



Department for Applied Statistics  
Johannes Kepler University Linz



## IFAS Research Paper Series 2009-46

# Bayesian Inference for Finite Mixtures of Univariate and Multivariate Skew Normal and Skew- $t$ Distributions

Sylvia Frühwirth-Schnatter<sup>a</sup> and  
Saumyadipta Pyne<sup>b</sup>

November 2009

---

<sup>a</sup>Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, Linz, Austria, [sylvia.fruehwirth-schnatter@jku.at](mailto:sylvia.fruehwirth-schnatter@jku.at)

<sup>b</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA, [saumyadipta\\_pyne@dfci.harvard.edu](mailto:saumyadipta_pyne@dfci.harvard.edu)

## Abstract

Skew Normal and skew- $t$  distributions 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 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 which draws from standard densities, only. This MCMC scheme is extended to mixtures of skew- $t$  distributions based on representing the skew- $t$  distribution as 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 normals and skew- $t$  mixture models, we could model non-Gaussian cell populations rigorously and directly without transformation or projection to lower dimensions.

*Key words:* Flow cytometry, Markov chain Monte Carlo, Gibbs sampling, kurtosis, 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  $\mathbf{Y}$  from a finite mixture distribution. This leads to the standard finite mixture model considered e.g. 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 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,

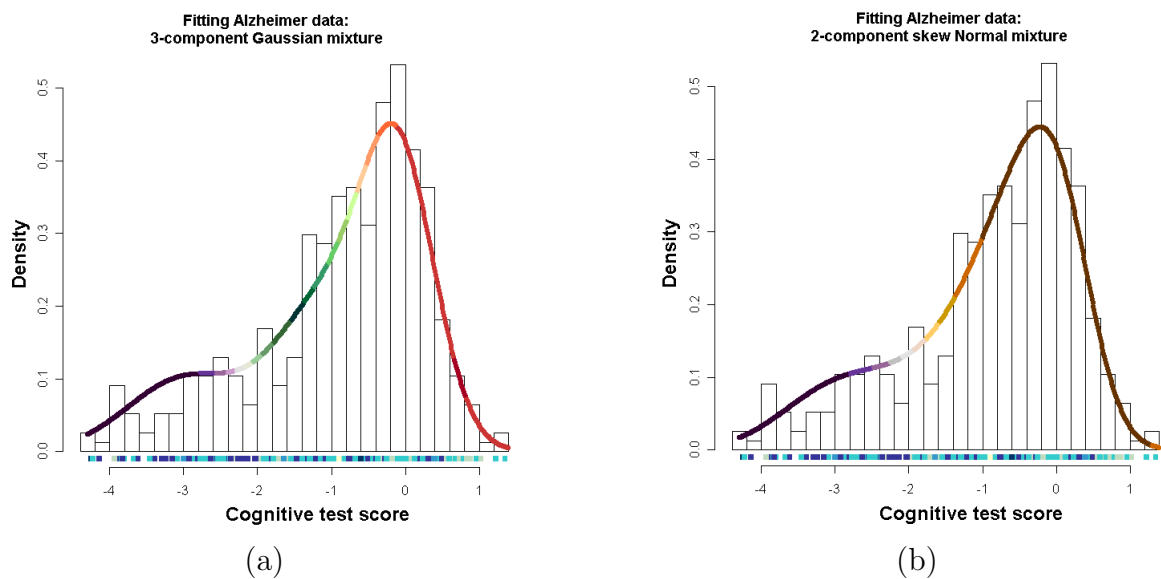
because multiple normal distributions may be necessary to capture skewness and kurtosis of a single cluster, thus leading to wrong inference about the number of clusters in data (Jasra, Stephens, Gallagher, and Holmes, 2006). Similarly, in the context of supervised learning, groups of observations represented 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 which will be analyzed in detail in Subsection 4.1. The left hand side of Figure 1 shows the result of fitting a three component mixture of normal distributions which correspond to the optimal number of components as will be demonstrated in Subsection 4.1. Interestingly, the bimodality of the fitted mixture indicates the presence of two clusters, however, the normal mixture needs two 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, Lee, and Ni (2004) which allow for heavy tails of each component. Very recently, application of finite mixture models have been to the univariate skew normal distribution (Lin, Lee, and Yen, 2007), to the univariate skew- $t$  distribution (Jasra et al., 2006; Lin, Lee, and Hsieh, 2007), to the univariate skew student- $t$ -normal distribution (Cabral, Bolfarine, and Pereira, 2008) as well as to the multivariate skew normal distribution (Lin, 2009a) and the multivariate skew- $t$  distribution (Lin, 2009b; Pyne, Hu, Wang, Rossin, Lin, Maier, Baecher-Allan, McLachlan, Tamayo, Hafler, De Jager, and Mesirov, 2009).

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 Alzheimer's Disease introduced above and will show that the optimal mixture of skew normal distributions needs only two components to fit the observed distribution, see the right hand side of Figure 1 . In addition, we consider

Figure 1: Gaussian and skew normal mixture modeling of Alzheimer’s disease data set. The histogram, common to plots (a) and (b), 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 blue point in the rug indicates more e4 alleles in a subject’s genotype implying higher risk factor for AD. In plot (a), fitting of a 3-component Gaussian mixture is shown with each component in distinct color (purple, green or red). In plot (b), fitting of a 2-component skew normal mixture is shown with each component in distinct color (purple or brown).



clustering multivariate flow cytometric data from Graft versus Host Disease. 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, Chattopadhyay, and Roederer, 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 two dimensions. 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 data.

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 MCMC. Towards this, we use a representation of the skew normal and the skew- $t$  distribution which 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 rather straightforward MCMC sampling scheme which involves a two-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 which 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 shortly 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 Alzheimer's Disease and multivariate cytometric data from Graft versus

## 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  $\alpha$ ,  $X \sim \mathcal{SN}(\alpha)$ , if the density takes the form  $p(x|\alpha) = 2\phi(x)\Phi(\alpha x)$ , where  $\phi(\cdot)$  and  $\Phi(\cdot)$  are, respectively, the pdf and the 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

$$X = \delta Z + \sqrt{1 - \delta^2} \varepsilon \quad (1)$$

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 arbitrary location and scale, a location parameter  $\xi \in \mathfrak{R}$  and a scale parameter  $\omega \in \mathfrak{R}^+$  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:

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

A stochastic representation of the  $\mathcal{SN}(\xi, \omega^2, \alpha)$  distribution is obtained by applying the affine transformation  $Y = \xi + \omega X$  to (1):

$$Y = \xi + \omega \delta Z + \omega \sqrt{1 - \delta^2} \varepsilon, \quad (3)$$

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 (1). The  $r$  components of a multivariate random variable  $\mathbf{X} = (X_1, \dots, X_r)' \in \mathfrak{R}^r$  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  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_r)' \sim \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Omega}_\varepsilon)$  is independent of  $Z$  and multivariate normal with arbitrary correlation matrix  $\boldsymbol{\Omega}_\varepsilon$ . Applying the affine transformation  $\mathbf{Y} = \boldsymbol{\xi} + \boldsymbol{\omega}\mathbf{X}$  with location parameter  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_r)' \in \mathfrak{R}^r$  and diagonal scale matrix  $\boldsymbol{\omega} = \text{Diag}(\omega_1, \dots, \omega_r)$  with  $\omega_j > 0$  immediately leads to the stochastic representation

$$Y_j = \xi_j + \omega_j \delta_j Z + \omega_j \sqrt{1 - \delta_j^2} \varepsilon_j. \quad (4)$$

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

$$f_{\mathcal{SN}}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = 2\phi_r(\mathbf{y} - \boldsymbol{\xi}; \boldsymbol{\Omega})\Phi(\boldsymbol{\alpha}'\boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\xi})), \quad (5)$$

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  $\boldsymbol{\Omega}_\varepsilon$  in the stochastic representation (4) through:

$$\boldsymbol{\Omega} = \boldsymbol{\omega}\bar{\boldsymbol{\Omega}}\boldsymbol{\omega}, \quad \boldsymbol{\alpha} = \frac{1}{\sqrt{1 - \boldsymbol{\delta}'\boldsymbol{\delta}}}\bar{\boldsymbol{\Omega}}^{-1}\boldsymbol{\delta}, \quad (6)$$

where  $\bar{\boldsymbol{\Omega}} = \boldsymbol{\Delta}\boldsymbol{\Omega}_\varepsilon\boldsymbol{\Delta} + \boldsymbol{\delta}\boldsymbol{\delta}'$  and  $\boldsymbol{\Delta} = \text{Diag}(\sqrt{1 - \delta_1^2}, \dots, \sqrt{1 - \delta_r^2})$ . The matrix  $\bar{\boldsymbol{\Omega}}$  is a correlation matrix, because  $\bar{\boldsymbol{\Omega}}_{jj} = (1 - \delta_j^2)(\boldsymbol{\Omega}_\varepsilon)_{jj} + \delta_j^2 = 1$ , 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  $(\boldsymbol{\delta}, \boldsymbol{\omega}, \boldsymbol{\Omega}_\varepsilon)$  in the stochastic representation (4) are obtained from:

$$\boldsymbol{\delta} = \frac{1}{\sqrt{1 + \boldsymbol{\alpha}'\bar{\boldsymbol{\Omega}}\boldsymbol{\alpha}}}\bar{\boldsymbol{\Omega}}\boldsymbol{\alpha}, \quad \boldsymbol{\Omega}_\varepsilon = \boldsymbol{\Delta}^{-1}\bar{\boldsymbol{\Omega}}\boldsymbol{\Delta}^{-1} - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}', \quad (7)$$

where  $\bar{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$  with  $\boldsymbol{\omega} = \text{Diag}(\boldsymbol{\Omega})^{1/2}$  being a diagonal matrix obtained from the diagonal elements of  $\boldsymbol{\Omega}$ ,  $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}_1, \dots, \tilde{\alpha}_r)'$  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

$$\mathbf{E}(\mathbf{Y}) = \boldsymbol{\xi} + \boldsymbol{\omega}\boldsymbol{\delta}\sqrt{\frac{2}{\pi}}. \quad (8)$$

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:

$$Y = \xi + \omega \frac{X}{\sqrt{W}}, \quad (9)$$

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

$$f_{\mathcal{ST}}(y; \xi, \omega^2, \alpha, \nu) = \frac{2}{\omega} t_\nu(x_y) T_{\nu+1} \left( \alpha x_y \sqrt{\frac{\nu+1}{\nu+x_y^2}} \right), \quad (10)$$

where  $x_y = (y - \xi)/\omega$  and  $t_\nu$  and  $T_\nu$  denote, respectively, the pdf and the cdf of a standard Student- $t$  distribution with  $\nu$  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:

$$\mathbf{Y} = \boldsymbol{\xi} + \frac{1}{\sqrt{W}} \mathbf{X}, \quad (11)$$



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

$$f_{\mathcal{ST}}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) = 2f_{t_r}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu)T_{\nu+r} \left( \boldsymbol{\alpha}' \boldsymbol{\omega}^{-1}(\mathbf{y} - \boldsymbol{\xi}) \sqrt{\frac{\nu+r}{\nu+Q_y}} \right), \quad (12)$$

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  $T_\nu$  denotes the cdf of the scalar standard Student- $t$  distribution as above. The skew- $t$  distribution converges to the skew normal distribution as  $\nu \rightarrow \infty$ . For any  $r \geq 1$ , the expectation of the skew- $t$  distribution, provided that  $\nu > 1$ , is given by:

$$\mathbb{E}(\mathbf{Y}) = \boldsymbol{\xi} + \boldsymbol{\omega} \boldsymbol{\mu}_X, \quad \boldsymbol{\mu}_X = \boldsymbol{\delta} \sqrt{\frac{\nu}{\pi}} \frac{\Gamma((\nu-1)/2)}{\Gamma(\nu/2)}. \quad (13)$$

### 3 Skew Normal and Skew- $t$ Finite Mixture Models

We consider univariate and multivariate finite mixture models where the component densities  $p(\mathbf{y}_i | \boldsymbol{\theta}_k)$ ,  $k = 1, \dots, K$ , 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  $\sum_{k=1}^K \eta_k = 1$ , e.g. for a mixture of scalar skew normal distributions:

$$p(y_i | \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K, \boldsymbol{\eta}) = \eta_1 f_{\mathcal{SN}}(y_i; \xi_1, \omega_1^2, \alpha_1) + \dots + \eta_K f_{\mathcal{SN}}(y_i; \xi_K, \omega_K^2, \alpha_K), \quad (14)$$

or for a mixture of multivariate skew- $t$  distributions:

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

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 rather 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 a hierarchical latent variable model, 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 (14) the application of (3) to each component density leads to 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 parametrization in terms of the component specific parameters  $\boldsymbol{\theta}_k^* = (\xi_k, \psi_k, \sigma_k^2)$ , where  $\psi_k = \omega_k \delta_k$  and  $\sigma_k^2 = \omega_k^2(1 - \delta_k^2)$ :

$$\begin{aligned} z_i &\sim \mathcal{TN}_{[0,\infty)}(0, 1), \\ y_i|(S_i = k) &= \xi_k + \psi_k z_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_k^2). \end{aligned} \tag{15}$$

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

$$\alpha_k = \frac{\psi_k}{\sigma_k}, \quad \omega_k^2 = \sigma_k^2 + \psi_k^2, \tag{16}$$

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

Representation (15) offers several advantages. First, a conditionally conjugate prior for  $\boldsymbol{\theta}_k^* = (\xi_k, \psi_k, \sigma_k^2)$  is available and, second, straightforward MCMC estimation using a two-block Gibbs sampler becomes feasible, see Subsection 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 et al. (2007), however,  $\xi_k$ ,  $\omega_k^2$  and  $\delta_k$  are sampled in different blocks and a Metropolis-Hastings algorithm is needed to sample  $\delta_1, \dots, \delta_K$ , while representation (15) 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  $\mathcal{SN}_r(\xi_k, \Omega_k, \alpha_k)$ ,  $k = 1, \dots, K$ , where the componentwise application of (4) leads to a mixture of random effects models with repeated measurements and a univariate truncated normal random effect:

$$\begin{aligned} z_i &\sim \mathcal{TN}_{[0,\infty)}(0, 1), \\ \mathbf{y}_i | (S_i = k) &= \boldsymbol{\xi}_k + \boldsymbol{\psi}_k z_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Sigma}_k), \end{aligned} \quad (17)$$

with  $z_1, \dots, z_N$  and  $\boldsymbol{\epsilon}_1, \dots, \boldsymbol{\epsilon}_N$  being mutually independent. We introduced the parametrization  $\boldsymbol{\theta}_k^* = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$ , 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 (7):  $\boldsymbol{\Sigma}_k = \boldsymbol{\omega}_k \boldsymbol{\Delta}_k (\boldsymbol{\Omega}_\varepsilon)_k \boldsymbol{\Delta}_k \boldsymbol{\omega}_k = \boldsymbol{\Omega}_k - \boldsymbol{\omega}_k \boldsymbol{\Delta}_k \tilde{\boldsymbol{\alpha}}_k (\boldsymbol{\omega}_k \boldsymbol{\Delta}_k \tilde{\boldsymbol{\alpha}}_k)'$ . 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:

$$\boldsymbol{\Omega}_k = \boldsymbol{\Sigma}_k + \boldsymbol{\psi}_k \boldsymbol{\psi}_k', \quad \boldsymbol{\alpha}_k = \frac{1}{\sqrt{1 - \boldsymbol{\psi}_k' \boldsymbol{\Omega}_k^{-1} \boldsymbol{\psi}_k}} \boldsymbol{\omega}_k \boldsymbol{\Omega}_k^{-1} \boldsymbol{\psi}_k. \quad (18)$$

For skew- $t$  mixtures we combine the stochastic representations (9) or (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:

$$w_i | (S_i = k) \sim \mathcal{G}\left(\frac{\nu_k}{2}, \frac{\nu_k}{2}\right), \quad (19)$$

$$z_i | w_i \sim \mathcal{TN}_{[0,\infty)}\left(0, \frac{1}{w_i}\right), \quad (20)$$

$$y_i | (S_i = k, w_i) = \xi_k + \psi_k z_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_k^2 / w_i), \quad (21)$$

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  $\mathcal{ST}_r(\xi_k, \Omega_k, \alpha_k, \nu_k)$ ,  $k = 1, \dots, K$ , has

a similar representation with a repeated measurements observation equation:

$$\mathbf{y}_i | (S_i = k, w_i) = \boldsymbol{\xi}_k + \boldsymbol{\psi}_k z_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_r \left( \mathbf{0}, \frac{1}{w_i} \boldsymbol{\Sigma}_k \right), \quad (22)$$

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 (15) or (17) a  $\mathcal{TN}_{[0, \infty)}(0, 1)$  random variable with  $1/\sqrt{w_i}$ .

As for skew normal mixtures, we use an alternative parametrization with component specific parameter  $\boldsymbol{\theta}_k^* = (\xi_k, \psi_k, \sigma_k^2, \nu_k)$  and  $\boldsymbol{\theta}_k^* = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \nu_k)$ , respectively. This allows Bayesian estimation through a three-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 “non-informative 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, Subsection 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  $\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*$  introduced in Subsection 3.1 rather than directly for  $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K$ . Using the representations discussed in Subsection 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  $\nu_k$  is a slight modification of a prior introduced by Juárez and Steel (2009) for Student- $t$  mixtures with  $\nu_1 = \dots = \nu_K$ . Further details for all priors are provided in Appendix A.

Following the seminal paper by Diebolt and Robert (1994), the most popular method for Bayesian estimation of finite mixtures is to apply Markov chain Monte Carlo methods based data augmentation and Gibbs sampling, 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 Subsection 3.1. We introduce the latent allocations  $\mathbf{S} = (S_1, \dots, S_N)$  and the latent random effects  $\mathbf{z} = (z_1, \dots, z_N)$  as missing data and add the latent scaling factors  $\mathbf{w} = (w_1, \dots, w_N)$  for skew- $t$  mixtures. MCMC sampling is based on following observations.

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

Second, conditional on  $\mathbf{S}$  (and  $\mathbf{w}$ ) we consider skew normal and skew- $t$  mixtures as a random effects model 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 Appendix B.1. This allows joint multi-move sampling of the latent variables  $\mathbf{S}$  and  $\mathbf{z}$ .

Third, conditional on  $\mathbf{S}$ ,  $\mathbf{z}$ , (and  $\mathbf{w}$ ) sampling of the transformed component specific parameters  $\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*$  (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. Joint sampling of  $\boldsymbol{\xi}_k$ ,  $\boldsymbol{\psi}_k$  and  $\boldsymbol{\Sigma}_k$  from a closed form posterior distribution is possible, because the priors introduced in Appendix A are conditionally conjugate.

As a result, MCMC estimation for skew normal mixtures is possible through a two-step

Gibbs sampler if the hyperparameters  $C_0$  and  $\mathbf{C}_0$  are fixed:

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

All conditional densities are of closed form, see Appendix B.2 which also contains details on how to sample  $C_0$  or  $\mathbf{C}_0$  under the hierarchical prior (31). For MCMC estimation of skew- $t$  mixtures a third step has to be added:

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

All conditional densities except  $p(\nu_1, \dots, \nu_K | \cdot)$  are of closed form, see Appendix B.3 for details.

Like for any finite mixture model, a non-identifiability 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 e.g. (14). 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 to deal with this problem, see, e.g., Celeux, Hurn, and Robert (2000), Stephens (2000b), and Jasra, Holmes, and Stephens (2005). Here we follow Frühwirth-Schnatter (2001) who suggested to add a random permutation step to the MCMC scheme and to post-process the resulting MCMC output to identify component-specific parameters, see Appendix B.4 for more details.

### 3.3 Selecting the Number of Components

Selecting the number of components of a finite mixture model is quite a challenge, see Frühwirth-Schnatter (2006, Chapter 5) for a recent review. Popular methods are implementing reversible jump MCMC, computing marginal likelihoods, or using 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 e.g. Dellaportas

and Papageorgiou (2006) for multivariate normal mixtures. Since adding skewness even complicates matters, we did not pursue RJMCMC.

Alternatively, the choice of  $K$  may be based 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  $K \leq 5$ , 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 iid 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 consider. One such criterion is  $\text{BIC}_K$  which is an asymptotic approximation to  $-2 \log p(\mathbf{y}|\mathcal{M}_K)$ :

$$\text{BIC}_K = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}_K, \mathcal{M}_K) + d_K \log N, \quad (23)$$

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 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  $\text{BIC}_K$  is known to be consistent (Keribin, 2000), although in small data sets it tends to choose models with too few components (Biernacki, Celeux, and Govaert, 2000). On the other hand, simulation studies reported in Biernacki and Govaert (1997), Biernacki et al. (2000), and McLachlan and Peel (2000, Section 6.11) show that  $\text{BIC}_K$  will overrate the number of clusters under misspecification of the component density, whereas several alternative criteria such as the  $\text{AWE}_K$  and the  $\text{ICL}_K$  criterion to be discussed below are able

to identify the correct number of clusters even when the component densities are misspecified. Thus  $\text{BIC}_K$  for clustering large data sets, where the component densities of the finite mixture model may not be correctly specified, is likely to be imprecise.

Approximate weight of evidence ( $\text{AWE}_K$ ) is derived in Banfield and Raftery (1993) as another approximation to minus twice the log of the marginal likelihood.  $\text{AWE}_K$  is described in Biernacki and Govaert (1997) as a criterion which penalizes the log of the complete-data likelihood function with model complexity:

$$\text{AWE}_K = -2 \log p(\mathbf{y}, \hat{\mathbf{S}} | \hat{\boldsymbol{\vartheta}}_K^C) + 2 d_K \left( \frac{3}{2} + \log N \right), \quad (24)$$

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  $\log p(\mathbf{y}, \mathbf{S} | \boldsymbol{\vartheta}_K) = \sum_{i=1}^N \log (\eta_{S_i} p(\mathbf{y}_i | \boldsymbol{\theta}_{S_i}))$ . Again, approximate estimators are determined as the posterior draw maximizing the complete-data likelihood function.

Biernacki et al. (2000) introduced the integrated classification likelihood criterion  $\text{ICL}_K$  which has been shown by McLachlan and Peel (2000, p.216) to be approximately equal to

$$\text{ICL-BIC}_K = \text{BIC}_K + 2\text{EN}(\hat{\boldsymbol{\vartheta}}_K). \quad (25)$$

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

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

and measures how well the finite mixture model defined by  $\boldsymbol{\vartheta}_K$  classifies the data into  $K$  distinct clusters. Thus the  $\text{ICL-BIC}_K$  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, Best, Carlin, and van der Linde (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, Forbes, Robert, and Titterton (2006). A first problem is the choice of the appropriate likelihood function which 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 et al. (2006) investigate in total 8 different DIC criteria.  $\text{DIC}_2$ , 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:

$$\text{DIC}_{2,K} = -4\mathbb{E}_{\boldsymbol{\vartheta}_K} (\log p(\mathbf{y}|\boldsymbol{\vartheta}_K)|\mathbf{y}) + 2\log p(\mathbf{y}|\hat{\boldsymbol{\vartheta}}_K^M, \mathbf{y}), \quad (26)$$

where the posterior mode estimator  $\hat{\boldsymbol{\vartheta}}_K^M$  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 et al. (2006) recommend using  $\text{DIC}_4$  which is based on computing first DIC for the complete-data likelihood function and then integrating with respect to the posterior  $p(\mathbf{S}|\mathbf{y}, \mathcal{M}_K)$ :

$$\text{DIC}_{4,K} = -4\mathbb{E}_{\boldsymbol{\vartheta}_K, \mathbf{S}} (\log p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta}_K)|\mathbf{y}) + 2\mathbb{E}_{\mathbf{S}} \left( \log p(\mathbf{y}, \mathbf{S}|\hat{\boldsymbol{\vartheta}}_K(\mathbf{S}))|\mathbf{y} \right). \quad (27)$$

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 et al. (2006) show that substituting  $\hat{\boldsymbol{\vartheta}}_K(\mathbf{S})$  by the posterior mode estimator  $\hat{\boldsymbol{\vartheta}}_K^M$ , an approximation to  $\text{DIC}_{4,K}$  is obtained which penalizes  $\text{DIC}_{2,K}$  by the expected entropy:

$$\text{DIC}_{4a,K} = \text{DIC}_{2,K} + 2\mathbb{E}_{\boldsymbol{\vartheta}_K} (\text{EN}(\boldsymbol{\vartheta}_K)|\mathbf{y}). \quad (28)$$

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

both  $\log p(\mathbf{y}|\boldsymbol{\vartheta}_K)$  as well as  $\text{EN}(\boldsymbol{\vartheta}_K)$  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

Alzheimer’s disease (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 (ROS), see Wilson, Bienias, Evans, and Bennett (2004) and the Memory and Aging Project (MAP), see Bennett, Schneider, Buchman, de Leon, Bienias, and Wilson (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 Appendix A with different sets of hyper parameters. For all priors,  $b_0^\psi = b_0^\xi = 0$  and  $g_0 = 0.5$ , while we consider four different settings for  $D^\xi = D^\psi$ ,  $c_0$  and  $\phi$  and two 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  $\xi$  and the skewness parameter  $\psi$ , prior 3 introduces stronger smoothing for the group specific variances  $\sigma_1^2, \dots, \sigma_K^2$ , 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  $\nu_k$  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

Table 1: Choosing the hyper parameters  $D^\xi$ ,  $D^\psi$ ,  $c_0$ , and  $\phi$  of the prior in skew normal and skew- $t$  mixture modeling of Alzheimer’s disease data set;  $R^2 = 1 - \phi/(c_0 - 1)$  is the prior expectation of explained heterogeneity;  $d$  is an additional hyper parameter for skew- $t$  mixtures

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

draws by using the MCMC schemes described in Appendix B.2 and B.3.

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

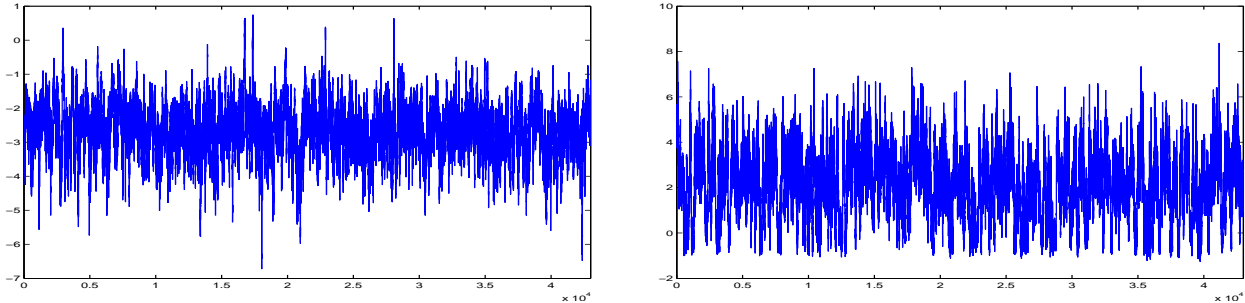
For skew normal mixtures, both the marginal likelihood as well as  $\text{BIC}_K$  select a model with two components for all priors considered. In contrast to that,  $\text{DIC}_{2,K}$  shows high sensitivity to prior choices and the selected number of components ranges from two to four. For skew- $t$  mixtures, we find that  $\text{BIC}_K$  favors a two component mixture, however, this model is outperformed by the two-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 two components for all priors. For skew- $t$  mixtures, sensitivity of  $\text{DIC}_{2,K}$  to prior choices is even higher and the selected number of components ranges from one to four.

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

Table 2: Selecting the number  $K$  of components in Gaussian and skew normal and skew- $t$  mixture modeling of Alzheimer's disease data set

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

Figure 2: Two component skew normal mixture modeling of Alzheimer’s disease data set. Posterior draws obtained under prior 1 for the skewness parameters  $\alpha_1$  (left hand side) and  $\alpha_2$  (right hand side) after identification



as in Frühwirth-Schnatter (2004) and  $p(\mathcal{M}_K)$  is the same as above. The (non-normalized) posterior probabilities  $\log(p(\mathbf{y}|\mathcal{M}_K)p(\mathcal{M}_K))$  together with  $\text{BIC}_K$  and  $\text{DIC}_{2,K}$  are reported in Table 2. The first two criteria select a normal mixture with three components, while  $\text{DIC}_{2,K}$  leads to choosing  $K = 4$ .

When the three component normal mixture is compared with the two component skew normal mixture the latter one is preferred by the marginal likelihood and  $\text{BIC}_K$ , 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 Figures 1a and 1b), the normal mixture needs two components to fit the skewness in the second cluster.

The two component skew normal mixture is identified for each prior as described in Appendix B.4. Figure 2 shows the resulting posterior draws of  $\alpha_1$  and  $\alpha_2$  for prior 1. The estimated parameters are reported for all priors in Table 3. Evidently, the skewness parameter  $\alpha_k$  is sensitive to selecting the prior information  $D^\xi$  and  $D^\psi$  which is much smaller under prior 2 than for the other priors. On the other hand, the expected cognitive score  $\mu_k$  and the group sizes  $\eta_k$  are insensitive to prior choices. For all priors, the first component has a much higher expected cognitive score  $\mu_k$  than the second one and exhibit considerable negative skewness. The skewness parameter  $\alpha_k$  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, Schneider, Barnes, Beckett, Aggarwal, Cochran, Berry-Kravis, Bach, Fox, Evans, and Bennett, 2002; Roses, 1997). There are 3 different allele polymorphisms of the gene in

Table 3: Two component skew normal mixture modeling of Alzheimer’s disease data set. Parameter estimation under different priors using posterior means (posterior standard deviations in parenthesis).

	$k$	$\xi_k$	$\omega_k^2$	$\alpha_k$	$\mu_k = E(Y S_i = k)$	$\eta_k$
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)

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.

First, we used the skew normal mixture model to classify each subject into one of the two 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 6 colors, 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 one skewed components into two 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.

In common practice, a flow cytometric data analyst looks at the high-dimensional flow readout in 2-dimensional 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, Feng, Ottinger, Foster, West, and Kepler, 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 multi-modal, multi-dimensional 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.

Table 4: Choosing  $K$  for the flow cytometric data.

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

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 Graft versus Host Disease (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, Gasparetto, Lee, Ribickas, Perkins, Janssen, Smiley, and Smith (2007), which may be referred to for further details. Brinkman et al. (2007) observed an increased proportion in the CD4+CD8 $\beta$ +CD3+ population to be correlated with the development of GvHD.

Recently, Lo, Brinkman, and Gottardo (2008) used an EM-based Student- $t$  mixture model, with Box-Cox transformation to diminish the asymmetry of populations in the sample. Their optimal model for the Brinkman et al. (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 8,691 cells. For Bayesian estimation we use the priors introduced in Appendix A 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  $K$



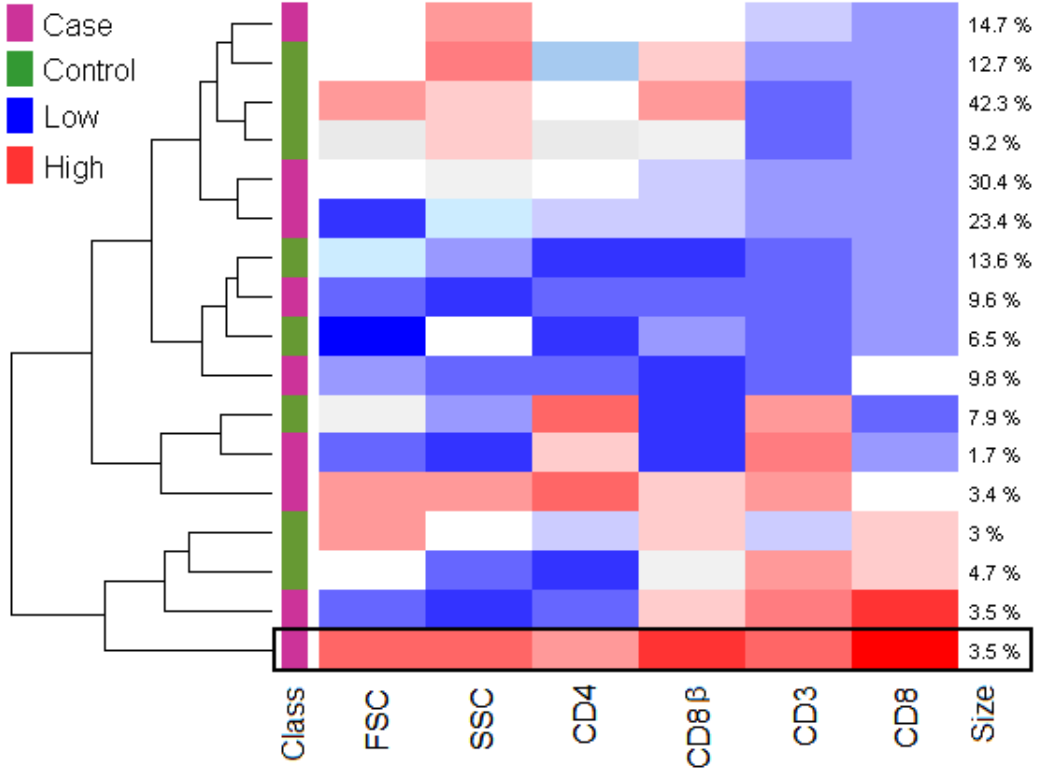
Table 5: Data set `GvHDB01case`; fitted skew- $t$  mixture with  $K = 9$  components; parameter estimation

$k$	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

several independent chains, each with 10,000 MCMC draws after a burn-in of 5,000 draws using the MCMC schemes described in Appendix B.2 and B.3. 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 (23). This guarantees that we are dealing with posterior draws from a modal region with high posterior probability, because  $-0.5BIC_K$  is a rough estimate of the marginal likelihood of a model where the parameters are restricted to each modal region.

To select the optimal number of components, various criteria introduced in Subsection 3.3 were computed both for skew normal and skew- $t$  mixtures. Since we would like to find well-separated clusters,  $ICL-BIC_K$  and  $DIC_{4a,K}$  as well as  $AWE_K$  are reported in Table 4. For both samples,  $ICL-BIC_K$  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

Figure 3: Skew- $t$  mixture modeling of GvHD case and control samples from Brinkman et al. (2007) data identifies component with unique marker signature. In the heatmap, each row represents the location of a six-dimensional 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 9 six-dimensional 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+CD8 $\beta$ +CD3+ signature.



clearly favor the optimal skew- $t$  mixture model over the optimal skew normal mixture for both samples.  $DIC_{4a,K}$  selects the same number of clusters as  $ICL-BIC_K$  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  $ICL-BIC_K$  and  $AWE_K$  were identified as described in Appendix B.4. Parameter estimates are reported in Table 5 for the case sample. We find that several components have a small degree of freedom  $\nu_k$  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 as described in Appendix B.4 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 et al. (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 one 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 FSC, high SSC) in the case sample with a clear and unique CD4+CD8 $\beta$ +CD3+CD8+ signature. Yet another component in the case sample of size 3.4% (represented in the fifth row from the bottom in Figure 3) may also be considered. Both components re-affirm the same GvHD specific signature reported by Brinkman et al. (2007).

## 5 Concluding Remarks

We studied multivariate mixtures which 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 and skew- $t$  distributions where the univariate random effect is substituted by a higher dimensional one, see e.g. Branco and Dey (2001) and 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 e.g. 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.

## A Choosing the Prior

To perform Bayesian inference for finite mixtures of skew normal or skew- $t$  mixtures, a prior has to be chosen for the weight distribution  $\boldsymbol{\eta}$  and for all component specific parameters. Concerning the weight distribution, we apply the commonly used Dirichlet distribution  $\boldsymbol{\eta} \sim \mathcal{D}(e_0, \dots, e_0)$  with  $e_0 = 4$ .

Using the representations discussed in Subsection 3.1, conditionally conjugate priors are available for all transformed component specific parameters except the degrees of freedom parameter. Representation (15), for instance, suggests following prior for  $\boldsymbol{\theta}_k^* = (\xi_k, \psi_k, \sigma_k^2)$  for skew normal mixtures:

$$(\xi_k \ \psi_k)' | \sigma_k^2 \sim \mathcal{N}_2(\mathbf{b}_0, \mathbf{B}_0 \sigma_k^2), \quad \sigma_k^2 \sim \mathcal{G}^{-1}(c_0, C_0), \quad (29)$$

with  $\mathbf{b}_0 = (b_0^\xi \ b_0^\psi)' \in \mathfrak{R}^2$  and  $\mathbf{B}_0 = \text{Diag}(D^\xi, D^\psi) \in \mathfrak{R}^{2 \times 2}$ . Similarly, for multivariate skew normal mixtures representation (17) suggests following prior for  $\boldsymbol{\theta}_k^* = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$ :

$$(\boldsymbol{\xi}_k' \ \boldsymbol{\psi}_k')' | \boldsymbol{\Sigma}_k \sim \mathcal{N}_{2r} \left( \left( \begin{array}{cc} (\mathbf{b}_0^\xi)' & (\mathbf{b}_0^\psi)' \end{array} \right)', \mathbf{B}_0 \otimes \boldsymbol{\Sigma}_k \right), \quad \boldsymbol{\Sigma}_k \sim \mathcal{W}_r^{-1}(c_0, \mathbf{C}_0), \quad (30)$$

where  $\mathbf{b}_0^\xi, \mathbf{b}_0^\psi \in \mathfrak{R}^r$  and  $\mathbf{B}_0 \otimes \boldsymbol{\Sigma}_k$  denotes the Kronecker product of  $\mathbf{B}_0$  and  $\boldsymbol{\Sigma}_k$ .

In both cases, we center the prior of the skewness parameter at the normal distribution by choosing  $b_0^\psi = 0$  or  $\mathbf{b}_0^\psi = \mathbf{0}_{r \times 1}$ . We center the prior of  $\xi_k$  and  $\boldsymbol{\xi}_k$  either at the mean of the data, i.e.  $b_0^\xi = \bar{y}$  or  $\mathbf{b}_0^\xi = \bar{\mathbf{y}}$ , or choose  $b_0^\xi = 0$  or  $\mathbf{b}_0^\xi = \mathbf{0}_{r \times 1}$ . The hyperparameters  $D^\xi$  and  $D^\psi$  control the prior information in  $\xi_k$  or  $\boldsymbol{\xi}_k$  and  $\psi_k$  or  $\boldsymbol{\psi}_k$  and are selected as small positive numbers, e.g.  $D^\xi = D^\psi = 0.1$ .

We choose  $c_0 = 2.5$  to bound  $\sigma_k^2$  away from zero, while for  $r > 1$   $c_0 = 2.5 + (r - 1)/2$  to bound the eigenvalues of  $\boldsymbol{\Sigma}_k$  away from zero. We choose  $C_0 = \phi s_y^2$  or  $\mathbf{C}_0 = \phi \mathbf{S}_y$ , where  $s_y^2$  and  $\mathbf{S}_y$  are, respectively, the sample variance and the sample covariance matrix of the data.  $\phi$  influences the prior expectation of the amount of heterogeneity explained by differences in the group means, see e.g. Frühwirth-Schnatter (2006, Section 6.3.2). Choosing  $\phi = 0.5$  corresponds to a prior expectation of 2/3 explained heterogeneity.

Among these hyperparameters, we found  $C_0$  and  $\mathbf{C}_0$  to be rather influential. For this

reasons, we combine prior (29) or (30) with a hierarchical prior,

$$C_0 \sim \mathcal{G}(g_0, G_0), \quad \mathbf{C}_0 \sim \mathcal{W}_r(g_0, \mathbf{G}_0), \quad (31)$$

where we select  $g_0 = 0.5 + (r - 1)/2$ ,  $\mathbf{G}_0 = g_0(\phi \mathbf{S}_y)^{-1}$  for  $r > 1$ , and  $G_0 = g_0/(\phi s_y^2)$ . Such hierarchical priors have been used by Richardson and Green (1997, Subsection 2.4) for normal mixtures and by Stephens (1997) for Student- $t$ - mixtures to reduce sensitivity with respect to choosing the prior of component specific scale parameters.

Finally, for skew- $t$  mixtures we assume that the degrees of freedom parameters  $\nu_1, \dots, \nu_K$  are a priori independent of the remaining parameters and  $p(\xi_k, \psi_k, \sigma_k^2)$  and  $p(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$  are selected as in (29) and (30). The prior on  $\nu_k$  has to be selected carefully in order to avoid improper posteriors, see e.g. Fonseca, Ferreira, and Migon (2008). We assume prior independence of  $\nu_1, \dots, \nu_K$  with

$$p(\nu_k) \propto \frac{(\nu_k - 1)}{(\nu_k - 1 + d)^3} I_{[1, \infty)}(\nu_k). \quad (32)$$

This prior was introduced in Juárez and Steel (2009) for Student- $t$  mixtures with  $\nu_1 = \dots = \nu_K$ . The median of this prior is equal to  $1 + d(1 + \sqrt{2})$ . We shift the prior away from 0, as it is advisable to avoid values for  $\nu_k$  that are close to 0, see Fernández and Steel (1999).

## B Details on MCMC Estimation

We provide details only for multivariate mixtures, univariate ones results for  $r = 1$ .

### B.1 The Conditional Posterior of the Truncated Normal Random Effects

Consider following random effects model with  $r \geq 1$  repeated measurements and truncated normal random effects:

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

where the parameters  $\boldsymbol{\xi}$ ,  $\boldsymbol{\psi}$ , and  $\boldsymbol{\Sigma}$  are known. The full conditional posterior density  $p(z_i|\mathbf{y}_i)$  of  $z_i$  given observation  $\mathbf{y}_i$  is given by:

$$\begin{aligned} p(z_i|\mathbf{y}_i) &\propto p(\mathbf{y}_i|z_i)p(z_i) \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\xi} - \boldsymbol{\psi}z_i)' \boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\xi} - \boldsymbol{\psi}z_i)\right\} \exp\left\{-\frac{z_i^2}{2}\right\} I_{\{z_i>0\}} \\ &\propto \exp\left\{-\frac{1}{2}\left(z_i^2(\boldsymbol{\psi}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi} + 1) - 2z_i\boldsymbol{\psi}'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\xi})\right)\right\} I_{\{z_i>0\}}. \end{aligned}$$

Completing squares yields:

$$\begin{aligned} z_i|\mathbf{y}_i &\sim \mathcal{TN}_{[0,\infty)}(a_i, A), \\ a_i &= A\boldsymbol{\psi}'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\xi}), \quad A = (1 + \boldsymbol{\psi}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi})^{-1}. \end{aligned} \tag{33}$$

## B.2 Mixtures of Skew Normal Distributions

MCMC estimation for skew normal mixtures is possible through a two-step Gibbs sampler if the hyperparameters  $C_0$  and  $\mathbf{C}_0$  are fixed:

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

A starting value for  $\mathbf{S}$  is determined using  $K$ -means clustering of  $\mathbf{y}_1, \dots, \mathbf{y}_N$ , while  $z_i = 0$  for  $i = 1, \dots, N$ .

**Step (a).** Let  $N_k = \#\{S_i = k\}$  be equal to the number of observations in group  $k$ . Sample the weights  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$  from a  $\mathcal{D}(e_1, \dots, e_K)$ -distribution, where  $e_k = e_0 + N_k$ ,  $k = 1, \dots, K$ .

Partition the observations  $\mathbf{y}_i$  and the regressors  $\mathbf{x}_i = (1 \ z_i)$ , for  $i = 1, \dots, N$  according to the indicators  $\mathbf{S}$  into  $K$  groups. For each  $k = 1, \dots, K$ , construct a regressor matrix  $\mathbf{X}_k \in \mathfrak{R}^{N_k \times 2}$  where the  $N_k$  rows are equal to all regressors  $\mathbf{x}_i$  where  $S_i = k$ . Similarly, construct an observation matrix  $\mathbf{y}_k \in \mathfrak{R}^{r \times N_k}$  where the  $N_k$  columns are equals to all observations  $\mathbf{y}_i$  where  $S_i = k$ . Sample  $\boldsymbol{\theta}_k^* = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$  from the conditional posterior  $p(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k|\mathbf{z}, \mathbf{S}, \mathbf{y}) = p(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k|\boldsymbol{\Sigma}_k, \mathbf{z}, \mathbf{S}, \mathbf{y})p(\boldsymbol{\Sigma}_k|\mathbf{z}, \mathbf{S}, \mathbf{y})$  which is obtained from combining the regression model (17)

with prior (30):

$$\begin{aligned}
(\boldsymbol{\xi}'_k \boldsymbol{\psi}'_k)' | \boldsymbol{\Sigma}_k, \mathbf{z}, \mathbf{S}, \mathbf{y} &\sim \mathcal{N}_{2r}(\text{vec}(\mathbf{b}_k), \mathbf{B}_k \otimes \boldsymbol{\Sigma}_k), \\
\boldsymbol{\Sigma}_k | \mathbf{z}, \mathbf{S}, \mathbf{y} &\sim \mathcal{W}_r^{-1}(c_k, \mathbf{C}_k), \\
\mathbf{b}_k &= \begin{pmatrix} \mathbf{b}_k^\xi & \mathbf{b}_k^\psi \end{pmatrix} = \left( \mathbf{y}_k \mathbf{X}_k + \begin{pmatrix} \frac{1}{D^\xi} \mathbf{b}_0^\xi & \frac{1}{D^\psi} \mathbf{b}_0^\psi \end{pmatrix} \right) \mathbf{B}_k, \\
\mathbf{B}_k &= (\mathbf{X}'_k \mathbf{X}_k + \mathbf{B}_0^{-1})^{-1}, \\
c_k &= c_0 + \frac{N_k}{2}, \\
\mathbf{C}_k &= \mathbf{C}_0 + \frac{1}{2} \left( \sum_{i:S_i=k} \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}'_i + \frac{1}{D^\xi} (\mathbf{b}_k^\xi - \mathbf{b}_0^\xi) (\mathbf{b}_k^\xi - \mathbf{b}_0^\xi)' + \frac{1}{D^\psi} (\mathbf{b}_k^\psi - \mathbf{b}_0^\psi) (\mathbf{b}_k^\psi - \mathbf{b}_0^\psi)' \right), \\
\boldsymbol{\varepsilon}_i &= \mathbf{y}_i - \mathbf{b}_k^\xi - z_i \mathbf{b}_k^\psi.
\end{aligned} \tag{34}$$

The symbol  $\text{vec}(\cdot)$  refers to the vector obtained by stacking all column of the matrix appearing as argument.

**Step(b).** Sample  $S_i$  independently for each  $i = 1, \dots, N$  from  $p(S_i | \mathbf{y}_i, \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \boldsymbol{\eta})$  which is equal to following discrete distribution:

$$p(S_i = k | \boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \eta_k, \mathbf{y}_i) \propto f_{SN}(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k) \eta_k. \tag{35}$$

$f_{SN}(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k)$  is the density of a multivariate skew normal distribution defined in (5) and  $(\boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k)$  are determined from  $(\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$  as in (18). Sample  $z_i$  independently for  $i = 1, \dots, N$  from  $p(z_i | S_i, \boldsymbol{\theta}_k^*, \mathbf{y}_i)$  using the truncated normal posterior (33) corresponding to the random effects model (17):

$$\begin{aligned}
z_i | S_i = k, \mathbf{y}_i, \boldsymbol{\theta}_k^* &\sim \mathcal{TN}_{[0, \infty)}(a_{i,k}, A_k), \\
a_{i,k} &= A_k \boldsymbol{\psi}'_k \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\xi}_k), \quad A_k = (1 + \boldsymbol{\psi}'_k \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\psi}_k)^{-1}.
\end{aligned} \tag{36}$$

**Hierarchical priors.** For the hierarchical prior (31) a further step has to be added to sample  $C_0$  or  $\mathbf{C}_0$  from following densities, where  $g_N = g_0 + Kc_0$ :

$$C_0 \sim \mathcal{G} \left( g_N, G_0 + \sum_{k=1}^K \frac{1}{\sigma_k^2} \right), \quad \mathbf{C}_0 \sim \mathcal{W}_r \left( g_N, \mathbf{G}_0 + \sum_{k=1}^K \boldsymbol{\Sigma}_k^{-1} \right).$$

### B.3 Mixtures of Skew- $t$ Distributions

MCMC estimation for skew- $t$  mixtures is possible through following three-step sampler if the hyperparameters  $C_0$  and  $\mathbf{C}_0$  are fixed:

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

MCMC estimation is started with  $w_i = 1, i = 1, \dots, N$ , and  $\nu_1 = \dots = \nu_K = 10$ , while starting value for  $\mathbf{S}$  and  $\mathbf{z}$  are selected as in Appendix B.2. For the hierarchical prior (31) a further step has to be added as described at the end of Appendix B.2.

**Step(a).** Sample  $\boldsymbol{\eta}$  as in Appendix B.2, Step(a). Sample  $(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$  for  $k = 1, \dots, K$ , from  $p(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k | \mathbf{z}, \mathbf{S}, \mathbf{w}, \mathbf{y}) = p(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k | \boldsymbol{\Sigma}_k, \mathbf{z}, \mathbf{S}, \mathbf{w}, \mathbf{y}) p(\boldsymbol{\Sigma}_k | \mathbf{z}, \mathbf{S}, \mathbf{w}, \mathbf{y})$  where:

$$\begin{aligned}
 (\boldsymbol{\xi}'_k \boldsymbol{\psi}'_k)' | \boldsymbol{\Sigma}_k, \mathbf{z}, \mathbf{S}, \mathbf{w}, \mathbf{y} &\sim \mathcal{N}_{2r}(\text{vec}(\mathbf{b}_k), \mathbf{B}_k \otimes \boldsymbol{\Sigma}_k), \\
 \boldsymbol{\Sigma}_k | \mathbf{z}, \mathbf{S}, \mathbf{w}, \mathbf{y} &\sim \mathcal{W}_r^{-1}(c_k, \mathbf{C}_k), \\
 \mathbf{b}_k &= \begin{pmatrix} \mathbf{b}_k^\xi & \mathbf{b}_k^\psi \end{pmatrix} = \left( \mathbf{y}_k^w \mathbf{X}_k^w + \begin{pmatrix} \frac{1}{D^\xi} \mathbf{b}_0^\xi & \frac{1}{D^\psi} \mathbf{b}_0^\psi \end{pmatrix} \right) \mathbf{B}_k, \\
 \mathbf{B}_k &= ((\mathbf{X}_k^w)' \mathbf{X}_k^w + \mathbf{B}_0^{-1})^{-1}, \\
 \mathbf{C}_k &= \mathbf{C}_0 + \frac{1}{2} \left( \sum_{i:S_i=k} w_i \boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' + \frac{1}{D^\xi} (\mathbf{b}_k^\xi - \mathbf{b}_0^\xi)(\mathbf{b}_k^\xi - \mathbf{b}_0^\xi)' + \frac{1}{D^\psi} (\mathbf{b}_k^\psi - \mathbf{b}_0^\psi)(\mathbf{b}_k^\psi - \mathbf{b}_0^\psi)' \right).
 \end{aligned} \tag{37}$$

$c_k$  and  $\boldsymbol{\varepsilon}_i$  are the same as in (34). The  $N_k$  rows of the matrix  $\mathbf{X}_k^w \in \mathfrak{R}^{N_k \times 2}$  are equal to all rescaled regressors  $\mathbf{x}_i = (\sqrt{w_i} \ \sqrt{w_i} z_i)$  where  $S_i = k$ . Similarly, the  $N_k$  columns of the matrix  $\mathbf{y}_k^w \in \mathfrak{R}^{r \times N_k}$  are equal to all rescaled observations  $\sqrt{w_i} \mathbf{y}_i$  where  $S_i = k$ .

**Step(b).** Sample  $S_i$  independently for each  $i = 1, \dots, N$  from  $p(S_i | \mathbf{y}_i, \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \boldsymbol{\eta})$  which is equal to following discrete distribution:

$$p(S_i = k | \boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \nu_k, \eta_k, \mathbf{y}_i) \propto f_{\mathcal{ST}}(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k, \nu_k) \eta_k.$$

$f_{\mathcal{ST}}(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k, \nu_k)$  is the density of the multivariate skew- $t$  distribution defined in (12) and  $(\boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k)$  are determined from  $(\boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$  as in (18). Sample  $z_i$  independently for  $i = 1, \dots, N$  from  $p(z_i | S_i, w_i, \boldsymbol{\theta}_k^*, \mathbf{y}_i)$  using the truncated normal posterior (33) corresponding to



the random effects model (22):

$$z_i | S_i = k, w_i, \mathbf{y}_i, \boldsymbol{\theta}_k^* \sim \mathcal{TN}_{[0, \infty)}(a_{i,k}, A_k/w_i),$$

where  $a_{i,k}$  and  $A_k$  are the same as in (36).

**Step(c).** Depending on the degree of data augmentation in the conditional density  $p(\nu_1, \dots, \nu_K | \cdot)$ , different Metropolis-Hastings steps to sample  $\nu_k$  result. The fastest algorithm is sampling  $\nu_k$  from  $p(\nu_k | \boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \mathbf{S}, \mathbf{w}, \mathbf{y})$  as Lin et al. (2007) did for Student- $t$  mixtures. However, we found that this works only, if the degree of freedom is small in all components. We observed tremendous inefficiency factors if some of the  $\nu_k$ s were larger than about 10. Sampling  $\nu_k$  from  $p(\nu_k | \boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \mathbf{S}, \mathbf{y})$  where  $\mathbf{w}$  is integrated out increases efficiency considerably. We gained additional efficiency by sampling  $\nu_k$  without conditioning on  $\mathbf{S}$  and  $\mathbf{w}$  from  $p(\nu_k | \boldsymbol{\theta}_{-k}^*, \boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\eta}, \mathbf{y})$  where  $\boldsymbol{\theta}_{-k}^*$  denotes all component specific parameters except  $\boldsymbol{\theta}_k^*$ . However, this sampler is the most time consuming one because it involves the computation of the observed-data likelihood function  $p(\mathbf{y} | \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \boldsymbol{\eta})$ .

To sample  $\nu_k$  for  $k = 1, \dots, K$  from  $p(\nu_k | \boldsymbol{\theta}_{-k}^*, \boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\eta}, \mathbf{y})$  we use a Metropolis-Hastings algorithm with a uniform log random walk proposal

$$\log(\nu_k^{new} - 1) \sim \mathcal{U}[\log(\nu_k - 1) - c_{\nu_k}, \log(\nu_k - 1) + c_{\nu_k}]$$

with fixed width parameter  $c_{\nu_k}$ . Accept  $\nu_k^{new}$  with probability

$$\min\left(1, \frac{p(\mathbf{y} | \boldsymbol{\theta}_{-k}^*, (\boldsymbol{\theta}_k^*)^{new}, \boldsymbol{\eta}) p(\nu_k^{new})(\nu_k^{new} - 1)}{p(\mathbf{y} | \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \boldsymbol{\eta}) p(\nu_k)(\nu_k - 1)}\right),$$

where  $(\boldsymbol{\theta}_k^*)^{new} = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \nu_k^{new})$ . The computation of the acceptance rate involves the computation of the observed-data likelihood function

$$p(\mathbf{y} | \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \boldsymbol{\eta}) = \prod_{i=1}^N \left( \sum_{k=1}^K \eta_k f_{\mathcal{ST}}(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k, \nu_k) \right).$$

Note that  $p(\mathbf{y} | \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \boldsymbol{\eta})$  may be computed efficiently by observing that the contribution of only one component density changes.

Finally, sample  $w_i$  independently for  $i = 1, \dots, N$  from  $p(w_i | \boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_K^*, \mathbf{z}, \mathbf{S}, \mathbf{y})$ . To derive this posterior observe that  $w_i$  appears both in the observation equation (22) as well as in the prior distribution of the random effect  $z_i$  given in (20):

$$\begin{aligned} p(w_i | \mathbf{y}_i, S_i = k, z_i, \boldsymbol{\theta}_k^*) &\propto p(\mathbf{y}_i | \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\psi}_k, w_i) p(z_i | w_i, \nu_k) p(w_i | \nu_k) \\ &\propto |w_i \boldsymbol{\Sigma}_k^{-1}|^{1/2} \exp\left(-\frac{w_i}{2} \boldsymbol{\varepsilon}_i' \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\varepsilon}_i\right) w_i^{1/2} \exp\left(-\frac{w_i z_i^2}{2}\right) w_i^{\nu_k/2-1} \exp\left(-\frac{w_i \nu_k}{2}\right), \end{aligned}$$

where  $\boldsymbol{\varepsilon}_i$  is the same as in (34). This is the kernel of following Gamma distribution:

$$w_i | \mathbf{y}_i, z_i, S_i = k, \boldsymbol{\theta}_k^* \sim \mathcal{G}\left(\frac{\nu_k + r + 1}{2}, \frac{\nu_k + z_i^2 + \text{tr}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \boldsymbol{\Sigma}_k^{-1})}{2}\right).$$

## B.4 Label Switching and Post-processing MCMC

To make sure that we explore all labelling subspaces we add a random permutation step as in Frühwirth-Schnatter (2001) to the MCMC scheme introduced in the previous subsection and perform post-processing of the MCMC output to handle label switching.

Following Celeux (1998), we use standard  $k$ -means clustering in the point process representation of the MCMC draws to identify the finite mixture model. For univariate skew normal and skew- $t$  mixtures we apply  $k$ -means clustering to  $(\xi_k, \alpha_k, \omega_k)$ . For multivariate mixtures  $k$ -means clustering is applied to the component means  $\boldsymbol{\mu}_k$  defined in (8) and (13), respectively.

The whole method is based on the idea that MCMC draws belonging to the same component will cluster around the same point in the point process representation of the underlying “true” mixture model (Stephens, 2000a). In cases where the simulation clusters are well-separated all classification sequences are a permutation of  $\{1, \dots, K\}$  and indicate how to rearrange the component specific parameters in order to obtain a unique labelling. This method not only allows to identify the component specific parameters, but also identifies a unique labelling of the allocations, see Frühwirth-Schnatter (2006, p. 96f).

## References

Arellano-Valle, R. B. and A. Azzalini (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 A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew  $t$ -distribution. *Journal of the Royal Statistical Society, Ser. B* 65, 367–389.
- Azzalini, A. and A. Dalla Valle (1996). The multivariate skew normal distribution. *Biometrika* 83, 715–726.
- Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49, 803–821.
- Bennett, D. A., J. A. Schneider, A. S. Buchman, C. M. de Leon, J. L. Bienias, and R. S. Wilson (2005). The Rush Memory and Aging Project: Study Design and Baseline Characteristics of the Study Cohort. *Neuroepidemiology* 25, 163–175.
- Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated classification likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 719–725.
- Biernacki, C. and G. Govaert (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics* 29, 451–457.
- Boedigheimer, M. J. and J. Ferbas (2008). Mixture modeling approach to flow cytometry data. *Cytometry Part A* 73A, 421–429.
- Branco, M. D. and D. K. Dey (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis* 79, 99–113.
- Brinkman, R. R., M. Gasparetto, S.-J. J. Lee, A. Ribickas, J. Perkins, W. Janssen, R. Smiley, and C. Smith (2007). Transplant. High content flow cytometry and temporal data analysis for defining a cellular signature of Graft versus Host Disease. *Biol Blood Marrow* 13, 691–700.
- Cabral, C. R. B., H. Bolfarine, and J. R. G. Pereira (2008). Bayesian density estimation using skew student- $t$ -normal mixtures. *Computational Statistics and Data Analysis* 52, 5075–5090.
- Celeux, G. (1998). Bayesian inference for mixture: The label switching problem. In P. J. Green and R. Rayne (Eds.), *COMPSTAT 98*, pp. 227–232. Heidelberg: Physica.

- Celeux, G., F. Forbes, C. P. Robert, and D. M. Titterton (2006). Deviance information criteria for missing data models. *Bayesian Analysis* 1, 651–674.
- Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* 95, 957–970.
- Chan, C., F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler (2008). Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry Part A* 73A, 693–701.
- Dellaportas, P. and I. Papageorgiou (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing* 16, 57–68.
- Diebolt, J. and C. P. Robert (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Ser. B* 56, 363–375.
- Fernández, C. and M. F. J. Steel (1999). Multivariate student- $t$  regression models: Pitfalls and inference. *Biometrika* 86, 153–167.
- Fonseca, T. C. O., M. A. R. Ferreira, and H. S. Migon (2008). Objective Bayesian analysis for the Student- $t$  regression model. *Biometrika* 95, 325–333.
- 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. (Ed.) (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., C. C. Holmes, and D. A. Stephens (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science* 20, 50–67.
- Jasra, A., D. A. Stephens, K. Gallagher, and C. C. Holmes (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 “On Bayesian analysis of mixtures with an unknown number of components.”. *Journal of the Royal Statistical Society, Ser. B* 59, 778–779.
- Juárez, M. A. and M. F. J. Steel (2009). Model-based clustering of non-Gaussian panel data based on skew- $t$  distributions. *Journal of Business & Economic Statistics* 27, to appear.
- Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya A* 62, 49–66.
- Kruglyak, L. and E. S. Lander (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* 139, 1421–1428.
- Liang, F. and W. H. Wong (2001). Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. *Journal of the American Statistical Association* 96, 653–666.
- Lin, T. I. (2009a). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis* 100, 257–265.
- Lin, T. I. (2009b). Robust mixture modeling using multivariate skew  $t$  distributions. *Statistics and Computing* XX, to appear.
- Lin, T. I., J. C. Lee, and W. J. Hsieh (2007). Robust mixture modeling using the skew  $t$ -distribution. *Statistics and Computing* 17, 81–92.
- Lin, T. I., J. C. Lee, and H. F. Ni (2004). Bayesian analysis of mixture modelling using the multivariate  $t$ -distribution. *Statistics and Computing* 14, 119–130.
- Lin, T. I., J. C. Lee, and S. Y. Yen (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica* 17, 909–927.
- Lo, K., R. R. Brinkman, and R. Gottardo (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry Part A* 73A, 321–332.
- McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. Wiley Series in Probability and Statistics. New York: Wiley.
- Meng, X.-L. and W. H. Wong (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 G. McLachlan (2000). Robust mixture modelling using the  $t$  distribution. *Statistics and Computing* 10, 339–348.
- Perfetto, S. P., P. K. Chattopadhyay, and M. Roederer (2004). Seventeen-colour flow cytometry: unravelling the immune system. *Nature Reviews Immunol.* 4, 648–55.
- Pyne, S., X. Hu, K. Wang, E. Rossin, T. Lin, L. M. Maier, C. Baecher-Allan, G. J. McLachlan, P. Tamayo, D. A. Hafler, P. L. De Jager, and J. P. Mesirov (2009). Automated high-dimensional flow cytometric data analysis. *Proc.Natl.Acad.Sci.USA* 106, 8519–8524.
- Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Ser. 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., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Ser. B* 64, 583–639.
- Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions*. Ph. D. thesis, University of Oxford.
- Stephens, M. (2000a). Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods. *The Annals of Statistics* 28, 40–74.
- Stephens, M. (2000b). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Ser. B* 62, 795–809.
- van Dyk, D. and X.-L. Meng (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* 10, 1–50.
- Wilson, R. S., J. L. Bienias, D.A. Evans, and D. A. Bennett (2004). The Religious Orders Study: Overview and Change in Cognitive and Motor Speed. *Aging, Neuropsychol, Cogn.* 11, 280–303.
- Wilson, R. S., J. A. Schneider, L. L. Barnes, L. A. Beckett, N. T. Aggarwal, E. J. Cochran, E. Berry-Kravis, J. Bach, J. H. Fox, D. A. Evans, and D. A. Bennett (2002). The Apolipoprotein E e4 allele and decline in different cognitive systems during a 6-year period. *Arch. Neurol.* 59, 1154–1160.