbol{\mu}_X = \boldsymbol{\delta}\sqrt{\frac{\nu}{\pi}}\frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}. \tag{2.13}$$

3. Skew-normal and skew-*t* finite-mixture models

We consider univariate and multivariate finite-mixture models where the component densities $p(\mathbf{y}_i|\boldsymbol{\theta}_k)$, $k = 1, \ldots, K$, arise either from a skew-normal or a skew-*t* distribution with component-specific parameter $\boldsymbol{\theta}_k$. The marginal distribution takes the form of a finite-mixture distribution with weights $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$ where $\sum_{k=1}^{K} \eta_k = 1$, for example, for a mixture of scalar skew-normal distributions:

$$p(y_i|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \boldsymbol{\eta}) = \eta_1 f_{\mathcal{SN}}(y_i; \xi_1, \omega_1^2, \alpha_1) + \cdots + \eta_K f_{\mathcal{SN}}(y_i; \xi_K, \omega_K^2, \alpha_K) \tag{3.1}$$

or for a mixture of multivariate skew-t distributions:

$$p(\mathbf{y}_i|\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \boldsymbol{\eta}) = \eta_1 f_{\mathcal{ST}}(\mathbf{y}_i; \boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\alpha}_1, \nu_1) + \cdots + \eta_K f_{\mathcal{ST}}(\mathbf{y}_i; \boldsymbol{\xi}_K, \boldsymbol{\Omega}_K, \boldsymbol{\alpha}_K, \nu_K).$$

Although this extension appears quite natural, the estimation of such a finite-mixture model results in a complex computational problem. In our subsequent Bayesian analysis, we combine the stochastic representations of the skew-normal and the skew-t distribution discussed in Section 2 with the standard hierarchical representation of a finite-mixture model in terms of a sequence of latent allocations. This leads to a straightforward MCMC sampling scheme.

3.1 *Finite mixture of random-effects model representation*

Like any other finite-mixture model, mixtures of skew-normal or skew-t distributions may be regarded as hierarchical latent variable models, where the distribution of the observations $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$ is specified conditional on latent allocations $\mathbf{S} = (S_1, \dots, S_N)$:

$$p(\mathbf{y}|\mathbf{S}, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) = \prod_{i=1}^{N} p(\mathbf{y}_i|S_i, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K) = \prod_{i=1}^{N} p(\mathbf{y}_i|\boldsymbol{\theta}_{S_i}),$$

where $\Pr(S_i = k|\boldsymbol{\eta}) = \eta_k$, $k = 1, \dots, K$ and $S_1, \dots, S_N$ are mutually independent. Conditional on S_i, the distribution underlying $p(\mathbf{y}_i|\boldsymbol{\theta}_{S_i})$ is represented as in Section 2 as a random-effects model. Thus, we obtain a representation of skew-normal or skew-t mixtures in terms of finite mixtures of random-effects models with truncated normal random effects.

For scalar skew-normal mixtures as defined in (3.1), the application of (2.3) to each component density leads to the following representation for $i = 1, \dots, N$,

$$\begin{aligned} z_i &\sim \mathcal{TN}_{[0,\infty)}(0, 1), \\ y_i|(S_i = k) &= \xi_k + \omega_k\delta_k z_i + \omega_k\sqrt{1-\delta_k^2}\varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1), \end{aligned}$$

where $z_1, \dots, z_N$ and $\varepsilon_1, \dots, \varepsilon_N$ are mutually independent. To implement our Bayesian approach, we introduce a new parameterization in terms of the component-specific parameters $\boldsymbol{\theta}_k^\star = (\xi_k, \psi_k, \sigma_k^2)$, where $\psi_k = \omega_k\delta_k$ and $\sigma_k^2 = \omega_k^2(1-\delta_k^2)$:

$$\begin{aligned} z_i &\sim \mathcal{TN}_{[0,\infty)}(0, 1), \\ y_i|(S_i = k) &= \xi_k + \psi_k z_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_k^2). \end{aligned} \tag{3.2}$$

The original parameter $\boldsymbol{\theta}_k = (\xi_k, \omega_k^2, \alpha_k)$ is recovered through

$$\alpha_k = \frac{\psi_k}{\sigma_k}, \quad \omega_k^2 = \sigma_k^2 + \psi_k^2 \tag{3.3}$$

because $\psi_k/\sigma_k = \omega_k\delta_k/\bigl(\omega_k\sqrt{1-\delta_k^2}\bigr) = \alpha_k$ and $\sigma_k^2 + \psi_k^2 = \omega_k^2(1-\delta_k^2) + \omega_k^2\delta_k^2 = \omega_k^2$.

Representation (3.2) offers several advantages. First, a conditionally conjugate prior for $\boldsymbol{\theta}_k^\star = (\xi_k, \psi_k, \sigma_k^2)$ is available and, second, straightforward estimation using a 2-block Gibbs sampler becomes feasible, see Section 3.2. A related representation with random-effects distribution $z_i \sim \mathcal{TN}_{[0,\infty)}(0, \omega_k^2)$ has been used in Lin, Lee, and Yen (2007), however, ξ_k, ω_k^2, and δ_k are sampled in different blocks and a Metropolis–Hastings algorithm is needed to sample $\delta_1, \dots, \delta_K$, while representation (3.2) allows to sample all component-specific parameters jointly from a closed-form posterior.

A similar representation is available for mixtures of multivariate skew-normal distributions $\mathcal{SN}_r(\boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k), k = 1, \dots, K$, where the componentwise application of (2.4) leads to a mixture of random-effects models with repeated measurements and a univariate truncated normal random effect:

$$
\begin{aligned}
&z_i \sim \mathcal{TN}_{[0,\infty)}(0, 1),\\
&\mathbf{y}_i|(S_i = k) = \boldsymbol{\xi}_k + \boldsymbol{\psi}_k z_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Sigma}_k),
\end{aligned} \tag{3.4}
$$

with $z_1, \dots, z_N$ and $\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N$ being mutually independent. We introduced the parameterization $\boldsymbol{\theta}_k^\star = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$, as we did for scalar skew-normal mixtures, where $\boldsymbol{\psi}_k = (\psi_{k1}, \dots, \psi_{kr})'$ with $\psi_{kj} = \omega_{kj}\delta_{kj}$ and $\boldsymbol{\Sigma}_k = \boldsymbol{\Omega}_k - \boldsymbol{\psi}_k\boldsymbol{\psi}_k'$. The form of $\boldsymbol{\Sigma}_k$ results from (2.7): $\boldsymbol{\Sigma}_k = \boldsymbol{\omega}_k\boldsymbol{\Delta}_k(\boldsymbol{\Omega}_{\boldsymbol{\varepsilon}})_k\boldsymbol{\Delta}_k\boldsymbol{\omega}_k = \boldsymbol{\Omega}_k - \boldsymbol{\omega}_k\boldsymbol{\Delta}_k\tilde{\boldsymbol{\alpha}}_k(\boldsymbol{\omega}_k\boldsymbol{\Delta}_k\tilde{\boldsymbol{\alpha}}_k)'$. The matrix $\boldsymbol{\Delta}_k\tilde{\boldsymbol{\alpha}}_k$ is a diagonal matrix with $(\boldsymbol{\Delta}_k\tilde{\boldsymbol{\alpha}}_k)_{jj} = \sqrt{1-\delta_{kj}^2}\delta_{kj}/\sqrt{1-\delta_{kj}^2} = \delta_{kj}$, therefore $\boldsymbol{\omega}_k\boldsymbol{\Delta}_k\tilde{\boldsymbol{\alpha}}_k = \boldsymbol{\psi}_k$. The original parameter $\boldsymbol{\theta}_k = (\boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k)$ is recovered from

$$
\boldsymbol{\Omega}_k = \boldsymbol{\Sigma}_k + \boldsymbol{\psi}_k\boldsymbol{\psi}_k', \quad \boldsymbol{\alpha}_k = \frac{1}{\sqrt{1 - \boldsymbol{\psi}_k'\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k}}\boldsymbol{\omega}_k\boldsymbol{\Omega}_k^{-1}\boldsymbol{\psi}_k. \tag{3.5}
$$

For skew-t mixtures, we combine the stochastic representations (2.9) or (2.11) with the random-effects representation of the skew-normal distribution. For a finite mixture of scalar skew-t distributions $\mathcal{ST}(\xi_k, \omega_k^2, \alpha_k, \nu_k)$, $k = 1, \dots, K$, this yields

$$
w_i|(S_i = k) \sim \mathcal{G}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right), \tag{3.6}
$$

$$
z_i|w_i \sim \mathcal{TN}_{[0,\infty)}\left(0, \frac{1}{w_i}\right), \tag{3.7}
$$

$$
y_i|(S_i = k, w_i) = \xi_k + \psi_k z_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_k^2/w_i), \tag{3.8}
$$

where $w_1, \dots, w_N$ are mutually independent as are $z_1, \dots, z_N$ and $\epsilon_1, \dots, \epsilon_N$ given $w_1, \dots, w_N$. A finite mixture of multivariate skew-t distributions $\mathcal{ST}_r(\boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k, \nu_k)$, $k = 1, \dots, K$, has a similar representation with a repeated measurements observation equation:

$$
\mathbf{y}_i|(S_i = k, w_i) = \boldsymbol{\xi}_k + \boldsymbol{\psi}_k z_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_r\left(\mathbf{0}, \frac{1}{w_i}\boldsymbol{\Sigma}_k\right), \tag{3.9}
$$

where $z_1, \dots, z_N$ and $\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N$ are mutually independent given $w_1, \dots, w_N$. The variance of the truncated normal random effect z_i depends on the latent scaling factor w_i which results from multiplying in (3.2) or (3.4) a $\mathcal{TN}_{[0,\infty)}(0, 1)$ random variable with $1/\sqrt{w_i}$.

As for skew-normal mixtures, we use an alternative parameterization with component-specific parameter $\boldsymbol{\theta}_k^\star = (\xi_k, \psi_k, \sigma_k^2, \nu_k)$ and $\boldsymbol{\theta}_k^\star = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \nu_k)$, respectively. This allows Bayesian estimation through a 3-block MCMC sampler where only sampling of the degrees of freedom parameters $\nu_1, \dots, \nu_K$ requires a Metropolis–Hastings step.

3.2 *Bayesian estimation*

To perform a Bayesian analysis, we first have to select a prior for the weight distribution $\boldsymbol{\eta}$ and the component-specific parameters. It should be noted that, in general, the prior distribution has to be selected carefully in the context of finite-mixture models. First of all, it is not possible to choose an improper prior because this leads to an improper posterior density (see, e.g. Frühwirth-Schnatter, 2006, Section 3.2). Furthermore, as noted by Jennison (1997), one should avoid trying to be as "noninformative as possible" by choosing large prior variances because the choice of the prior of the parameters strongly affects the posterior of the number of components K which will be considered in Section 3.3 for selecting K. For this reason, we extend the hierarchical priors introduced by Richardson and Green (1997, Section 2.4) in the context of mixtures of normals and by Stephens (1997) in the context of mixtures of t-distributions to skew-normal and skew-*t* mixtures. Such hierarchical priors are known to reduce sensitivity with respect to choosing the prior variances.

Concerning the weight distribution, we apply the commonly used Dirichlet distribution $\boldsymbol{\eta} \sim \mathcal{D}(e_0, \dots, e_0)$. Nobile (2004) showed that the parameter e_0 exercises considerable influence on the posterior distribution of K because this parameter strongly affects the link between the marginal likelihoods of finite-mixture models with $K-1$ and K components. Frühwirth-Schnatter (2006, Section 5.3.2) demonstrated that this link is reduced considerably by selecting e_0 larger than $e_0 = 1$, which is the value commonly used in the literature.

Concerning the component-specific parameters, we specify priors for the transformed parameters $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$ introduced in Section 3.1 rather than directly for $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$. Using the representations discussed in Section 3.1, conditionally conjugate priors taking the form of normal–gamma distributions are available for all transformed component-specific parameters except the degrees of freedom parameters $\nu_1, \dots, \nu_K$. The prior on ν_k is a slight modification of a prior introduced by Juárez and Steel (2010) for Student-*t* mixtures with $\nu_1 = \cdots = \nu_K$. Further details for all priors are provided in Section A of the supplementary material available at *Biostatistics* online.

Following the seminal paper by Diebolt and Robert (1994), the most popular method for Bayesian estimation of finite mixtures is to apply MCMC methods (see Frühwirth-Schnatter, 2006, Section 3.5, for an extensive review). This approach is extended to skew-normal and skew-*t* mixtures using the representations introduced in Section 3.1. We introduce the latent allocations $\mathbf{S} = (S_1, \dots, S_N)$ and the latent random effects $\boldsymbol{z} = (z_1, \dots, z_N)$ as missing data and add the latent scaling factors $\boldsymbol{w} = (w_1, \dots, w_N)$ for skew-*t* mixtures. MCMC sampling is based on the following observations.

First, as for more conventional finite-mixture models, it is possible to sample the allocations $\mathbf{S}$ given the component-specific parameters $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$ and the weights $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$ without conditioning on the other latent variables $\boldsymbol{z}$ (and $\boldsymbol{w}$) because the component densities are available in closed form, see Section 2.

Second, conditional on $\mathbf{S}$ (and $\boldsymbol{w}$), we consider skew-normal and skew-*t* mixtures as random-effects models with a normal observation equation and a truncated normal random effect. A nice property of such a model is that the full conditional of the random effect z_i given the observation $\mathbf{y}_i$ is available in closed form, see Section B.1 of the supplementary material available at *Biostatistics* online. This allows joint multi-move sampling of the latent variables $\mathbf{S}$ and $\boldsymbol{z}$.

Third, conditional on $\mathbf{S}, \boldsymbol{z}$, (and $\boldsymbol{w}$) sampling of the transformed component specific parameters $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$ (except the degrees of freedom) reduces to Bayesian inference for a finite mixture of regression models with known allocations. For each group k, $(\boldsymbol{\xi}_k' \ \boldsymbol{\psi}_k')'$ is a regression coefficient and $\boldsymbol{\Sigma}_k$ is an error covariance matrix in a regression model with a conditionally conjugate prior, allowing joint sampling of $\boldsymbol{\xi}_k$, $\boldsymbol{\psi}_k$, and $\boldsymbol{\Sigma}_k$ from a closed-form posterior distribution.

As a result, MCMC estimation for skew-normal mixtures is possible through a 2-step Gibbs sampler if all hyperparameters are fixed:

(a) Sample $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$ and $\boldsymbol{\eta}$ conditional on $\boldsymbol{z}$, $\mathbf{S}$, and $\mathbf{y}$.
(b) Sample $\boldsymbol{z}$ and $\mathbf{S}$ jointly conditional on $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$, $\boldsymbol{\eta}$, and $\mathbf{y}$.

All conditional densities are of closed form, see Section B.2 of the supplementary material available at *Biostatistics* online, which also contains details on sampling under a hierarchical prior. For MCMC estimation of skew-*t* mixtures, a third step has to be added:

(a) Sample $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$ (except $\nu_1, \dots, \nu_K$) and $\boldsymbol{\eta}$ conditional on $\boldsymbol{z}$, $\mathbf{S}$, $\boldsymbol{w}$, and $\mathbf{y}$.
(b) Sample $\boldsymbol{z}$ and $\mathbf{S}$ jointly conditional on $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$, $\boldsymbol{\eta}$, $\boldsymbol{w}$, and $\mathbf{y}$.
(c) Sample $\nu_1, \dots, \nu_K$ and $\boldsymbol{w}$ conditional on $\mathbf{y}$ and the remaining parameters.

All conditional densities except $p(\nu_1, \dots, \nu_K|\cdot)$ are of closed form, see Section B.3 of the supplementary material available at *Biostatistics* online for details.

Like for any finite-mixture model, a nonidentifiability problem is present because the labeling of the components in the mixture density may be changed without changing the likelihood $p(\mathbf{y}|\boldsymbol{\vartheta})$, see, for example, (3.1). This might cause label switching during MCMC sampling which makes it difficult to estimate component-specific parameters from the MCMC output. Various methods have been suggested in the literature do deal with this problem, (see, e.g. Celeux *and others*, 2000; Stephens, 2000; Jasra *and others*, 2005). Here, we follow Frühwirth-Schnatter (2001) who suggested to add a random permutation step to the MCMC scheme and to postprocess the resulting MCMC output to identify component-specific parameters, see Section B.4 of the supplementary material available at *Biostatistics* online for more details.

3.3 *Selecting the number of components*

Selecting the number of components of a finite-mixture model is a challenge (see Frühwirth-Schnatter, 2006, Chapter 5, for a recent review). Popular methods implement reversible jump MCMC, compute marginal likelihoods, or use model choice criteria.

Reversible jump MCMC was introduced by Richardson and Green (1997) to select the number of components for univariate mixtures of normal distributions. This method is based on creating a Markov chain that moves between finite mixtures with different number of components while retaining detailed balance that ensures the correct limiting distribution. Those moves have to be based on carefully selected degenerate proposal densities. The design of suitable proposals for higher-dimensional mixtures is quite a challenge, see, for example, Dellaportas and Papageorgiou (2006) for multivariate normal mixtures. Since adding skewness even complicates matters, we did not pursue reversible jump MCMC.

Alternatively, the choice of K may be based on the posterior probability $p(\mathcal{M}_K|\mathbf{y})$ of a finite-mixture model $\mathcal{M}_K$ with K components, given by $p(\mathcal{M}_K|\mathbf{y}) \propto p(\mathbf{y}|\mathcal{M}_K)p(\mathcal{M}_K)$, where $p(\mathbf{y}|\mathcal{M}_K)$ is the marginal likelihood and $p(\mathcal{M}_K)$ is the prior probability of $\mathcal{M}_K$, for instance, a truncated Poisson distribution (see, e.g. Nobile, 2004).

Also the computation of the marginal likelihood $p(\mathbf{y}|\mathcal{M}_K)$ turns out to be challenging for skew-normal and skew-*t* mixtures. For moderate K, say $K \leqslant 5$, we follow Frühwirth-Schnatter (2004) who demonstrates that the technique of bridge sampling (Meng and Wong, 1996) is a useful method of computing the marginal likelihood of a finite-mixture model and is superior to alternative sampling-based approaches such as importance sampling (Neal, 2001). Like importance sampling, bridge sampling is based on an i.i.d. sample from an importance density, however, this sample is combined with the MCMC draws from the posterior density in an appropriate way. An important advantage of bridge sampling over importance sampling is that the variance of the resulting estimator depends on a ratio that is bounded regardless of the tail behavior of the underlying importance density.

For larger values of K, all simulation-based estimators including bridge sampling turned out to be unstable. For such mixtures, model choice criteria may be considered. One such criterion is the Bayesian information criterion (BIC_K) that is an asymptotic approximation to $-2\log p(\mathbf{y}|\mathcal{M}_K)$:

$$\mathrm{BIC}_K = -2\log p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}_K, \mathcal{M}_K) + d_K \log N, \tag{3.10}$$

where $d_K = (2r+1)K - 1 + Kr(r+1)/2 = d_K^N$ for skew-normal mixtures and $d_K = d_K^N + K$ for skew-t mixtures. $\hat{\boldsymbol{\vartheta}}_K$ is an approximate maximum likelihood (ML) estimator of $\boldsymbol{\vartheta}_K = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \boldsymbol{\eta})$ obtained by maximizing the log of the observed-data likelihood function $\log p(\mathbf{y}|\boldsymbol{\vartheta}_K, \mathcal{M}_K)$ over the MCMC draws. If the distribution family underlying the component densities is correctly specified, then BIC_K is known to be consistent (Keribin, 2000), although in small data sets, it tends to choose models with too few components (Biernacki *and others*, 2000). On the other hand, simulation studies reported in Biernacki and Govaert (1997), Biernacki *and others* (2000), and McLachlan and Peel (2000, Section 6.11) show that BIC_K will overrate the number of clusters under misspecification of the component density, whereas several alternative criteria such as the approximate weight of evidence (AWE_K) and the integrated classification likelihood (ICL_K) criterion to be discussed below are able to identify the correct number of clusters even when the component densities are misspecified. Thus, BIC_K for clustering large data sets, where the component densities of the finite-mixture model may not be correctly specified, is likely to be imprecise.

AWE_K is derived in Banfield and Raftery (1993) as another approximation to minus twice the log of the marginal likelihood. AWE_K is described in Biernacki and Govaert (1997) as a criterion that penalizes the log of the complete-data likelihood function with model complexity:

$$\mathrm{AWE}_K = -2\log p(\mathbf{y}, \hat{\mathbf{S}}|\hat{\boldsymbol{\vartheta}}_K^C) + 2d_K\left(\frac{3}{2} + \log N\right), \tag{3.11}$$

where $\hat{\boldsymbol{\vartheta}}_K^C$ and $\hat{\mathbf{S}}$ are determined jointly as that combination of parameters and allocations that maximize the log of the complete-data likelihood $\log\ p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta}_K) = \sum_{i=1}^N \log(\eta_{S_i} p(\mathbf{y}_i|\boldsymbol{\theta}_{S_i}))$. $\hat{\boldsymbol{\vartheta}}_K^C$ is approximated by the posterior draw maximizing the complete-data likelihood function.

Biernacki *and others* (2000) introduced the ICL_K that has been shown by McLachlan and Peel (2000, p. 216) to be approximately equal to

$$\mathrm{ICL{-}BIC}_K = \mathrm{BIC}_K + 2\mathrm{EN}(\hat{\boldsymbol{\vartheta}}_K). \tag{3.12}$$

$\mathrm{EN}(\boldsymbol{\vartheta}_K)$ is the entropy defined by

$$\mathrm{EN}(\boldsymbol{\vartheta}_K) = -\sum_{i=1}^N \sum_{k=1}^K \Pr(S_i = k|\mathbf{y}_i, \boldsymbol{\vartheta}_K) \log \Pr(S_i = k|\mathbf{y}_i, \boldsymbol{\vartheta}_K)$$

and measures how well the finite-mixture model defined by $\boldsymbol{\vartheta}_K$ classifies the data into K distinct clusters. Thus, the $\mathrm{ICL{-}BIC}_K$ criterion penalizes not only model complexity but also the failure of the model to provide a classification into well-separated clusters.

Recently, the deviance information criterion (DIC) introduced by Spiegelhalter *and others* (2002) became a popular criterion for Bayesian model selection because it is easily computed from the MCMC draws. However, the application of DIC to finite-mixture models is not without problems as discussed recently by Celeux *and others* (2006). A first problem is the choice of the appropriate likelihood function that could either be the observed-data likelihood function $\log\ p(\mathbf{y}|\boldsymbol{\vartheta}_K)$, the complete-data likelihood function $\log\ p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta}_K)$, or the conditional likelihood $p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta}_K)$. Second, the calculation of DIC requires an estimator of the unknown parameter $\boldsymbol{\vartheta}_K$ which may suffer from label switching as discussed above, making DIC unstable. Finally, DIC involving the complete-data or the conditional likelihood requires some way of handling the problem that $\mathbf{S}$ is unobserved, either by integrating with respect to the posterior $p(\mathbf{S}|\mathbf{y}, \mathcal{M}_K)$ or by using an estimator of $\mathbf{S}$ where once more the label switching problem has to be addressed.

In reaction to these difficulties, Celeux *and others* (2006) investigate in total 8 different DIC criteria. DIC_2, for instance, focuses on the marginal distribution of the data and considers the allocations as nuisance parameters. Consequently, it is based on the observed-data likelihood:

$$\mathrm{DIC}_{2,K} = -4E_{\boldsymbol{\vartheta}_K}(\log p(\mathbf{y}|\boldsymbol{\vartheta}_K)|\mathbf{y}) + 2\log p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}^{M}_{K}, \mathbf{y}), \tag{3.13}$$

where the posterior mode estimator $\hat{\boldsymbol{\vartheta}}^{M}_{K}$ which is invariant to label switching is obtained from the observed-data posterior $p(\boldsymbol{\vartheta}_K|\mathbf{y}, \mathcal{M}_K)$.

Based on several simulation studies, Celeux *and others* (2006) recommend using DIC_4 which is based on computing first DIC for the complete-data likelihood function and then integrating with respect to the posterior $p(\mathbf{S}|\mathbf{y}, \mathcal{M}_K)$:

$$\mathrm{DIC}_{4,K} = -4E_{\boldsymbol{\vartheta}_K,\mathbf{S}}(\log p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta}_K)|\mathbf{y}) + 2E_{\mathbf{S}}(\log p(\mathbf{y}, \mathbf{S}|\hat{\boldsymbol{\vartheta}}_K(\mathbf{S}))|\mathbf{y}). \tag{3.14}$$

The application of this criterion requires the computation of the complete-data estimator $\hat{\boldsymbol{\vartheta}}_K(\mathbf{S})$ for each draw from the posterior $p(\mathbf{S}|\mathbf{y}, \mathcal{M}_K)$ which is straightforward only for simple mixture models, where the complete-data posterior $p(\boldsymbol{\theta}_k|\mathbf{y}, \mathbf{S})$ is of closed form. However, this is not the case for the class of skew finite mixtures. Celeux *and others* (2006) show that substituting $\hat{\boldsymbol{\vartheta}}_K(\mathbf{S})$ by the posterior mode estimator $\hat{\boldsymbol{\vartheta}}^{M}_{K}$, an approximation to $\mathrm{DIC}_{4,K}$ is obtained which penalizes $\mathrm{DIC}_{2,K}$ by the expected entropy:

$$\mathrm{DIC}_{4a,K} = \mathrm{DIC}_{2,K} + 2E_{\boldsymbol{\vartheta}_K}(\mathrm{EN}(\boldsymbol{\vartheta}_K)|\mathbf{y}). \tag{3.15}$$

For skew finite mixtures, both $\mathrm{DIC}_{2,K}$ as well as $\mathrm{DIC}_{4a,K}$ are easily estimated from the MCMC draws from the posterior $p(\boldsymbol{\vartheta}_K|\mathbf{y}, \mathcal{M}_K)$ by substituting all expectations $E_{\boldsymbol{\vartheta}_K}(\cdot|\mathbf{y})$ by the average over the MCMC draws. Note that the label switching is not a problem here because both $\log p(\mathbf{y}|\boldsymbol{\vartheta}_K)$ as well as $\mathrm{EN}(\boldsymbol{\vartheta}_K)$ are invariant to changing the labeling of the groups.

4. Applications

When the observations in a study generate asymmetric data, even moderately imprecise models could lead to erroneous classification of the subjects. As shown in the following examples, we address this problem with the help of precise skew mixture modeling.

4.1 *Skew-normal mixture modeling of Alzheimer's disease data*

AD is a complex disease that has multiple genetic as well as environmental risk factors. It is commonly characterized by loss of a wide range of cognitive abilities with aging. For the present analysis, the data set consists of 451 subjects from the cohorts of the Religious Orders Study, see Wilson *and others* (2004) and the Memory and Aging Project, see Bennett *and others* (2005). The level of cognition of the subjects was clinically evaluated proximate to their death based on tests of cognitive functions and summarized by a mean global cognition score, with higher scores suggesting better cognition capabilities. The genetic risk factor Apolipoprotein E (ApoE) polymorphism was determined by genotyping the DNA from the subjects' blood.

Since the distribution of global cognition scores appeared to be skewed, see again Figure 1, we applied skew-normal and skew-t mixture models with $K = 1, \dots, 4$ components. Bayesian analysis is based on the priors introduced in Section A of the supplementary material available at *Biostatistics* online with different sets of hyper parameters. For all priors, $b_0^{\psi} = b_0^{\xi} = 0$ and $g_0 = 0.5$, while we consider 4

Table 1. *Choosing the hyper parameters D^{ξ}, D^{ψ}, c_0, and ϕ of the prior in skew-normal and skew-t mixture modeling of AD data set; $R^2 = 1-\phi/(c_0-1)$ is the prior expectation of explained heterogeneity; d is an additional hyperparameter for skew-t mixtures*

	$D^{\xi} = D^{\psi}$	c_0	ϕ	R^2	d	Median of ν_k
Prior 1	0.1	2.5	0.5	2/3	$9/(1+\sqrt{2})$	10
Prior 2	0.01	2.5	0.5	2/3	$9/(1+\sqrt{2})$	10
Prior 3	0.1	5	4/3	2/3	$9/(1+\sqrt{2})$	10
Prior 4	0.1	2.5	1	1/3	$9/(1+\sqrt{2})$	10
Prior 5	0.1	2.5	0.5	2/3	$4/(1+\sqrt{2})$	5

different settings for $D^{\xi} = D^{\psi}$, c_0, and ϕ and 2 different values for d for skew-t mixtures, see Table 1 for details. Compared to the other priors, prior 2 introduces considerably smaller prior information for the location parameter ξ and the skewness parameter ψ, prior 3 introduces stronger smoothing for the group specific variances $\sigma_1^2, \ldots, \sigma_K^2$, and prior 4 assumes a smaller prior expectation of the heterogeneity explained by differences in the group locations. Prior 5 applies only to skew-t mixtures and reduces the prior median of ν_k by 50% compared to the other priors.

For each K and each prior, we generate 50 000 MCMC draws after a burn-in of 10 000 draws by using the MCMC schemes described in Sections B.2 and B.3 of the supplementary material available at *Biostatistics* online.

To select the optimal K, marginal likelihoods $p(\mathbf{y}|\mathcal{M}_K)$ are computed for each prior as described in Section 3.3 and are combined with a truncated $\mathcal{P}(2)$-prior for K. The resulting (nonnormalized) posterior probabilities $\log(p(\mathbf{y}|\mathcal{M}_K)p(\mathcal{M}_K))$ are reported in Table 2. The same table reports BIC_K and $\mathrm{DIC}_{2,K}$ for the various priors. Although BIC_K is independent from the prior, differences in the estimated values of BIC_K occur caused by random fluctuations of the approximate ML estimator across MCMC runs. Table 2 reports the smallest BIC_K among all MCMC runs. Table 2 does not report the remaining criteria introduced in Section 3.3 because, regardless of the prior, AWE_K, ICL$-\mathrm{BIC}_K$, as well as $\mathrm{DIC}_{4a,K}$ selected a model with $K = 1$ which, however, contradicts common knowledge of AD classification.

For skew-normal mixtures, both the marginal likelihood as well as BIC_K select a model with 2 components for all priors considered. In contrast to that, $\mathrm{DIC}_{2,K}$ shows high sensitivity to prior choices and the selected number of components ranges from 2 to 4. For skew-t mixtures, we find that BIC_K favors a 2-component mixture, however, this model is outperformed by the 2-component skew-normal mixture. In contrast to skew-normal mixtures, model selection based on marginal likelihoods is sensitive to prior choices. Under prior 1 and prior 5, the marginal likelihood rejects skew-t mixtures in favor of a single skew-t distribution. For prior 2 and prior 4, $K = 2$ is selected, while prior 3 leads to choosing $K = 3$. Upon comparison of all models, we find a preference for a skew-normal mixture with 2 components for all priors. For skew-t mixtures, sensitivity of $\mathrm{DIC}_{2,K}$ to prior choices is even higher and the selected number of components ranges from 1 to 4.

For comparison, we also fitted finite mixtures of normal distributions where the priors are selected similarly as in Richardson and Green (1997). The marginal likelihoods are computed as in Frühwirth-Schnatter (2004) and $p(\mathcal{M}_K)$ is the same as above. The (nonnormalized) posterior probabilities $\log(p(\mathbf{y}|\mathcal{M}_K)p(\mathcal{M}_K))$ together with BIC_K and $\mathrm{DIC}_{2,K}$ are reported in Table 2. The first 2 criteria select a normal mixture with 3 components, while $\mathrm{DIC}_{2,K}$ leads to choosing $K = 4$.

When the 3-component normal mixture is compared with the 2-component skew-normal mixture, the latter one is preferred by the marginal likelihood and BIC_K, regardless of the prior. Figure 1 shows that the fitted density is practically the same for both finite mixtures. While one of the clusters is comparable for both mixtures (see the leftmost cluster in the left-hand and the right-hand side of Figure 1), the normal mixture needs 2 components to fit the skewness in the second cluster.

Table 2. *Selecting the number K of components in Gaussian and skew-normal and skew-t mixture modeling of AD data set*

		K			
Skew-normal mixtures		1	2	3	4
BIC_K		1376.25	**1363.13**	1385.29	1404.09
$\log(p(\mathbf{y}\|\mathcal{M}_K)p(\mathcal{M}_K))$	Prior 1	–690.11	**–684.28**	–686.97	–692.80
	Prior 2	–690.88	**–682.68**	–688.74	–697.69
	Prior 3	–690.80	**–683.79**	–685.94	–691.60
	Prior 4	–691.16	**–683.89**	–686.48	–692.87
$\text{DIC}_{2,K}$	Prior 1	1363.98	1345.40	1345.89	**1344.68**
	Prior 2	1363.81	**1335.14**	1337.65	1335.24
	Prior 3	1363.94	1340.54	**1335.23**	1351.47
	Prior 4	1364.10	**1343.33**	1346.66	1351.14
Skew-*t* mixtures		1	2	3	4
BIC_K		1382.32	**1375.76**	1406.15	1436.53
$\log(p(\mathbf{y}\|\mathcal{M}_K)p(\mathcal{M}_K))$	Prior 1	**–690.65**	–695.38	–692.89	–698.74
	Prior 2	–693.48	**–687.82**	–692.00	–700.87
	Prior 3	–693.20	–698.49	**–692.43**	–696.77
	Prior 4	–693.58	**–691.77**	–693.25	–698.12
	Prior 5	**–693.55**	–696.95	–696.93	–700.47
$\text{DIC}_{2,K}$	Prior 1	1363.48	1369.19	1359.07	**1350.87**
	Prior 2	1363.73	1341.41	**1340.78**	1343.30
	Prior 3	1364.23	1375.39	**1344.26**	1347.87
	Prior 4	1364.53	1372.33	1355.57	**1354.62**
	Prior 5	**1364.87**	1388.21	1378.25	1367.83
Normal mixtures		1	2	3	4
BIC_K		1473.93	1371.69	**1369.09**	1378.97
$\log(p(\mathbf{y}\|\mathcal{M}_K)p(\mathcal{M}_K))$		–740.70	–686.87	**−685.83**	–686.30
$\text{DIC}_{2,K}$		1465.70	1350.60	1354.37	**1345.76**

Bold values indicate the selected number K.

The 2-component skew-normal mixture is identified for each prior as described in Section B.4 of the supplementary material available at *Biostatistics* online. Figure 2 shows the resulting posterior draws of α_1 and α_2 for prior 1. The estimated parameters are reported for all priors in Table 3. Evidently, the skewness parameter α_k is sensitive to selecting the prior information D^{ξ} and D^{ψ} which is much smaller under prior 2 than for the other priors. On the other hand, the expected cognitive score μ_k and the group sizes η_k are insensitive to prior choices. For all priors, the first component has a much higher expected cognitive score μ_k than the second one and exhibit considerable negative skewness. The skewness parameter α_k is positive for the second component, however, strongly depends on the prior and exhibits very large standard errors.

Among the genetic risk factors for AD, the pivotal role of ApoE gene is well established (Wilson *and others*, 2002; Roses, 1997). There are 3 different allele polymorphisms of the gene in general population—e2, e3, and e4—and the number of copies of e4 is linked to increased risk of early onset of the disease. Hence, an individual with the homozygous alleles e44 (i.e. both alleles are e4) carries greater risk than one with heterozygous e34 (i.e. an e3 and an e4); the latter, in turn, has greater risk than e24 (which however has normal risk similar to e33) as those with e2 alleles have reduced risk of early onset of AD.

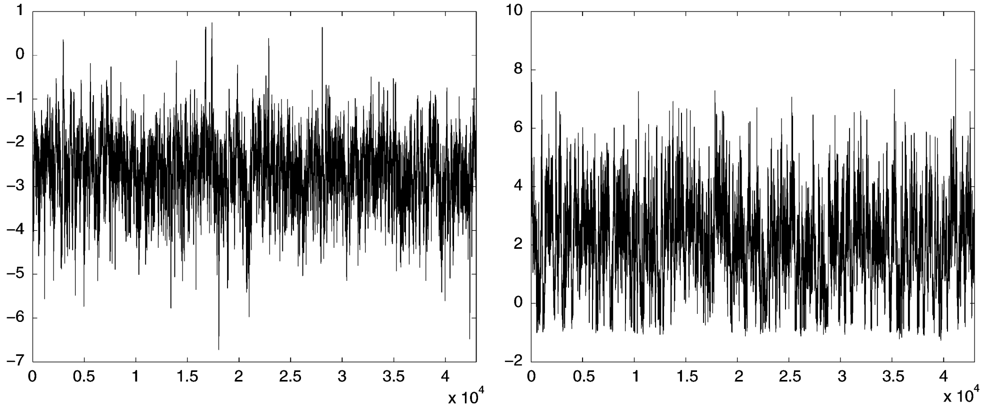

Fig. 2. Two-component skew-normal mixture modeling of AD data set. Posterior draws obtained under prior 1 for the skewness parameters α_1 (left-hand side) and α_2 (right-hand side) after identification.

Table 3. *Two-component skew-normal mixture modeling of AD data set. Parameter estimation under different priors using posterior means (posterior standard deviations in parenthesis)*

	k	ξ_k	ω_k^2	α_k	$\mu_k = E(Y\|S_i = k)$	η_k
Prior 1	1	0.36 (0.11)	1.26 (0.37)	–2.61 (0.78)	–0.46 (0.096)	0.767 (0.061)
	2	–3.55 (0.43)	2.20 (1.30)	2.06 (1.48)	–2.65 (0.34)	0.233 (0.061)
Prior 2	1	0.44 (0.07)	1.54 (0.32)	–3.63 (0.89)	–0.51 (0.076)	0.777 (0.048)
	2	–4.09 (0.10)	3.87 (1.39)	8.31 (2.96)	–2.57 (0.294)	0.223 (0.048)
Prior 3	1	0.38 (0.10)	1.38 (0.36)	–2.80 (0.73)	–0.49 (0.101)	0.782 (0.055)
	2	–3.88 (0.24)	2.59 (1.22)	3.47 (1.31)	–2.70 (0.331)	0.218 (0.055)
Prior 4	1	0.35 (0.12)	1.27 (0.36)	–2.58 (0.79)	–0.47 (0.086)	0.77 (0.054)
	2	–3.75 (0.30)	2.49 (1.23)	2.85 (1.44)	–2.65 (0.301)	0.23 (0.054)

First, we used the skew-normal mixture model to classify each subject into one of the 2 components. To test how this classification is related to the genetic risk factor, we assigned the genotype labels into 2 classes: lower risk {e22, e23, e33} and higher risk {e24, e34, e44}. Under prior 1, for instance, we found that 84.5% of the lower risk subjects, as opposed to only 28.4% of the higher risk subjects, were assigned to the component with the higher expected cognitive score. On the other hand, 71.6% of the higher risk subjects, but only 15.5% of the lower risk subjects, were assigned to the component with the lower expected cognitive score. This clearly indicates consistent classification of the cognition scores of the subjects based on their genetic risk factors. The genotype labels are also plotted in Figure 1 in grayscales, from the lightest to the darkest in the sequence {e22, e23, e33, e24, e34, e44}, as rugplot for visual perception of the classification.

Further, to test the classification induced by the normal mixture model with 3 components, we assigned the genotype labels into 3 classes: reduced risk {e22, e23}, normal risk {e33, e24}, and increased risk {e34, e44}. We found that 26.87%, 38.06%, and 35.07%, respectively, of the higher risk subjects were classified into the left, the middle, and the right normal components in Figure 1; the same numbers for the lower risk subjects were 15.48%, 32.26%, and 52.26%, respectively. In contrast to the precise classification by the 2-component skew-normal mixture model, the classification by the 3-component normal mixture model is weak, which may be attributed to the spurious splitting of 1 skewed component into 2 symmetric ones.

4.2 *Multivariate skew-t mixture modeling of flow cytometric data*

A flow cytometer is an instrument that measures the expression of proteins on the surface of and within individual cells in a given sample. Fluorescently tagged antibodies are used as markers to bind the corresponding proteins and thus measure the amounts expressed for each cell in terms of fluorescence intensities. This produces a high-throughput sample in which each cell is represented by a high-dimensional data point where a dimension corresponds a particular marker.

Typically, a flow cytometric data analyst looks at the high-dimensional flow readout in 2D projections and manually identifies (or “gates”) the cell populations of interest. A flow cytometric sample is generally understood as a mixture of different cell populations which express in the form of immuno-phenotypic clusters under different conditions such as disease and control. Therefore, finite-mixture modeling approach to cluster the cell populations in terms of their protein expression provides a natural interpretation to the mixture components. Moreover, it provides automation, rigor, and reproducibility in flow data analysis.

Often cell populations in flow cytometric readouts suffer from considerable presence of non-Gaussian characteristics, such as prominent skewness and large number of outliers. Therefore, while Gaussian mixture modeling is not unprecedented in flow data analysis (Boedigheimer and Ferbas, 2008; Chan *and others*, 2008), it neither models formally the skewness in flow populations nor is robust against large number of outliers. Both skewness and outliers cause inaccurate inference by Gaussian mixture modeling due to fitting more components than the true number of clusters present in the data. Given the multimodal, multidimensional, and asymmetric nature of flow cytometric cell populations, it appears to be a perfectly suitable and most useful application for multivariate skew-*t* finite-mixture modeling.

In the following example, we used skew-*t* mixture modeling to do a comparative analysis of a peripheral blood sample from a subject who developed GvHD following blood and marrow transplantation with a control sample from a subject who underwent similar transplant but did not develop signs of the disease. The samples were obtained from publicly available data due to the study of Brinkman *and others* (2007), which may be referred to for further details. Brinkman *and others* (2007) observed an increased proportion in the CD4+CD8β+CD3+ population to be correlated with the development of GvHD.

Recently, Lo *and others* (2008) used an expectation-maximization-based Student-*t* mixture model, with Box–Cox transformation to diminish the asymmetry of populations in the sample. Their optimal model for the Brinkman *and others* (2007), GvHD data had 12 components based on BIC, a count that exceeded our optimal model (see below). In this respect, it may be noted that finding a suitable transformation to adequately correct the skew in data is known to be difficult (Kruglyak and Lander, 1995), and thus the resulting modeling with symmetric *t* densities could lead to inaccurate inference.

In contrast, we fit 6-variate finite mixtures of skew-normal and skew-*t* distributions over a range of $K = 1, \dots, 14$ components to the case sample GvHDB01case containing a population of 12 442 cells and the control sample GvHDB06control containing a population of 8691 cells. For Bayesian estimation, we use the priors introduced in Section A of the supplementary material available at *Biostatistics* online with $\mathbf{b}_0^{\psi} = \mathbf{0}_{6\times 1}$, $\mathbf{b}_0^{\xi} = \bar{\mathbf{y}}$, $D^{\xi} = D^{\psi} = 0.1$, $c_0 = 5.5$, $g_0 = 0.5$, $\phi = 0.5$, and $d = 9/(1+\sqrt{2})$ for the skew-*t* mixtures.

Since, we expect the posterior density to have many local modes, we generated for each K several independent chains, each with 10 000 MCMC draws after a burn-in of 5000 draws using the MCMC schemes described in Sections B.2 and B.3 of the supplementary material available at *Biostatistics* online. In fact, it turned out that the various chains converged to different modal regions of the parameter space. For further inference we selected for each value of K, the chain with the smallest BIC_K, computed as in (3.10). This guarantees that we are dealing with posterior draws from a modal region with high posterior probability because $-0.5\mathrm{BIC}_K$ is a rough estimate of the marginal likelihood of a model where the parameters are restricted to each modal region.

Table 4. *Choosing K for the flow cytometric data*

	Data set GvHDB01case			Data set GvHDB06control			
Skew-*t*							
	$K = 8$	$K = 9$	$K = 10$	$K = 7$	$K = 8$	$K = 9$	$K = 10$
$\text{ICL}-\text{BIC}_K$	–15126.6	**–17810.1**	–17580.3	7651.2	**6823.7**	6942.7	9054.95
AWE_K	–12077.4	**–14378.5**	–13766.1	10230.3	**9772.7**	10261.7	12743.90
$\text{DIC}_{4a,K}$	–16757.8	–19568.9	–19898.1	5719.8	5068.2	**4368.5**	6315.3
Skew-normal							
	$K = 11$	$K = 12$	$K = 13$	$K = 12$	$K = 13$	$K = 14$	
ICL-BIC_K	–16050.1	**–16543.8**	–15510.2	11307.5	**10320.4**	11793.8	
AWE_K	–11973.6	**–12095.7**	–10690.6	15609.5	**14981.8**	16814.6	
$\text{DIC}_{4a,K}$	–16643.2	–17883.8	–18781.8	8369.7	**7562.4**	8259.87	

Bold values indicate the selected number K.

To select the optimal number of components, various criteria introduced in Section 3.3 were computed both for skew-normal and skew-*t* mixtures. Since we would like to find well-separated clusters, $\text{ICL}-\text{BIC}_K$ and $\text{DIC}_{4a,K}$ as well as AWE_K are reported in Table 4. For both samples, $\text{ICL}-\text{BIC}_K$ and AWE_K select the same number of clusters both for skew-normal and skew-*t* mixtures. For the GvHDB01case sample, these criteria select $K = 9$ for skew-*t* mixtures and $K = 12$ for skew-normal mixtures. For the GvHDB06control sample, these criteria select $K = 8$ for skew-*t* mixtures and $K = 13$ for skew-normal mixtures. Both criteria clearly favor the optimal skew-*t* mixture model over the optimal skew-normal mixture for both samples. $\text{DIC}_{4a,K}$ selects the same number of clusters as $\text{ICL}-\text{BIC}_K$ and AWE_K only when a skew-normal mixture is fitted to the GvHDB06control sample. In all other cases, the number of selected clusters is higher.

The skew-*t* mixtures selected by $\text{ICL}-\text{BIC}_K$ and AWE_K were identified as described in Section B.4 of the supplementary material available at *Biostatistics* online. Parameter estimates are reported in Table 5 for the case sample. We find that several components have a small degree of freedom ν_k and that some, but not all, skewness parameters $\alpha_{k,j}$ are different from 0. A similar result is obtained for the control sample (not reported).

The MCMC draws obtained from relabeling are used for further inference as shown in the heatmap in Figure 3. Using a totally unsupervised 6-variate skew-*t* mixture modeling, the present method succeeded in discovering the signature specified by Brinkman *and others* (2007) with fewer components, see Figure 3. The case and control samples were optimally modeled with 8 and 9 skew-*t* components respectively, as shown in the heatmap. Each row in the heatmap represents 1 component from either sample. Whether a component belonged to case or control sample is marked by a pink or a green label. It is likely that superior modeling by skew-*t* mixture over symmetric *t* mixture led to a smaller number of components. Among the 17 components from both samples, grouped by the similarity of their locations, an outstanding one marked with a rectangle (bottom row in Figure 3) represented a 3.5% cell population of live cells (high Forward Scatter [FSC], high Side Scatter [SSC]) in the case sample with a clear and unique CD4+CD8β+CD3+CD8+ signature. Yet another component in the case sample of size 3.4% (fifth row from the bottom in Figure 3) may also be considered. Both components reaffirm the same GvHD-specific signature reported by Brinkman *and others* (2007).

5. Concluding remarks

We studied multivariate mixtures that introduce for each component a skewness parameter of the same dimension as the observations. A more flexible mixture could be based on more general skew-normal

Table 5. *Data set GvHDB01case; fitted skew-t mixture with $K = 9$ components; parameter estimation*

k	1	2	3	4	5	6	7	8	9
$E(\mu_{k,1}\|\mathbf{y})$	2.72	2.22	2.8	2.11	2.29	2.5	2.37	2.29	2.52
$E(\mu_{k,2}\|\mathbf{y})$	2.59	1.85	2.7	2.26	1.85	2.3	1.96	1.85	2.59
$E(\mu_{k,3}\|\mathbf{y})$	1.81	0.54	1.62	0.92	1.43	1.1	0.54	0.48	1.16
$E(\mu_{k,4}\|\mathbf{y})$	1.28	0.44	1.93	0.93	0.295	0.95	0.39	1.42	1.09
$E(\mu_{k,5}\|\mathbf{y})$	1.7	0.46	2.02	0.78	1.97	0.64	0.37	1.89	0.91
$E(\mu_{k,6}\|\mathbf{y})$	1.46	0.82	2.93	0.81	0.95	0.68	1.43	2.6	0.93
$E(\alpha_{k,1}\|\mathbf{y})$	–0.3	–2.26	–0.19	–4.56	0.48	0.08	0.04	–1.55	–5.08
SD$(\alpha_{k,1}\|\mathbf{y})$	0.44	1.86	0.63	1.58	0.22	0.09	0.11	0.81	2
$E(\alpha_{k,2}\|\mathbf{y})$	0.27	–0.85	–1.47	0.55	0.16	0.4	0.04	0.84	1.02
SD$(\alpha_{k,2}\|\mathbf{y})$	0.4	1.21	0.51	0.25	0.21	0.19	0.13	0.34	0.35
$E(\alpha_{k,3}\|\mathbf{y})$	–1.37	0.09	–1.93	0.07	–6.16	–0.8	–0.01	0.67	–0.68
SD$(\alpha_{k,3}\|\mathbf{y})$	0.5	0.23	0.53	0.2	0.84	0.21	0.17	0.27	0.43
$E(\alpha_{k,4}\|\mathbf{y})$	–0.09	0.13	0.24	–0.15	0.02	0.28	0.21	–1.82	–1.05
SD$(\alpha_{k,4}\|\mathbf{y})$	0.28	0.25	0.33	0.24	0.22	0.27	0.16	0.72	0.54
$E(\alpha_{k,5}\|\mathbf{y})$	–2.1	0.33	0.34	–0.32	0.25	–0.07	0.1	–0.05	–0.02
SD$(\alpha_{k,5}\|\mathbf{y})$	0.41	0.34	0.27	0.21	0.23	0.23	0.21	0.14	0.21
$E(\alpha_{k,6}\|\mathbf{y})$	1.74	–0.18	0.03	0.07	0.63	–0.28	–4.81	–0.23	–0.1
SD$(\alpha_{k,6}\|\mathbf{y})$	0.71	0.35	0.12	0.13	0.4	0.32	1.22	0.17	0.19
Med$(\nu_k\|\mathbf{y})$	7.3	12.3	22.2	19.5	24.8	48.9	497	3.9	18.1
$E(100\eta_k\|\mathbf{y})$	3.4	9.6	3.5	23.4	1.7	30.4	9.8	3.5	14.7

SD, standard deviation.

and skew-*t* distributions where the univariate random effect is substituted by a higher dimensional one (see, e.g. Branco and Dey, 2001; Arellano-Valle and Azzalini, 2006). Our MCMC scheme may be easily extended to such a mixture.

Although our MCMC scheme is quite efficient, we see scope for improvement. Parameter expansion similar in spirit to van Dyk and Meng (2001) could be implemented by running MCMC for an expanded, unidentified model with the random effects distributed as $z_i \sim \mathcal{TN}_{[\alpha_k,\infty)}(\alpha_k, \beta_k)$. To improve mixing for multimodal posteriors in the context of clustering high-dimensional data sets ideas from evolutionary Monte Carlo as discussed, for example, in Liang and Wong (2001) could be considered.

Finally, we end by highlighting the potential application of the robust and precise data modeling by our method to high-throughput and high-dimensional platforms such as flow cytometry.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknoweldgments

We thank the anonymous reviewers and the editor for numerous useful suggestions and comments. Furthermore, Saumyadipta Pyne would like to thank Philip De Jager and Joshua Shulman for details about the Alzheimer's data. *Conflict of Interest:* None declared.

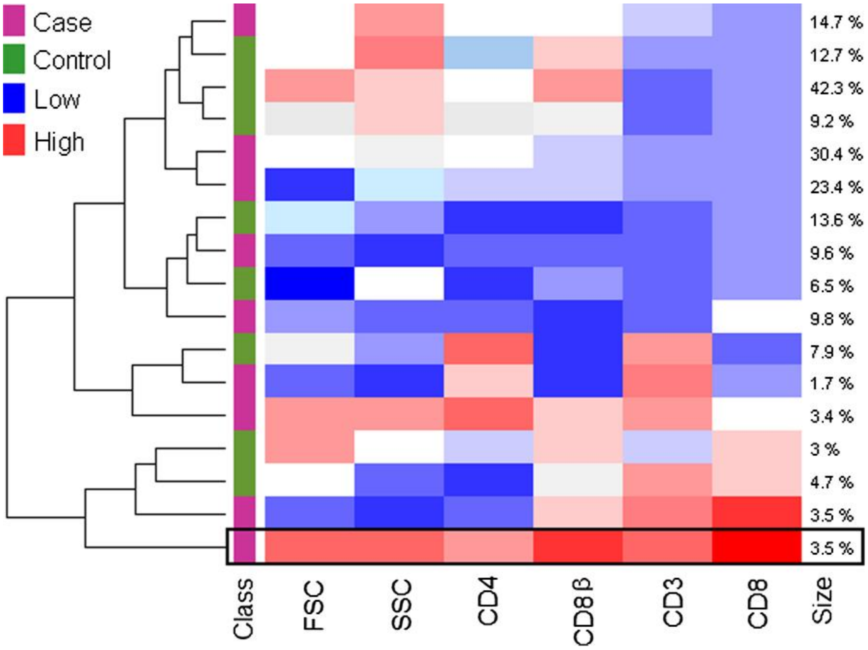

Fig. 3. Skew-t mixture modeling of GvHD case and control samples from Brinkman *and others* (2007) data identifies component with unique marker signature. In the heatmap, each row represents the location of a 6D cluster from either the case or the control sample and each column represents a particular marker. Whether a component belonged to the case or the control sample is marked by a pink or a green label. Based on ICL−BIC, the control sample was optimally modeled with nine 6D skew-t components, and the case sample with 8 components. The red, blue, and white colors denote high, low, and medium expression, respectively. Among all components, the one marked with a rectangle represents live cells (high FSC, high SSC) from the case sample with a unique CD4+CD8β+CD3+ signature.

References

Arellano-Valle, R. B. and Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* **33**, 561–574.

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

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* **46**, 199–208.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society, Series B* **65**, 367–389.

Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew normal distribution. *Biometrika* **83**, 715–726.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821.

Bennett, D. A., Schneider, J. A., Buchman, A. S., de Leon, C. M., Bienias, J. L. and Wilson, R. S. (2005). The rush memory and aging project: study design and baseline characteristics of the study cohort. *Neuroepidemiology* **25**, 163–175.

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 719–725.

BIERNACKI, C. AND GOVAERT, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics* **29**, 451–457.

BOEDIGHEIMER, M. J. AND FERBAS, J. (2008). Mixture modeling approach to flow cytometry data. *Cytometry Part A* **73**, 421–429.

BRANCO, M. D. AND DEY, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis* **79**, 99–113.

BRINKMAN, R., GASPARETTO, M., LEE, S.-J., RIBICKAS, A., PERKINS, J., JANSSEN, W., SMILEY, R. AND SMITH, C. (2007). High content flow cytometry and temporal data analysis for defining a cellular signature of graft versus host disease. *Biology of Blood and Marrow Transplantation* **13**, 691–700.

CABRAL, C., BOLFARINE, H. AND PEREIRA, J. (2008). Bayesian density estimation using skew student-t-normal mixtures. *Computational Statistics and Data Analysis* **52**, 5075–5090.

CELEUX, G., FORBES, F., ROBERT, C. P. AND TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis* **1**, 651–674.

CELEUX, G., HURN, M. AND ROBERT, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95**, 957–970.

CHAN, C., FENG, F., OTTINGER, J., FOSTER, D., WEST, M. AND KEPLER, T. (2008). Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A* **73**, 693–701.

DELLAPORTAS, P. AND PAPAGEORGIOU, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing* **16**, 57–68.

DIEBOLT, J. AND ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B* **56**, 363–375.

FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* **96**, 194–209.

FRÜHWIRTH-SCHNATTER, S. (2004). Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *The Econometrics Journal* **7**, 143–167.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.

GENTON, M. G. (editor) (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Boca Raton, FL. Chapman & Hall/CRC.

HENZE, N. (1986). A probabilistic representation of the skew-normal distribution. *Scandinavian Journal of Statistics* **13**, 271–275.

JASRA, A., HOLMES, C. C. AND STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science* **20**, 50–67.

JASRA, A., STEPHENS, D. A., GALLAGHER, K. AND HOLMES, C. C. (2006). Bayesian mixture modelling in geochronology via Markov chain Monte Carlo. *Mathematical Geology* **38**, 269–300.

JENNISON, C. (1997). Discussion of the paper by Richardson and Green. *Journal of the Royal Statistical Society, Series B* **59**, 778–779.

JUÁREZ, M. A. AND STEEL, M. F. J. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business and Economic Statistics* **28**, 52–66.

KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhya A* **62**, 49–66.

KRUGLYAK, L. AND LANDER, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

LIANG, F. AND WONG, W. H. (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association* **96**, 653–666.

LIN, T. I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis* **100**, 257–265.

LIN, T. I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing* (in press).

LIN, T. I., LEE, J. C. AND HSIEH, W. J. (2007). Robust mixture modeling using the skew t-distribution. *Statistics and Computing* **17**, 81–92.

LIN, T. I., LEE, J. C. AND NI, H. F. (2004). Bayesian analysis of mixture modelling using the multivariate t-distribution. *Statistics and Computing* **14**, 119–130.

LIN, T. I., LEE, J. C. AND YEN, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica* **17**, 909–927.

LO, K., BRINKMAN, R. R. AND GOTTARDO, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A* **73**, 321–332.

MCLACHLAN, G. J. AND PEEL, D. (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley.

MENG, X.-L. AND WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* **6**, 831–860.

NEAL, R. N. (2001). Annealed importance sampling. *Statistics and Computing* **11**, 125–139.

NOBILE, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics* **32**, 2044–2073.

PEEL, D. AND MCLACHLAN, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* **10**, 339–348.

PERFETTO, S. P., CHATTOPADHYAY, P. K. AND ROEDERER, M. (2004). Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunology* **4**, 648–655.

PYNE, S., HU, X., WANG, K., ROSSIN, E., LIN, T., MAIER, L. M., BAECHER-ALLAN, C., MCLACHLAN, G. J., TAMAYO, P., HAFLER, D. A. *and others* (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of National Academy of Sciences of the United States of America* **106**, 8519–8524.

RICHARDSON, S. AND GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B* **59**, 731–792.

ROSES, A. D. (1997). A model for susceptibility polymorphisms for complex diseases: apolipoprotein E and Alzheimer disease. *Neurogenetics* **1**, 3–11.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. AND VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–639.

STEPHENS, M. (1997). Bayesian methods for mixtures of normal distributions, [PhD. Thesis]. University of Oxford, Oxford.

STEPHENS, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B* **62**, 795–809.

VAN DYK, D. AND MENG, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**, 1–50.

WILSON, R., BIENIAS, J., EVANS, D. AND BENNETT, D. (2004). The religious orders study: overview and change in cognitive and motor speed. *Aging, Neuropsychology, and Cognition* **11**, 280–303.

WILSON, R., SCHNEIDER, J., BARNES, L., BECKETT, L., AGGARWAL, N., COCHRAN, E., BERRY-KRAVIS, E., BACH, J., FOX, J., EVANS, D. *and others* (2002). The apolipoprotein E e4 allele and decline in different cognitive systems during a 6-year period. *Archives of Neurology* **59**, 1154–1160.

[*Received December 23, 2009; revised December 23, 2009; accepted for publication December 24, 2009*]