Supplementary Material to

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

Sylvia Frühwirth-Schnatter[a] and Saumyadipta Pyne[b]

[a]*Department of Applied Statistics and Econometrics, Johannes Kepler Universität Linz, Linz, Austria,* sylvia.fruehwirth-schnatter@jku.at

[b]*Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA,* saumyadipta_pyne@dfci.harvard.edu

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 $\boldsymbol{\eta} \sim \mathcal{D}\left(e_0, \ldots, e_0\right)$ 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 $\boldsymbol{\theta}_k^\star = (\xi_k, \psi_k, \sigma_k^2)$ for skew normal mixtures:

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

with $\mathbf{b}_0 = \left(\begin{array}{cc} b_0^\xi & b_0^\psi \end{array} \right)' \in \Re^2$ and $\mathbf{B}_0 = \mathrm{Diag}\big(D^\xi, D^\psi\big) \in \Re^{2\times 2}$. Similarly, for multivariate skew normal mixtures representation (17) suggests following prior for $\boldsymbol{\theta}_k^\star = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$:

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

where $\mathbf{b}_0^\xi, \mathbf{b}_0^\psi \in \Re^r$ 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 ξ_k and $\boldsymbol{\xi}_k$ either at the mean of the data, i.e. $b_0^\xi = \overline{y}$ or $\mathbf{b}_0^\xi = \overline{\mathbf{y}}$, or choose $b_0^\xi = 0$ or $\mathbf{b}_0^\xi = \mathbf{0}_{r\times 1}$. The hyperparameters D^ξ and D^ψ control the prior information in ξ_k or $\boldsymbol{\xi}_k$ and ψ_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 σ_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. ϕ 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 hierarchial prior,

$$C_0 \sim \mathcal{G}\left(g_0, G_0\right), \qquad \mathbf{C}_0 \sim \mathcal{W}_r\left(g_0, \mathbf{G}_0\right), \tag{31}$$

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 $\nu_1, \ldots, \nu_K$ are apriori 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 ν_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 $\nu_1, \ldots, \nu_K$ with

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

This prior was introduced in Juárez and Steel (2009) for Student-t mixtures with $\nu_1 = \ldots = \nu_K$. 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 ν_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 Posterior of the Truncated Normal Random Effects

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

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

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:

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

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 $\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_K^\star$ and $\boldsymbol{\eta}$ conditional on $\boldsymbol{z}$, $\mathbf{S}$ and $\mathbf{y}$.

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

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

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

Partition the observations $\mathbf{y}_i$ and the regressors $\mathbf{x}_i = (1\ z_i)$, for $i = 1, \ldots, N$ according to the indicators $\mathbf{S}$ into K groups. For each $k = 1, \ldots, K$, construct a regressor matrix $\mathbf{X}_k \in \Re^{N_k \times 2}$ where the N_k rows are equal to all regressors $\mathbf{x}_i$ where $S_i = k$. Similarly, construct an observation matrix $\mathbf{y}_k \in \Re^{r \times N_k}$ where the N_k columns are equals to all observations $\mathbf{y}_i$ where $S_i = k$. Sample $\boldsymbol{\theta}_k^\star = (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k)$ from the conditional posterior $p(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k | \boldsymbol{z}, \mathbf{S}, \mathbf{y}) = p(\boldsymbol{\xi}_k, \boldsymbol{\psi}_k | \boldsymbol{\Sigma}_k, \boldsymbol{z}, \mathbf{S}, \mathbf{y}) p(\boldsymbol{\Sigma}_k | \boldsymbol{z}, \mathbf{S}, \mathbf{y})$ which is obtained from combining the regression model (17) with prior (30):

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

The symbol 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 $i = 1, \ldots, N$ from $p(S_i | \mathbf{y}_i, \boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_K^\star, \boldsymbol{\eta})$ which is equal to following discrete distribution:

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

$f_{\mathcal{SN}}\left(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k\right)$ 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 $i = 1, \ldots, N$ from $p(z_i | S_i, \boldsymbol{\theta}_k^\star, \mathbf{y}_i)$ using the truncated normal posterior (33) corresponding to the random effects model (17):

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

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 $g_N = g_0 + Kc_0$:

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

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 $\boldsymbol{\theta}_1^\star, \dots, \boldsymbol{\theta}_K^\star$ (except $\nu_1, \dots, \nu_K$) and $\boldsymbol{\eta}$ conditional on $\boldsymbol{z}$, $\mathbf{S}$, $\boldsymbol{w}$ and $\mathbf{y}$.

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

(c) Sample $\nu_1, \dots, \nu_K$ and $\boldsymbol{w}$ conditional on $\mathbf{y}$ and the remaining parameters.

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

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

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

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

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

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

$f_{\mathcal{ST}}\left(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k, \nu_k\right)$ 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 $i = 1, \ldots, N$ from $p(z_i|S_i, w_i, \boldsymbol{\theta}_k^\star, \mathbf{y}_i)$ using the truncated normal posterior (33) corresponding to the random effects model (22):

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

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 $p(\nu_1, \ldots, \nu_K|\cdot)$, different Metropolis-Hastings steps to sample ν_k result. The fastest algorithm is sampling ν_k from $p(\nu_k|\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \mathbf{S}, \boldsymbol{w}, \mathbf{y})$ as Lin, Lee, and Hsieh (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 ν_ks were larger than about 10. Sampling ν_k from $p(\nu_k|\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \mathbf{S}, \mathbf{y})$ where $\boldsymbol{w}$ is integrated out increases efficiency considerably. We gained additional efficiency by sampling ν_k without conditioning on $\mathbf{S}$ and $\boldsymbol{w}$ from $p(\nu_k|\boldsymbol{\theta}_{-k}^\star, \boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\eta}, \mathbf{y})$ where $\boldsymbol{\theta}_{-k}^\star$ denotes all component specific parameters expect $\boldsymbol{\theta}_k^\star$. However, this sampler is the most time consuming one because it involves the computation of the observed-data likelihood function $p(\mathbf{y}|\boldsymbol{\theta}_1^\star, \ldots, \boldsymbol{\theta}_K^\star, \boldsymbol{\eta})$.

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

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

with fixed width parameter c_{ν_k}. Accept ν_k^{new} with probability

$$\min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^\star_{-k}, (\boldsymbol{\theta}^\star_k)^{new}, \boldsymbol{\eta})p(\nu_k^{new})(\nu_k^{new}-1)}{p(\mathbf{y}|\boldsymbol{\theta}^\star_1, \ldots, \boldsymbol{\theta}^\star_K, \boldsymbol{\eta})p(\nu_k)(\nu_k-1)}\right),$$

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

$$p(\mathbf{y}|\boldsymbol{\theta}^\star_1, \ldots, \boldsymbol{\theta}^\star_K, \boldsymbol{\eta}) = \prod_{i=1}^N \left(\sum_{k=1}^K \eta_k f_{\mathcal{ST}}\left(\mathbf{y}_i; \boldsymbol{\xi}_k, \boldsymbol{\Omega}_k, \boldsymbol{\alpha}_k, \nu_k\right)\right).$$

Note that $p(\mathbf{y}|\boldsymbol{\theta}^\star_1, \ldots, \boldsymbol{\theta}^\star_K, \boldsymbol{\eta})$ may be computed efficiently by observing that the contribution of only one component density changes.

Finally, sample w_i independently for $i = 1, \ldots, N$ from $p(w_i|\boldsymbol{\theta}^\star_1, \ldots, \boldsymbol{\theta}^\star_K, \boldsymbol{z}, \mathbf{S}, \mathbf{y})$. 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):

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

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

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

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, 2000). In cases where the simulation clusters are well-separated all classification sequences are a permutation of $\{1, \ldots, K\}$ 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

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.

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. (2006). *Finite Mixture and Markov Switching Models.* New York: Springer.

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.

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.

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.

Stephens, M. (1997). *Bayesian Methods for Mixtures of Normal Distributions.* Ph. D. thesis, University of Oxford.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods. *The Annals of Statistics 28*, 40–74.