# Supplementary materials for Location and scale mixtures of Gaussians with flexible tail behaviour: properties, inference and application to multivariate clustering

by Wraith, D. & Forbes, F.

## A: Multiple Scaled Generalised Hyperbolic distributions (MSGH)

Standard Generalised Hyperbolic distributions can be seen as location and scale mixtures where the weight variable follows a Generalized Inverse Gaussian (GIG) distribution. The GIG distribution depends on three parameters and is given by

$$\begin{aligned} f_W(w;\lambda,\gamma,\delta) &= \mathcal{GIG}(w;\lambda,\gamma,\delta) \\ &= \left(\frac{\gamma}{\delta}\right)^{\lambda} \frac{w^{\lambda-1}}{2K_\lambda(\delta\gamma)} \exp(-\frac{1}{2}(\delta^2/w+\gamma^2 w)) , \end{aligned} \tag{36}$$

where $K_r(x)$ is the modified Bessel function of the third kind of order $r$ evaluated at $x$ [1]. Depending on the parameter choice for the GIG, special cases of the GH family include: the multivariate GH distribution with hyperbolic margins ($\lambda = 1$) [Schmidt et al., 2006]; the Normal Inverse Gaussian ($\lambda = -1/2$) distribution [Barndorff-Nielsen et al., 1982]; the multivariate hyperbolic ($\lambda = \frac{M+1}{2}$) distribution [Barndorff-Nielsen, 1977]; the hyperboloid ($\lambda = 0$) distribution [Jensen, 1981]; the hyperbolic skew-t ($\lambda = -\nu, \gamma = 0$) distribution [Aas et al., 2005]; and the Normal Gamma ($\lambda > 0, \boldsymbol{\mu} = 0, \delta = 0$) distribution [Griffin and Brown, 2010] amongst others.

The standard location and scale representation (1) is generalised into a multiple scale version

$$\begin{aligned} p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A},\boldsymbol{\beta},\boldsymbol{\theta}) = &\int_0^\infty \ldots \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu}+\boldsymbol{D}\boldsymbol{\Delta}_{\mathrm{w}}\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathrm{w}}\boldsymbol{A}\boldsymbol{D}^T) \\ &\times f_{\mathbf{w}}(w_1\ldots w_M;\boldsymbol{\theta})\, \mathrm{d}w_1 \ldots dw_M \ , \end{aligned} \tag{37}$$

where $\boldsymbol{D}$ is the matrix of eigenvectors of the scale matrix $\boldsymbol{\Sigma}$, $\boldsymbol{A}$ is a diagonal matrix with the corresponding eigenvalues, $\boldsymbol{\Delta}_{\mathbf{w}} = \mathrm{diag}(w_1, \ldots w_M)$, and the weights are assumed to be independent *i.e.* $f_{\mathbf{w}}(w_1\ldots, w_M;\boldsymbol{\theta}) = f_{W_1}(w_1;\boldsymbol{\theta}_1)\ldots f_{W_M}(w_M;\boldsymbol{\theta}_M)$.

[1] The modified Bessel function (see Appendix in Jorgensen [1982]) is $K_r(x) = 1/2\int_0^\infty y^{r-1}\exp(-\frac{1}{2}x(y+y^{-1}))\, dy$

Equation (37) can be equivalently written as

$$\begin{aligned} p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D},\boldsymbol{A},\boldsymbol{\beta},\boldsymbol{\theta}) = \prod_{m=1}^{M}\int_0^{\infty} & \mathcal{N}_1([\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m; w_m A_m[\boldsymbol{D}^T\boldsymbol{\beta}]_m, w_m A_m) \\ & \times f_{W_m}(w_m)\, \mathrm{d}w_m \end{aligned} \tag{38}$$

where $[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m$ denotes the $m$th component of vector $\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})$ and $A_m$ the $m$th diagonal element of the diagonal matrix $\boldsymbol{A}$ (or equivalently the $m$th eigenvalue of $\boldsymbol{\Sigma}$).

To simulate from the MSGH distribution, it is possible to use eq. (12) in the manuscript or

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{A}^{1/2}[X_1\sqrt{W_1},\ldots,X_M\sqrt{W_M}]^T \tag{39}$$

where $\boldsymbol{X} \sim \mathcal{N}(0,\boldsymbol{I}_M)$ and $W_m \sim \mathcal{GIG}(\lambda_m,\gamma_m,\delta_m)$ (for $m = 1,\ldots,M$).

## B: Multiple Scaled Normal Inverse Gaussian distribution (MSNIG)

By setting $\lambda = -1/2$ in the GIG distribution we recover the Inverse Gaussian (IG) distribution,

$$\begin{aligned} f_W(w;\gamma,\delta) &= \mathcal{IG}(w;\gamma,\delta) \quad (40)\\ &= \frac{\delta}{w^{3/2}\sqrt{2\pi}}\exp(\delta\gamma)\,\exp(-\frac{1}{2}(\delta^2/w+\gamma^2 w))\ , \quad (41) \end{aligned}$$

which (when used as the mixing distribution) leads to the NIG distribution

$$\begin{aligned} p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta},\gamma,\delta) &= \mathcal{NIG}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta},\gamma,\delta) \\ &= \int_0^{\infty}\mathcal{N}_M(\mathbf{y};\boldsymbol{\mu}+w\boldsymbol{\Sigma}\boldsymbol{\beta},w\boldsymbol{\Sigma})\ \mathcal{IG}(w;\gamma,\delta)dw \\ &= \frac{\delta}{2^{\frac{M-1}{2}}}\exp(\delta\gamma+(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\beta})\left(\frac{\alpha}{\pi q(\mathbf{y})}\right)^{\frac{M+1}{2}} K_{\frac{M+1}{2}}(\alpha q(\mathbf{y})) \end{aligned}$$

where $\alpha$ and $q$ are defined as in definitions (4) and (3) in the manuscript.

Therefore, in the case of the MSNIG where $W_m \sim \mathcal{GIG}(\lambda_m = -1/2,\gamma_m,\delta_m) = \mathcal{IG}(\gamma_m,\delta_m)$, expressions (12) and (13) simplify into

$$\begin{aligned} E[\mathbf{Y}_{MSNIG}] &= \boldsymbol{\mu} + \boldsymbol{D}E[\Delta_W]\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta} \\ &= \boldsymbol{\mu} + \boldsymbol{D}\mathrm{diag}\left(\frac{\delta_1}{\gamma_1},\ldots,\frac{\delta_M}{\gamma_M}\right)\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta} \quad (42) \end{aligned}$$

$$\begin{aligned} Var[\mathbf{Y}_{MSNIG}] =& \boldsymbol{D}\mathrm{diag}\left(\frac{\delta_1}{\gamma_1},\ldots,\frac{\delta_M}{\gamma_M}\right)\boldsymbol{A}\boldsymbol{D}^T \quad (43)\\ & + \boldsymbol{D}\mathrm{diag}\left(\frac{\delta_1}{\gamma_1^3}[\boldsymbol{D}^T\boldsymbol{\beta}]_1^2,\ldots,\frac{\delta_M}{\gamma_M^3}[\boldsymbol{D}^T\boldsymbol{\beta}]_M^2\right)\boldsymbol{A}\boldsymbol{D}^T \\ =& \boldsymbol{D}\mathrm{diag}\left(\frac{\delta_m A_m}{\gamma_m}\right)\left(1+\frac{[\boldsymbol{D}^T\boldsymbol{\beta}]_m^2\boldsymbol{A}_m}{\gamma_m^2}\right)\boldsymbol{D}^T \quad (44) \end{aligned}$$

## C: Tail behaviour of the multiple scaled GH

The tail behaviour of the multiple scaled GH is similar to the GH with tails governed by a combined algebraic and exponential form. Let us assume that $\boldsymbol{D} = \boldsymbol{I}_M$ for simplification. Such a case can be easily recovered after a rotation. Then, it is straightforward to see (*e.g.* Aas and Hobaek Haff [2006]) that the density function of the MSGH distribution is equivalent to:

$$\mathcal{MSGH}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \sim const\ |y_m|^{\lambda_m - 1} \exp(y_m\beta_m - \alpha_m A_m^{-1/2}|y_m|)\ , \quad \text{as } |y_m| \to \infty \quad (45)$$

where $\alpha_m^2 = \gamma_m^2 + A_m\beta_m^2$ (Equation (8) of the manuscript).

Hence, the multiple scaled GH, like the GH distribution, is said to be *semi-heavy* tailed, which means that its tail behaviour is characterized by exponential instead of power decay. Alternative parameterisations of the GH permit the possibility of heavier tails [Aas and Hobaek Haff, 2006]. The parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ govern the tail behaviour of the density with smaller values of $\boldsymbol{\gamma}$ implying heavier tails, and larger values lighter tails. For our multiple scaled GH distributions, when all $\delta_m, \gamma_m$ tend to infinity with $\delta_m/\gamma_m$ tending to 1, the distribution tends to the multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Sigma\beta}, \boldsymbol{\Sigma})$. This is easily seen from the characteristic function (see Section D).

A difference between the tail behaviour of the GH and the multiple scaled GH can also be seen in measures of the tail dependency [Coles et al., 1999]. In applications, strong tail dependence is important for modelling the dependency/association of potentially extreme events (*e.g.* in finance, meteorology). In Figure 1 we compare the tail dependency of the Gaussian, *t*-distribution, standard GH and multiple scaled GH using a $\chi(q)$ plot [Coles et al., 1999] and simulated values from each distribution with $\boldsymbol{\mu} = [0, 0]^T, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ (equivalently $\boldsymbol{A} = diag(3/2, 1/2)$ and $\xi = \pi/4$) , $\boldsymbol{\beta} = [0, 0]^T, \boldsymbol{\gamma} = \boldsymbol{\delta} = [1, 1]^T$ (or $\nu = 1$) and $\boldsymbol{\lambda} = [-1/2, -1/2]^T$ (NIG). The function $\chi(q)$ can be interpreted as a quantile dependent measure of dependence with $\chi(q) = 0$ indicating independence and $\chi(q) = 1$ perfect dependence. Tail dependence is determined by the limit of $\chi(q)$ when $q$ tends to 1. In particular, the sign of $\chi(q)$ determines whether the variables are positively or negatively associated at quantile level $q$.

To compute $\chi(q)$, we used Coles et al. [1999] and the R package 'evd' [Team, 2011]. We assume that the data are *i.i.d.* random vectors with common bivariate distribution function $G$, and we define the random vector $[X, Y]^T$ to be distributed according to $G$.

The $\chi(q)$ plot is a plot of $q$ in (0,1) (interpreted as a quantile level) against empirical estimates of function

$$\chi(q) = 2 - \log(p(F_X(X) < q, F_Y(Y) < q))/\log(q) \quad (46)$$

where $F_X$ and $F_Y$ are the marginal distribution functions. The quantity $\chi(q)$ is bounded by

$$2 - \log(2q - 1)/\log(q) \leq \chi(q) \leq 1$$

where the lower bound is interpreted as $-\infty$ for $q \leq 1/2$ and zero for $q = 1$.

From Figure 1, we see that the multiple scaled NIG has stronger tail dependence than the standard NIG. By comparison (and for reference), it is well known that the Gaussian distribution has no tail dependence, and the *t*-distribution has a stronger tail dependence than both the Gaussian, the standard NIG and multiple scaled NIG. We also computed for illustration the empirical upper tail

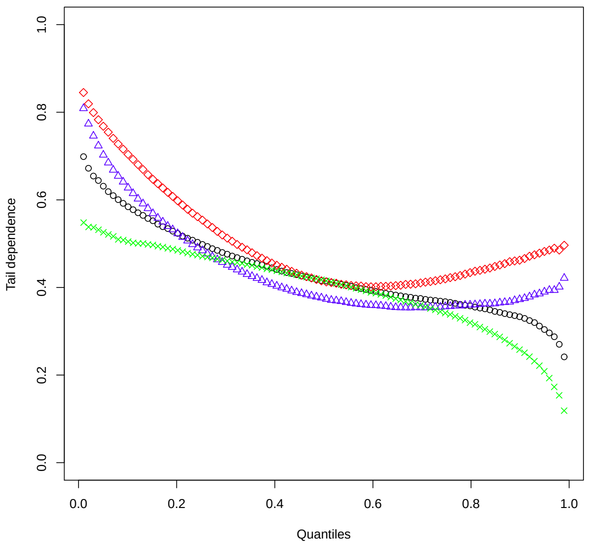


Figure 1: Comparison of tail dependence using $\chi(q)$. X-axis: quantiles levels. Y-axis: Estimate of $\chi(q)$. Gaussian distribution (Green), Standard NIG distribution (black), Multiple scaled NIG distribution (blue), $t$ distribution (Red)

dependence coefficient (as in Coles et al. [1999]) for each distribution and obtained 0.50, 0.42, 0.24 for the Student, MSNIG and NIG distribution respectively. The theoretical upper tail coefficient is 0 in the Gaussian case.

## D: MSGH Characteristic function and marginals

### Characteristic function

Denote by $\phi_{\mathbf{Y}}$ the characteristic function of a random vector $\mathbf{Y}$. It follows from (39) that, $\forall \mathbf{t} \in \mathbb{R}^M$, $\phi_{\mathbf{Y}}(\mathbf{t}) = E[\exp(i\mathbf{t}^T\mathbf{Y})] = E[E[\exp(i\mathbf{t}^T\mathbf{Y})|W]] = \exp(i\mathbf{t}^T\boldsymbol{\mu}) \prod_{m=1}^{M} \phi_{W_m}(u_m(\mathbf{t}))$ .

where $u_m(\mathbf{t}) = [\boldsymbol{A}^{1/2}\boldsymbol{D}^T\mathbf{t}]_m([\boldsymbol{A}^{1/2}\boldsymbol{D}^T\boldsymbol{\beta}]_m + \frac{i}{2}[\boldsymbol{A}^{1/2}\boldsymbol{D}^T\mathbf{t}]_m)$ and $\phi_{W_m}$ is the characteristic function of $W_m$.

In the Generalised Hyperbolic case $\phi_{W_m}$ is the characteristic function of a 1-dimensional $\mathcal{GIG}(\lambda_m, \gamma_m, \delta_m)$ distribution, which is

$$\phi_{W_m}(t) = \left(\frac{\gamma_m}{\gamma_m - 2it}\right)^{\lambda_m} \frac{K_{\lambda_m}(\delta_m\sqrt{\gamma_m^2 - 2it})}{K_{\lambda_m}(\delta_m\gamma_m)} \ . \tag{47}$$

The particular case of the multiple scaled NIG follows easily by setting $\lambda_m = -1/2$, which permits a simpler form

$$\phi_{W_m}(t) = \exp(\delta_m\gamma_m - \delta_m\sqrt{\gamma_m^2 - 2it}) \ . \tag{48}$$

The characteristic function is useful in practice for the computation of marginals as detailed in the next paragraph.

## Marginals

Using (9), marginals are easy to sample from but computing their pdfs involves, in general, numerical integration. An efficient and simple algorithm to compute such marginal pdfs in most cases can be derived according to Shephard [1991]. The derivation in Shephard [1991] is based on the inversion formula of the characteristic function which in the univariate case is:

$$\begin{aligned} f_Y(y) =& \frac{1}{2\pi} \int_0^\infty (\exp(ity)\phi_Y(-t) + \exp(-ity)\phi_Y(t))dt \\ =& \frac{1}{\pi} \int_0^\infty Re(\exp(-ity)\phi_Y(t))dt \end{aligned} \tag{49}$$

using the hermitian property of characteristic functions $\phi_Y(-t) = \overline{\phi_Y(t)}$ (the over line means the complex conjugate).

As an illustration, Figure 2 shows plots of the pdf of some 1-D marginals and a comparison with 1-D NIG distributions. From Figure 2 we can see that the marginals of the proposed multiple scaled NIG (MSNIG) distribution deviate slightly from the standard NIG distribution according to the specification of $\mathbf{\Sigma}$. The marginals of the MSNIG distribution are exactly 1-D standard NIG distributions in the diagonal scale matrix case.

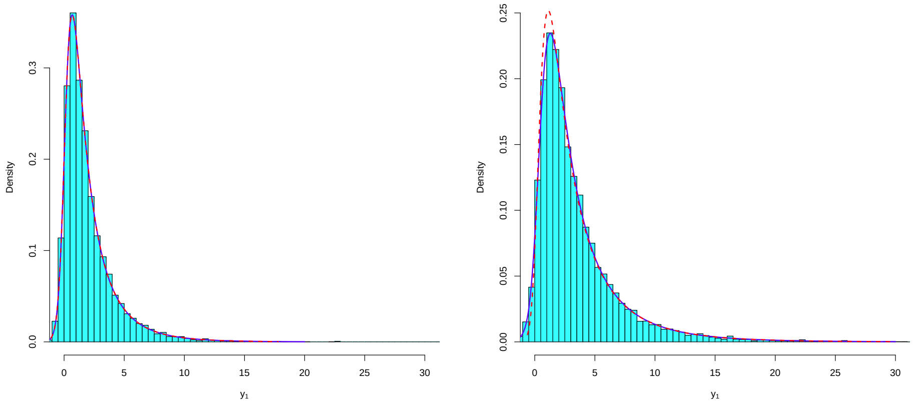


Figure 2: Histogram and density plots of the marginal $\mathbf{Y}_1$ of a bivariate NIG distribution with $\boldsymbol{\mu} = [0,0]^T$, $\boldsymbol{\gamma} = \boldsymbol{\delta} = \boldsymbol{\beta} = [2,2]^T$, and (left) diagonal $\mathbf{\Sigma}$ with diagonal entries equal to 1 or (right) $\mathbf{\Sigma}$ with diagonal entries equal to 1 and other entries to 0.5. Histograms and blue solid lines denote the multiple scaled NIG and red dashed lines the standard NIG.

For marginals of dimension greater than 1, we can also easily derive the characteristic function and use a simple multidimensional inversion formula. Let $\mathcal{I}$ be a subset of $\{1, \dots, M\}$ of size $I$ and write $\boldsymbol{Y}_{\mathcal{I}} = \{Y_m, m \in \mathcal{I}\}$ and $\boldsymbol{t}_{\mathcal{I}} = \{t_m, m \in \mathcal{I}\}$. The characteristic function of the marginal variable $\boldsymbol{Y}_{\mathcal{I}}$ is

$$\phi_{\boldsymbol{Y}_{\mathcal{I}}}(\boldsymbol{t}_{\mathcal{I}}) = \prod_{m\in\mathcal{I}} \exp(it_m\mu_m) \prod_{d=1}^{M} \phi_{W_d}(u_d(\boldsymbol{t}_{\mathcal{I}})) \ , \tag{50}$$

with $u_d(\boldsymbol{t}_{\mathcal{I}}) = (\sum\limits_{m\in\mathcal{I}} t_m[\boldsymbol{D}\boldsymbol{A}^{1/2}]_{md} \ [\boldsymbol{A}^{1/2}\boldsymbol{D}^T\boldsymbol{\beta}]_d) + \frac{i}{2}(\sum\limits_{m\in\mathcal{I}} t_m[\boldsymbol{D}\boldsymbol{A}^{1/2}]_{md})^2$ .

It follows that the density of $\boldsymbol{Y}_{\mathcal{I}}$ via the multidimensional inversion formula (see *e.g.* Shephard [1991]) is:

$$f_{\boldsymbol{Y}_{\mathcal{I}}}(\boldsymbol{y}_{\mathcal{I}}) = (2\pi)^{-I} \int\limits_{-\infty}^{\infty} \dots \int\limits_{-\infty}^{\infty} \exp(-i\boldsymbol{t}_{\mathcal{I}}^T\boldsymbol{y}_{\mathcal{I}}) \ \phi_{\boldsymbol{Y}_{\mathcal{I}}}(\boldsymbol{t}_{\mathcal{I}}) \ d\boldsymbol{t}_{\mathcal{I}} \tag{51}$$

When $I = 2$, and decomposing $\mathbb{R}^2$ into four quadrants,

$$f_{\boldsymbol{Y}_{\mathcal{I}}}(\boldsymbol{y}_{\mathcal{I}}) = 2 \ (2\pi)^{-2} \int\limits_{0}^{\infty} \int\limits_{-\infty}^{\infty} Re(\exp(-i\boldsymbol{t}_{\mathcal{I}}^T\boldsymbol{y}_{\mathcal{I}}) \ \phi_{\boldsymbol{Y}_{\mathcal{I}}}(\boldsymbol{t}_{\mathcal{I}})) \ d\boldsymbol{t}_{\mathcal{I}}. \tag{52}$$

This formula also generalizes easily in higher dimensions.

For illustration, Figure 3 shows the bivariate marginal $[Y_1, Y_2]^T$ for a 3 dimensional $[Y_1, Y_2, Y_3]^T$ following a MSNIG distribution with $\boldsymbol{\mu} = [0, 0, 0]^T$, $\boldsymbol{\gamma} = \boldsymbol{\delta} = [3, 3, 3]^T$, $\boldsymbol{\beta} = [-6, 2, 2]^T$ and $\boldsymbol{\Sigma}$ so that its diagonal entries are 1 and other entries are 0.5.

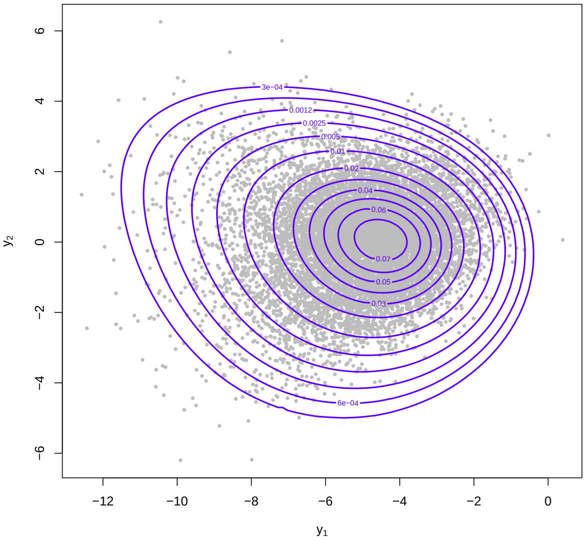


Figure 3: $[Y_1, Y_2]^T$ distribution when $[Y_1, Y_2, Y_3]^T$ follows a multiple scaled trivariate NIG distribution with $\boldsymbol{\mu} = [0, 0, 0]^T$, $\boldsymbol{\gamma} = \boldsymbol{\delta} = [3, 3, 3]^T$, $\boldsymbol{\beta} = [-6, 2, 2]^T$ and $\boldsymbol{\Sigma}$ so that its diagonal entries are 1 and other entries are 0.5. Contours are superimposed on points sampled from the distribution using equation (9).

# E: Algorithm for updating $\boldsymbol{D}$ in the EM algorithm (computing $\boldsymbol{D}^{(r+1)}$)

Using the equivalent parameterization (16), the goal is to minimize with respect to $\boldsymbol{D}$ the following quantity, where $\tilde{\boldsymbol{A}}$, $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\beta}}$ have been fixed to current estimations namely $\tilde{\boldsymbol{A}}^{(r)}$, $\boldsymbol{\mu}^{(r+1)}$ and $\tilde{\boldsymbol{\beta}}^{(r+1)}$,

$$\begin{aligned}
\boldsymbol{D}^{(r+1)} &= \arg\min_{\boldsymbol{D}} f(\boldsymbol{D}) \\
\text{where} \quad f(\boldsymbol{D}) &= \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}\boldsymbol{T}_i^{(r)}\tilde{\boldsymbol{A}}^{-1(r)}\boldsymbol{D}^T\boldsymbol{V}_i) + \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}\boldsymbol{S}_i^{(r)}\tilde{\boldsymbol{A}}^{-1(r)}\boldsymbol{D}^T\boldsymbol{B}_i) \\
&\quad - 2(\sum_{i=1}^{N} \text{trace}(\boldsymbol{D}\tilde{\boldsymbol{A}}^{-1}\boldsymbol{D}^T\boldsymbol{C}_i)
\end{aligned}$$

where $\boldsymbol{V}_i = (\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})(\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})^T, \boldsymbol{B}_i = \tilde{\boldsymbol{\beta}}^{(r+1)}\tilde{\boldsymbol{\beta}}^{T(r+1)}$, $\boldsymbol{C}_i = (\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})\tilde{\boldsymbol{\beta}}^{T(r+1)}$. Similarly to Celeux and Govaert [1995, see Appendix 2], we can derive from Flury and Gautschi [1986] the algorithm below.

**Step 1.** We start from an initial solution $\boldsymbol{D}^0 = [\boldsymbol{d}_1^0, \ldots, \boldsymbol{d}_M^0]$ where the $\boldsymbol{d}_m^0$'s are $M$-dimensional orthonormal vectors.
**Step 2.** For any couple $(l, m) \in \{1, \ldots, M\}^2$ with $l \neq m$, the couple of vectors $(\boldsymbol{d}_l, \boldsymbol{d}_m)$ is replaced with $(\boldsymbol{\delta}_l, \boldsymbol{\delta}_m)$ where $\boldsymbol{\delta}_l = [\boldsymbol{d}_l, \boldsymbol{d}_m]\boldsymbol{v}_1$ and $\boldsymbol{\delta}_m = [\boldsymbol{d}_l, \boldsymbol{d}_m]\boldsymbol{v}_2$ with $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ two orthonormal vectors of $\mathbb{R}^2$ such that $\boldsymbol{v}_1$ is the eigenvector associated to the smallest eigenvalue of the matrix

$$\begin{aligned}
\boldsymbol{M} \;=\; & \sum_{i=1}^{N}(\frac{t_{il}^{(r)}}{A_l^{(r)}} - \frac{t_{im}^{(r)}}{A_m^{(r)}})[d_l, d_m]^T\boldsymbol{V}_i[d_l, d_m] + \sum_{i=1}^{N}(\frac{s_{il}^{(r)}}{A_l^{(r)}} - \frac{s_{im}^{(r)}}{A_m^{(r)}})[d_l, d_m]^T\boldsymbol{B}_i[d_l, d_m] \\
& -2\sum_{i=1}^{N}(\frac{1}{A_l^{(r)}} - \frac{1}{A_m^{(r)}})[d_l, d_m]^T\boldsymbol{C}_i[d_l, d_m]
\end{aligned}$$

**Step 2** is repeated until it produces no decrease of the criterion $f(\boldsymbol{D})$.

Although not considered in this work, in a model-based clustering context, additional information for an efficient implementation can be found in Lin [2014].

## F: Details and corollary for the update of A in the EM algorithm

To update $\tilde{\boldsymbol{A}}$ we have to minimize the following quantity

$$\begin{aligned}
\tilde{\boldsymbol{A}}^{(r+1)} =& \arg\min_{\tilde{\boldsymbol{A}}} \Bigg\{ \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}^{(r+1)} \boldsymbol{T}_i^{(r)} \tilde{\boldsymbol{A}}^{-1} \boldsymbol{D}^{(r+1)T} \boldsymbol{V}_i) \\
&+ \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}^{(r+1)} \boldsymbol{S}_i^{(r)} \tilde{\boldsymbol{A}}^{-1} \boldsymbol{D}^{(r+1)T} \boldsymbol{B}_i^{(r)}) - \\
&2 \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}^{(r+1)} \tilde{\boldsymbol{A}}^{-1} \boldsymbol{D}^{(r+1)T} \boldsymbol{C}_i) + N \log|\tilde{\boldsymbol{A}}| \Bigg\} \\
=& \arg\min_{\tilde{\boldsymbol{A}}} \Bigg\{ \text{trace} \Bigg[ \sum_{i=1}^{N} (T_i^{(r)1/2} \boldsymbol{D}^{(r+1)T} \boldsymbol{V}_i \boldsymbol{D}^{(r+1)} \boldsymbol{T}_i^{(r)1/2} + \\
&\boldsymbol{S}_i^{(r)1/2} \boldsymbol{D}^{(r+1)T} \boldsymbol{B}_i \boldsymbol{D}^{(r+1)} \boldsymbol{S}_i^{(r)1/2} - \boldsymbol{D}^{(r+1)T} (\boldsymbol{C}_i + \boldsymbol{C}_i^T) \boldsymbol{D}^{(r+1)}) \tilde{\boldsymbol{A}}^{-1} \Bigg] \\
&+ N \text{log} |\tilde{\boldsymbol{A}}| \Bigg\} \\
=& \arg\min_{\tilde{\boldsymbol{A}}} \Bigg\{ \text{trace}((\sum_{i=1}^{N} \boldsymbol{M}_i)\ \tilde{\boldsymbol{A}}^{-1}) + N \text{log} \tilde{\boldsymbol{A}} \Bigg\}
\end{aligned}$$

where $\boldsymbol{M}_i = \boldsymbol{T}_i^{(r)1/2} \boldsymbol{D}^{(r+1)T} \boldsymbol{V}_i \boldsymbol{D}^{(r+1)} \boldsymbol{T}_i^{(r)1/2} + \boldsymbol{S}_i^{(r)1/2} \boldsymbol{D}^{(r+1)T} \boldsymbol{B}_i \boldsymbol{D}^{(r+1)} \boldsymbol{S}_i^{(r)1/2} - \boldsymbol{D}^{(r+1)T} (\boldsymbol{C}_i + \boldsymbol{C}_i^T) \boldsymbol{D}^{(r+1)}$ and $\boldsymbol{M}_i$ is a symmetric positive definite matrix.

For the update of $\tilde{\boldsymbol{A}}$, we can then use the following corollary (see Corollary A-2 in Celeux and Govaert [1995]) with $\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{M}_i$.

***Corollary 3.2:*** *The $M \times M$ diagonal matrix $\boldsymbol{A}$ minimizing trace$(\boldsymbol{S}\boldsymbol{A}^{-1})$ + $\alpha log|\boldsymbol{A}|$ where $\boldsymbol{S}$ is a $M \times M$ symmetric definite positive matrix and $\alpha$ is a positive real number is $\boldsymbol{A} = \frac{diag(\boldsymbol{S})}{\alpha}$*

By setting $\boldsymbol{D}$ and $\boldsymbol{\mu}$ to their current estimations $\boldsymbol{D}^{(r+1)}$ and $\boldsymbol{\mu}^{(r+1)}$ we then get,

$$\tilde{\boldsymbol{A}}^{(r+1)} = \frac{\text{diag}(\boldsymbol{S})}{N} \tag{53}$$

where

$$\begin{aligned}
\boldsymbol{S} =& \sum_{i=1}^{N} (\boldsymbol{T}_i^{(r)1/2} \boldsymbol{D}^{(r+1)T} \boldsymbol{V}_i \boldsymbol{D}^{(r+1)} \boldsymbol{T}_i^{(r)1/2} + \boldsymbol{S}_i^{(r)1/2} \boldsymbol{D}^{(r+1)T} \boldsymbol{B}_i \boldsymbol{D}^{(r+1)} \boldsymbol{S}_i^{(r)1/2} \\
&- \boldsymbol{D}^{(r+1)T} (\boldsymbol{C}_i + \boldsymbol{C}_i^T) \boldsymbol{D}^{(r+1)}) .
\end{aligned} \tag{54}$$

## G: Mixture setting and estimation

The results in sections 3.1 and 3.2 of the manuscript can be extended to cover the case of $K$-component mixture of multiple scaled NIG distributions. With the usual notation for the proportions $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$ and $\boldsymbol{\psi}_k = \{\boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \delta_k\}$ for $k = 1 \dots K$, we consider,

$$p(\mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k \mathcal{MSNIG}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{D}_k, \boldsymbol{A}_k, \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, \delta_k)$$

where $k$ indicates the $k$th component of the mixture and $\boldsymbol{\phi} = \{\boldsymbol{\pi}, \boldsymbol{\psi}\}$ with $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \dots \boldsymbol{\psi}_K\}$ the mixture parameters. In the EM framework, an additional variable $\boldsymbol{Z}$ is introduced to identify the missing class labels, where $\{Z_1, \dots, Z_N\}$ define the component of origin of the data $\{\mathbf{y}_1, \dots, \mathbf{y}_N\}$. In the light of the characterization of multiple scaled distributions, an equivalent modelling is: $\forall i \in \{1 \dots N\}$,
$\mathbf{Y}_i | \boldsymbol{W}_i = \boldsymbol{w}_i, Z_i = k \sim \mathcal{N}_M(\boldsymbol{\mu}_k + \boldsymbol{D}_k \boldsymbol{\Delta}_{\mathbf{w}_i} \boldsymbol{A}_k \boldsymbol{D}_k^T \boldsymbol{\beta}_k, \boldsymbol{D}_k \boldsymbol{\Delta}_{\mathbf{w}_i} \boldsymbol{A}_k \boldsymbol{D}_k^T)$ and $\boldsymbol{W}_i | Z_i = k \sim \mathcal{IG}(\gamma_{1k}, \delta_k) \otimes \dots \otimes \mathcal{IG}(\gamma_{Mk}, \delta_k)$ , where $\boldsymbol{\Delta}_{\mathbf{w}_i} = \text{diag}(w_{i1}, \dots, w_{iM})$. Inference using the EM algorithm with two sets of missing variables $\boldsymbol{Z} = \{Z_1, \dots, Z_N\}$ and $\boldsymbol{W} = \{\boldsymbol{W}_1, \dots, \boldsymbol{W}_N\}$ to fit such mixtures, is similar to the individual ML estimation.

Denote the parameters of the mixture in the equivalent parameterization (16) by $\boldsymbol{\phi} = \{\boldsymbol{\pi}, \tilde{\boldsymbol{\psi}}\}$ with $\tilde{\boldsymbol{\psi}} = \{\tilde{\boldsymbol{\psi}}_1, \dots \tilde{\boldsymbol{\psi}}_K\}$ the mixture parameters with $\tilde{\boldsymbol{\psi}}_k = \{\boldsymbol{\mu}_k, \boldsymbol{D}_k, \tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\gamma}}_k\}$ for $k = 1 \dots K$. For mixtures the EM algorithm iterates over the following two steps.

**E-step**

We denote by $\tau_{ik}^{(r)}$ the posterior probability that $\mathbf{y}_i$ belongs to the $k$th component of the mixture given the current estimates of the mixture parameters $\boldsymbol{\phi}^{(r)}$,

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \mathcal{MSNIG}(\mathbf{y}_i; \boldsymbol{\psi}_k^{(r)})}{p(\mathbf{y}; \boldsymbol{\phi}^{(r)})} \tag{55}$$

The conditional expectation of the complete data log-likelihood $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(r)})$ decomposes into three parts

$$Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(r)}) = Q_1(\boldsymbol{\pi}; \boldsymbol{\phi}^{(r)}) + Q_2(\tilde{\boldsymbol{\gamma}}; \boldsymbol{\phi}^{(r)}) + Q_3(\boldsymbol{\mu}, \boldsymbol{D}, \tilde{\boldsymbol{A}}, \tilde{\boldsymbol{\beta}}, ; \boldsymbol{\phi}^{(r)}) \tag{56}$$

with

$$Q_1(\boldsymbol{\pi}; \boldsymbol{\phi}^{(r)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik}^{(r)} \log \pi_k \tag{57}$$

$$Q_2(\tilde{\boldsymbol{\gamma}}; \boldsymbol{\phi}^{(r)}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik}^{(r)} \sum_{m=1}^{M} E_{W_{im}}[\log \mathcal{IG}(W_{im}; \tilde{\gamma}_{km}, 1) | \mathbf{y}_i, \boldsymbol{\phi}^{(r)}] \tag{58}$$

and

$$
\begin{aligned}
Q_3(\tilde{\gamma};\boldsymbol{\phi}^{(r)}) =& \sum_{i=1}^{N}\sum_{k=1}^{K}\tau_{ik}^{(r)}E_{\boldsymbol{W}_i}[\log\mathcal{N}_M(\boldsymbol{\mu}_k + \\
&\boldsymbol{D}_k\boldsymbol{\Delta}_{\mathbf{w}_i}\tilde{\boldsymbol{A}}_k\boldsymbol{D}_k^T\tilde{\boldsymbol{\beta}}_k, \boldsymbol{D}_k\boldsymbol{\Delta}_{\mathbf{w}_i}\tilde{\boldsymbol{A}}_k\boldsymbol{D}_k^T)|Z_i = k, \mathbf{y}_i, \boldsymbol{\phi}^{(r)}] \\
=& \sum_{i=1}^{N}\sum_{k=1}^{K}\tau_{ik}^{(r)}E_{\boldsymbol{W}_i}[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{D}_k\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{D}_k^T\tilde{\boldsymbol{\beta}}_k)^T\boldsymbol{D}_k\tilde{\boldsymbol{A}}^{-1}\boldsymbol{\Delta}_{\mathbf{w}_i}^{-1}\boldsymbol{D}_k^T \\
&\times(\mathbf{y}_i - \boldsymbol{\mu}_k - \boldsymbol{D}_k\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{D}_k^T\tilde{\boldsymbol{\beta}}_k)|Z_i = k, \mathbf{y}_i, \boldsymbol{\phi}^{(r)}] - \frac{1}{2}\log|\tilde{\boldsymbol{A}}_k|
\end{aligned}
\tag{59}
$$

ignoring constants.

Similarly to the E-step in Section 3.1, the quantities required for the E-step are given by,

$$
\begin{aligned}
s_{ikm}^{(r)} &= E[W_{im}|Z_i = k, \mathbf{y}_i; \boldsymbol{\phi}^{(r)})] = \frac{\phi_{ikm}^{(r)}K_0(\phi_{ikm}^{(r)}\hat{\alpha}_{km}^{(r)})}{\hat{\alpha}_{km}^{(r)}K_{-1}(\phi_{ikm}^{(r)}\hat{\alpha}_{km}^{(r)})} \\
t_{ikm}^{(r)} &= E[W_{im}^{-1}|Z_i = k, \mathbf{y}_i; \boldsymbol{\phi}^{(r)})] = \frac{\hat{\alpha}_{km}^{(r)}K_{-2}(\phi_{ikm}^{(r)}\hat{\alpha}_{km}^{(r)})}{\phi_{ikm}^{(r)}K_{-1}(\phi_{ikm}^{(r)}\hat{\alpha}_{km}^{(r)})}
\end{aligned}
$$

where

$$
\begin{aligned}
\phi_{ikm}^{(r)} &= \sqrt{1 + \frac{[\boldsymbol{D}_k^{(r)T}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r)})]_m^2}{\tilde{\boldsymbol{A}}_{km}^{(r)}}} \\
\hat{\alpha}_{km}^{(r)} &= \sqrt{\tilde{\boldsymbol{\gamma}}_{km}^{2(r)} + \frac{[\boldsymbol{D}_k^{(r)T}\tilde{\boldsymbol{\beta}}_k^{(r)}]_m^2}{\tilde{\boldsymbol{A}}_{km}^{(r)}}}
\end{aligned}
$$

**M-step**

**Updating the $\pi_k$'s**. The update of $\boldsymbol{\pi}$ is standard: for $k \in \{1, \ldots, K\}$, $\pi_k^{(r+1)} = \frac{n_k}{N}$ where $n_k = \sum_{i=1}^{N}\tau_{ik}^{(r)}$.

**Updating the $\boldsymbol{\mu}_k$'s**. It follows from the expression of $Q_3$ that for $k \in \{1, \ldots, K\}$, fixing $\boldsymbol{D}_k$ to the current estimation $\boldsymbol{D}_k^{(r)}$, leads for all $m = 1, \ldots, M$ to

$$
\begin{aligned}
\boldsymbol{\mu}_{km}^{(r+1)} =& \left(\frac{\sum_{i=1}^{N}\tau_{ik}\boldsymbol{T}_{ik}^{(r)}\boldsymbol{D}_k^{(r)T}}{n_k} - n_k\,(\sum_{i=1}^{N}\tau_{ik}\boldsymbol{S}_{ik}^{(r)})^{-1}\right)^{-1} \\
&\left(\frac{\sum_{i=1}^{N}\tau_{ik}\boldsymbol{T}_{ik}^{(r)}\boldsymbol{D}_k^{(r)T}\mathbf{y}_i}{n_k} - \sum_{i=1}^{N}\tau_{ik}\mathbf{y}_i\,(\sum_{i=1}^{N}\tau_{ik}\boldsymbol{S}_{ik}^{(r)})^{-1}\right)
\end{aligned}
$$

where $\boldsymbol{T}_{ik}^{(r)} = \mathrm{diag}(t_{ik1}^{(r)}, ..., t_{ikM}^{(r)})$ and $\boldsymbol{S}_{ik}^{(r)} = \mathrm{diag}(s_{ik1}^{(r)}, \ldots, s_{ikM}^{(r)})$.

**Updating the $\tilde{\boldsymbol{\beta}}_k$'s.** Similarly, it follows from the expression of $Q_3$ that for $k \in \{1, \ldots, K\}$, fixing $\boldsymbol{D}_k$ and $\boldsymbol{\mu}_k$ to their current estimation $\boldsymbol{D}_k^{(r)}$ and $\boldsymbol{\mu}_k^{(r)}$, leads to

$$\tilde{\boldsymbol{\beta}}_k^{(r+1)} = \boldsymbol{D}_k^{(r)}(\sum_{i=1}^{N} \tau_{ik}^{(r)} \boldsymbol{S}_{ik}^{(r)})^{-1} \boldsymbol{D}_k^{(r)T} \sum_{i=1}^{N} \tau_{ik}^{(r)} (\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})$$

**Updating the $\boldsymbol{D}_k$'s**

The parameter $\boldsymbol{D}_k$ is obtained by minimizing

$$\begin{aligned} \boldsymbol{D}_k^{(r+1)} = & \arg\min_{\boldsymbol{D}_k} \Bigg( \sum_{i=1}^{N} \mathrm{trace}(\boldsymbol{D}_k \boldsymbol{T}_{ik}^{(r)} \tilde{\boldsymbol{A}}_k^{(r)-1} \boldsymbol{D}_k^T \boldsymbol{V}_{ik}) \\ & + \sum_{i=1}^{N} \mathrm{trace}(\boldsymbol{D}_k \boldsymbol{S}_{ik}^{(r)} \tilde{\boldsymbol{A}}_k^{(r)-1} \boldsymbol{D}_k^T \boldsymbol{B}_{ik}) - 2(\sum_{i=1}^{N} \mathrm{trace}(\boldsymbol{D}_k \tilde{\boldsymbol{A}}_k^{(r)-1} \boldsymbol{D}_k^T \boldsymbol{C}_{ik}) \Bigg) \end{aligned}$$

where $\boldsymbol{V}_{ik} = \tau_{ik}^{(r)}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})^T$, $\boldsymbol{B}_{ik} = \tau_{ik}^{(r)} \tilde{\boldsymbol{\beta}}_k^{(r+1)} \tilde{\boldsymbol{\beta}}_k^{(r+1)T}$ and $\boldsymbol{C}_{ik} = \tau_{ik}^{(r)}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})\tilde{\boldsymbol{\beta}}_k^{(r+1)T}$

The parameter $\boldsymbol{D}_k$ can be updated using an algorithm derived from Flury and Gautschi [see Flury and Gautschi, 1986, and section C].

**Updating the $\tilde{\boldsymbol{A}}_k$'s.** We have to minimize the following quantity:

$$\tilde{\boldsymbol{A}}_k^{(r+1)} = \arg\min_{\tilde{\boldsymbol{A}}_k} (\mathrm{trace}(\sum_{i=1}^{N} \boldsymbol{M}_{ik} \tilde{\boldsymbol{A}}_k^{-1}) + \alpha_k \log|\tilde{\boldsymbol{A}}_k|)$$

where $\boldsymbol{M}_{ik} = \boldsymbol{T}_{ik}^{(r)1/2} \boldsymbol{D}_k^{(r+1)T} \boldsymbol{V}_{ik} \boldsymbol{D}_k^{(r+1)} \boldsymbol{T}_{ik}^{(r)1/2} + \boldsymbol{S}_{ik}^{(r)1/2} \boldsymbol{D}_k^{(r+1)T} \boldsymbol{B}_{ik} \boldsymbol{D}_k^{(r+1)} \boldsymbol{S}_{ik}^{(r)1/2} - \boldsymbol{D}_k^{(r+1)T}(\boldsymbol{C}_{ik} + \boldsymbol{C}_{ik}^T)\boldsymbol{D}_k^{(r+1)}$ is a symmetric positive definite matrix and $\alpha_k = \sum_{i=1}^{N} \tau_{ik}^{(r)}$

Using Corollary (see Section 3) leads for all $m = 1, \ldots, M$ to

$$\begin{aligned} \tilde{\boldsymbol{A}}_{km}^{(r+1)} = & \frac{1}{\sum\limits_{i=1}^{N} \tau_{ik}^{(r)}} \sum_{i=1}^{N} \tau_{ik}^{(r)} \Bigg( [\boldsymbol{D}_k^{(r+1)T}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})]_m^2 t_{ikm}^{(r)} + [\boldsymbol{D}_k^{(r+1)T} \tilde{\boldsymbol{\beta}}_k^{(r+1)}]_m^2 s_{ikm}^{(r)} \\ & - 2[\boldsymbol{D}_k^{(r+1)T}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})]_m [\boldsymbol{D}_k^{(r+1)T} \tilde{\boldsymbol{\beta}}_k^{(r+1)}]_m \Bigg) \end{aligned}$$

**Updating the $\tilde{\boldsymbol{\gamma}}_k$'s.** To update $\tilde{\boldsymbol{\gamma}}_k$ we have to minimize,

$$\tilde{\boldsymbol{\gamma}}_k^{(r+1)} = \arg\min_{\tilde{\boldsymbol{\gamma}}} \Bigg\{ \sum_{i=1}^{N} \tau_{ik}^{(r)} \sum_{m=1}^{M} \frac{1}{2} \tilde{\gamma}_{km}^2 s_{ikm}^{(r)} - \tilde{\gamma}_{km}) \Bigg\}$$

which leads for all $m = 1, \ldots, M$ to

$$\tilde{\gamma}_{km}^{(r+1)} = \frac{n_k}{\sum_{i=1}^{N} \tau_{ik} s_{ikm}}$$

To transform the estimated parameters back to the original ones, $\delta_k = |\tilde{\boldsymbol{A}}_k|^{\frac{1}{2M}}, \gamma_{km} = \tilde{\boldsymbol{\gamma}}_{km}/\delta_k, \boldsymbol{\beta}_k = \boldsymbol{D}_k \tilde{\boldsymbol{A}}_k^{-1} \boldsymbol{D}_k^T \tilde{\boldsymbol{\beta}}_k, \boldsymbol{A}_k = \tilde{\boldsymbol{A}}_k / |\tilde{\boldsymbol{A}}_k|^{\frac{1}{M}}$

## H: Simulated Data - Clusters of MSNIG distributions

In this section, we assess the classification performance of a mixture of MSNIG distributions using a simulated dataset and compare the results to estimation using mixtures of respectively standard multivariate NIG, MSGH$^{TFBM}$ and Coalesced GH distributions. We consider the case of two clusters each sampled from a 2-dimensional MSNIG distribution which are slightly separated from each other, using the parameter values outlined in Table 1. For both clusters the sample size is 500. A plot of the simulated data is shown in Figure 4 with the observations belonging to each cluster labelled by different colours.

Table 1: Parameter values for simulated dataset

| Parameters | Cluster 1 | Cluster 2 |
|---|---|---|
| $\boldsymbol{\mu}$ | (0.0, 0.0) | (-4.0, 0.0) |
| $\boldsymbol{\beta}$ | (0.0, -10.0) | (-3.0, -5.0) |
| $\boldsymbol{D}$ | $\begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}$ | $\begin{pmatrix} \cos(3\pi/8) & -\sin(3\pi/8) \\ \sin(3\pi/8) & \cos(3\pi/8) \end{pmatrix}$ |
| $\boldsymbol{A}$ | $\text{diag}(3/2, 2/3)$ | $\text{diag}(3/2, 2/3)$ |
| $\boldsymbol{\gamma}$ | (2.0, 2.0) | (2.0, 2.0) |
| $\delta$ | 1.0 | 1.0 |

The classification results for the MSNIG, NIG, MSGH$^{TFBM}$ and Coalesced GH mixtures over 30 simulated datasets (using the parameters in Table 1) are summarized in Table 2 where we report the Adjusted Rand Indices (ARI) [Hubert and Arabie, 1985] and Brier [Brier, 1950] scores. In contrast to the ARI, the Brier score incorporates the uncertainty of the classification and lower values represent a better classification. Figure 5 also shows the box plots for the Adjusted Rand Index of NIG, MSNIG, MSGH$^{TFBM}$, and Coalesced GH mixtures. The fitted contour and observations assigned to each cluster for MSNIG and NIG for one of the simulated datasets are shown in Figure 6. From these results we can see quite clearly that the classification performance of the MSNIG is very good with an ARI of 0.89. Compared to the results of the NIG (ARI=0.79), the difference in the classification performance appears to be in the tails of the two clusters with the MSNIG better capturing the heavy tails of both clusters. The results for MSGH$^{TFBM}$ and Coalesced GH appeared to be worse than for the NIG (ARI=0.70 and 0.74, respectively). Similar results to the

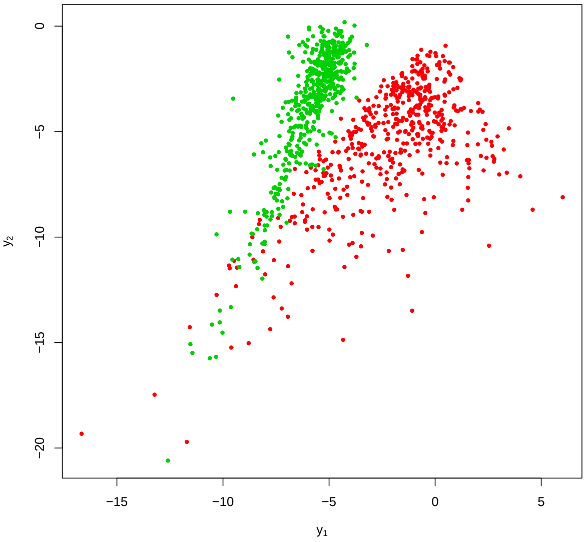


Figure 4: Simulated data showing the true classes

NIG were found for the skew-t and skew normal [Sahu et al., 2003, Lee and McLachlan, 2012, Lin, 2010] using the R package **mixsmsn** [Cabral et al., 2012].

In general, the results can be used to demonstrate that small changes in the tail behaviour of the true clusters can have a significant effect on the ability to accurately classify observations.

Table 2: Classification results for the MSNIG, NIG, MSGH$^{TFBM}$ and Coalesced GH mixtures for 30 simulated datasets. The average ARI and Brier score are reported with their standard deviations in parenthesis.

| Measure | MSNIG | NIG | MSGH$^{TFBM}$ | Coalesced GH |
|---|---|---|---|---|
| ARI | 0.89 (0.02) | 0.79 (0.03) | 0.70 (0.09) | 0.74 (0.12) |
| Brier score | 0.04 (0.01) | 0.09 (0.01) | 0.15 (0.05) | 0.12 (0.07) |

# I: Petroleum data

This data consists of 655 petroleum samples collected from the Montrose quadrangle of Western Colorado. The samples consist of log-concentration readings for a number of chemical elements, and are part of a multivariate dataset originally described by Cook and Johnson [1981]. The dataset is often used to compare and contrast different copula approaches [Genest and Rivest, 1993]. For ease of analysis and presentation we concentrate on two of the elements Cobalt ($Co$) and Uranium ($U$). Figure 7 provides a scatterplot of the data overlaid with contour lines for the standard NIG (red dashed) and multiple scaled NIG (blue) displayed. From the contour lines we can see that the multiple scaled NIG provides a better fit to the data and this is also evidenced by significantly

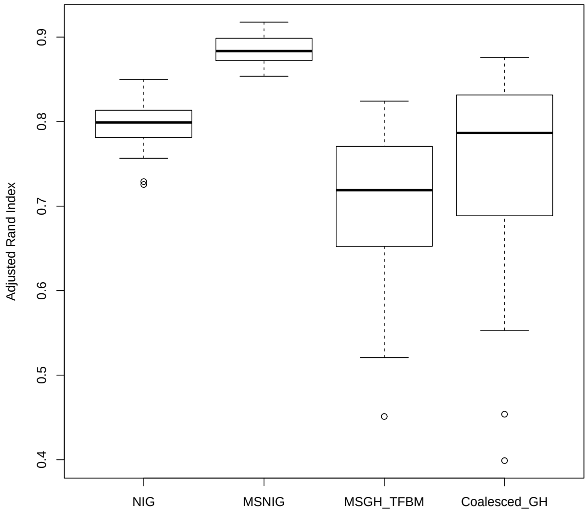


Figure 5: Boxplots of the ARI for 30 simulated datasets fitted respectively with a 2 component NIG, MSNIG, MSGH$^{TFBM}$, and Coalesced GH mixture.

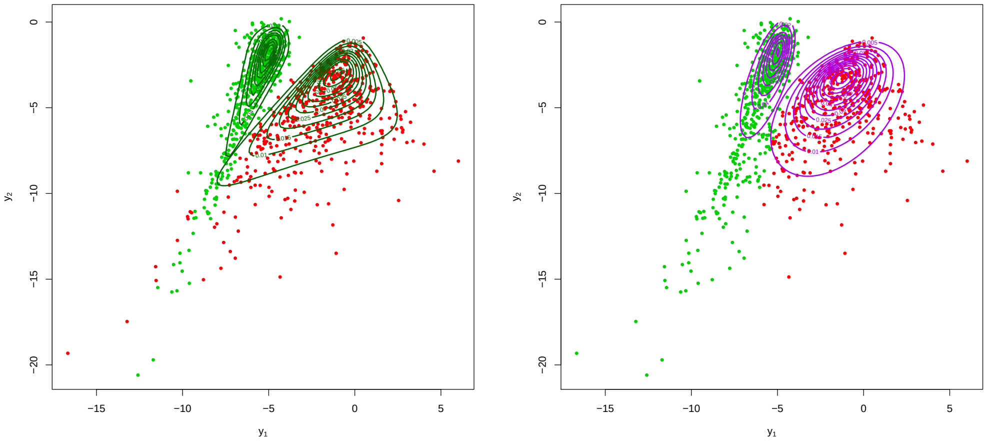


Figure 6: Plot of fitted contour and estimated classes for MSNIG (left) and NIG (right) from one of the simulated datasets

higher likelihood and BIC estimates for the multiple scaled NIG ($\mathcal{L} = 207.5$, BIC = -357) compared to the standard NIG ($\mathcal{L} = 168.4$, BIC = -331).

Table 3: Estimated parameters for MSNIG and NIG on the Petroleum data ($Co$ v. $U$)

| Parameters | MSNIG | NIG |
|---|---|---|
| $\boldsymbol{\mu}$ | (0.96,0.35) | (0.99,0.46) |
| $\boldsymbol{\beta}$ | (2.73,13.57) | (2.10,5.25) |
| $\boldsymbol{D}$ | $\begin{pmatrix} 0.06 & -0.99 \\ 0.99 & 0.06 \end{pmatrix}$ | - |
| $\boldsymbol{A}$ | $\mathrm{diag}(1.08, 0.93)$ | - |
| $\Sigma$ | - | $\begin{pmatrix} 0.51 & -0.01 \\ -0.01 & 1.97 \end{pmatrix}$ |
| $\boldsymbol{\gamma}$ | (8.17,14.69) | 8.77 |
| $\delta$ | 0.28 | 0.33 |
| Log-like | 207.6 | 168.4 |

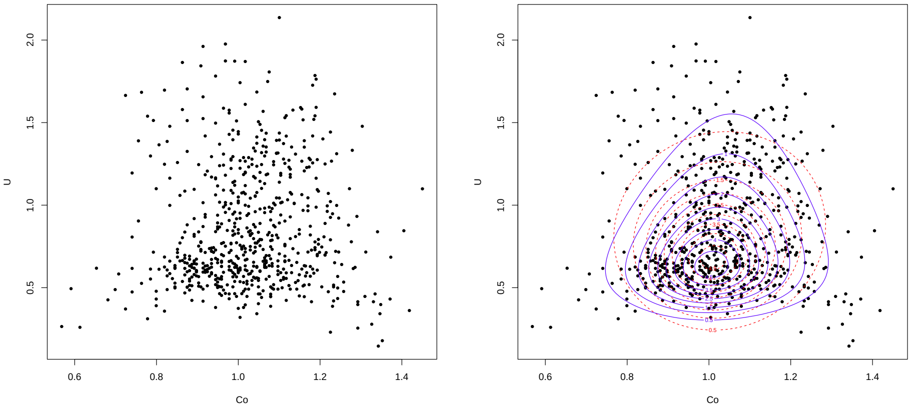
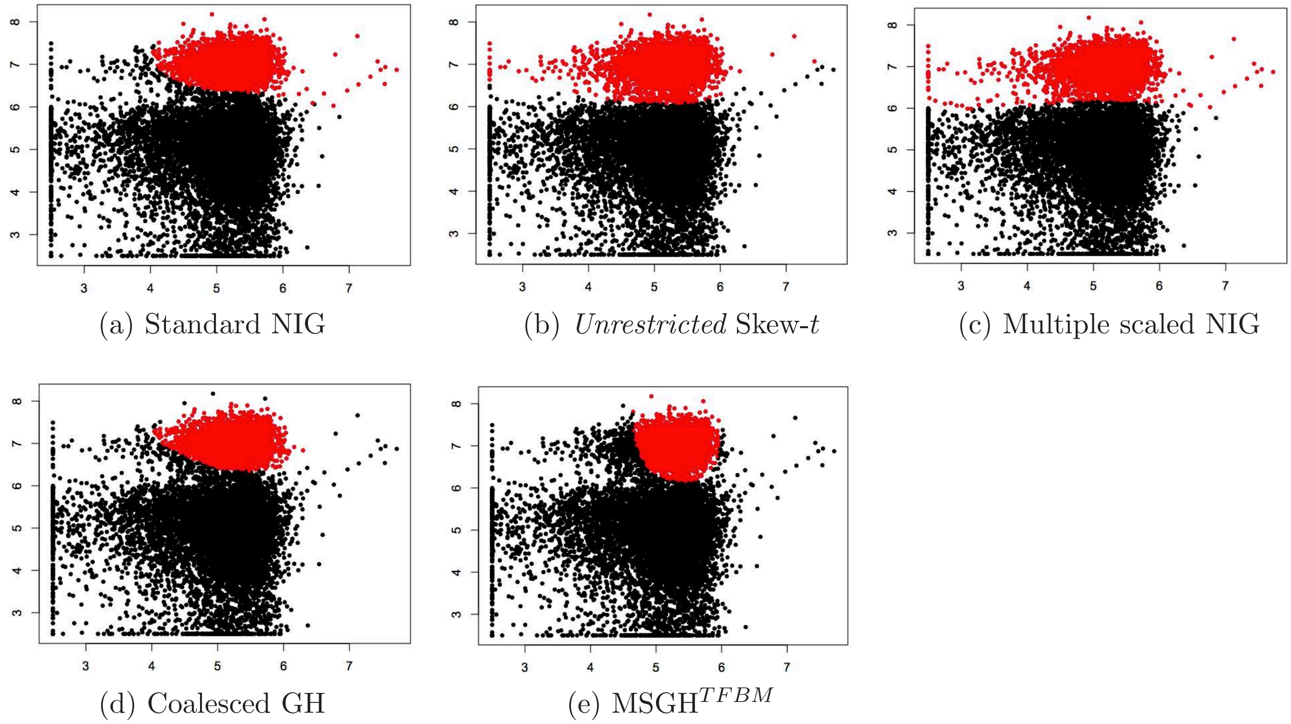


Figure 7: Scatterplot of petroleum data ($Co$ v. $U$). Right panel: Comparison of standard NIG (red, dashed line) versus MSNIG (blue).

## J: Mixture of coalesced GH distributions

In this section, we provide further results obtained on the Lymphoma data set (Section 4.3 of the manuscript) using the **MixGHD** R package [Tortora et al., 2014a] implementing the model described in the arXiv paper [Tortora et al., 2014b]. The contour plots are shown in Figures 5 and 6 (e,f) of the manuscript. Figure 8 (d,e) below shows the corresponding classification results for the plots in the manuscript Figure 5 (e,f). For comparison we provide classifications for the models in Figure 5 (a,b,d) of the manuscript. The coalesced GH mixture is clearly providing less satisfying classification results.



(a) Standard NIG

(b) *Unrestricted* Skew-$t$

(c) Multiple scaled NIG

(d) Coalesced GH

(e) MSGH$^{TFBM}$

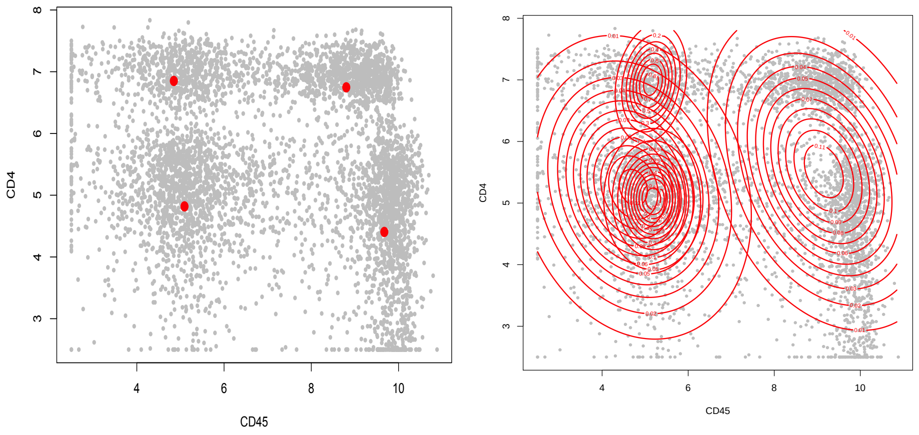
Figure 8: Lymphoma data, $CD4$ v. $ZAP70$. Classification results for: (a) Standard NIG [Karlis and Santourian, 2009]; (b) *Unrestricted* Skew-$t$ [Sahu et al., 2003]; (c) Multiple scaled NIG; (d) Coalesced GH [Tortora et al., 2014b] and (e) Multiple scaled GH [Tortora et al., 2014b].

For the second Lymphoma data sets (Figure 6 in the manuscript), we also provide a complementary plot below showing that the coalesced GH mixture starts from a reasonable initialization of the cluster centers (a). This suggests that initialization issues were probably not responsible for the not very satisfying results obtained in (b) and shown in Figure 6 (e) of the manuscript.

Then, we ran the **MixGHD** package with $\lambda$ set to -1/2, which corresponds to NIG distributions. The resulting MSNIG$^{TFBM}$ distribution (Figure 10) does not behave much better than its MSGH$^{TFBM}$ generalization. Also we observed (Figure 11) that the GH parameterization proposed in [Browne and McNicholas, 2013] provided results very close to the standard NIG distribution.

## K: Application in flow cytometry: Lymphoma

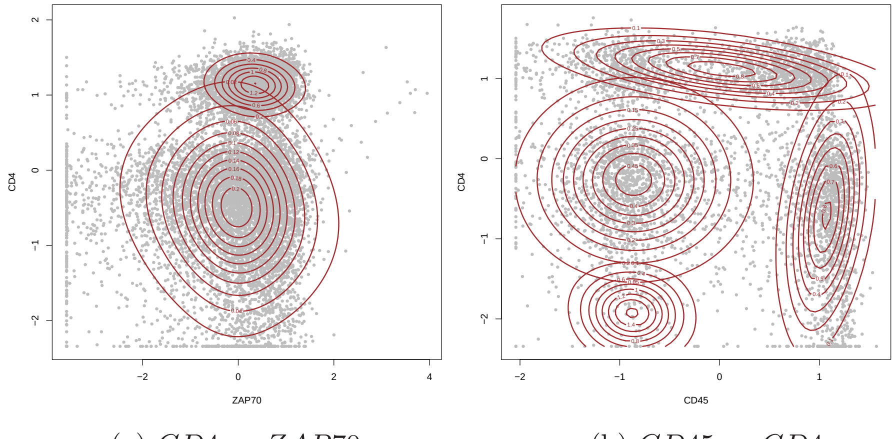
In this section, we compare the classification performance of the different approaches on a flow cytometry problem using lymphoma data where the true group labels are known (through manual gating). This data set is different from the lymphoma data set used in section 4.1 of the manuscript. The data is available in the R package **EMMIXuskew** and is a sample from the Diffuse Large B-cell Lymphoma (DLBCL) dataset from [Aghaeepour et al, 2013]. The original data contained



(a) initial centers

(b) final results

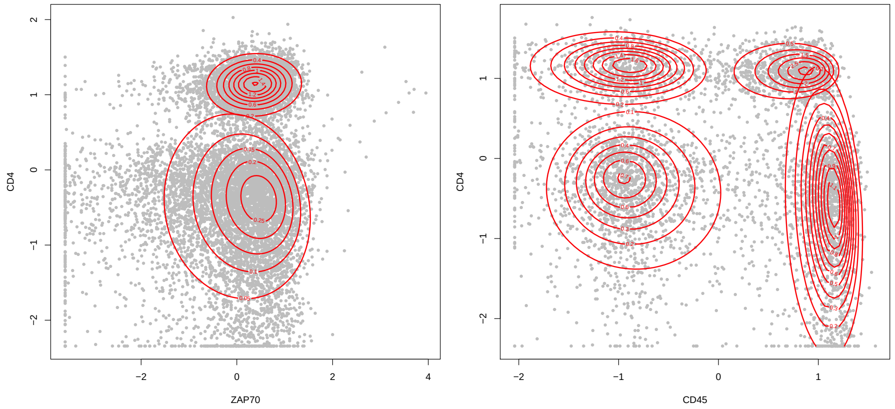
Figure 9: Lymphoma data, $CD45$ v. $CD4$. (a) Initial cluster centers for the mixture of coalesced GH distributions [Tortora et al., 2014b] leading to the results in (b).



(a) $CD4$ v. $ZAP70$

(b) $CD45$ v. $CD4$

Figure 10: Lymphoma data results for mixture of MSNIG$^{TFBM}$ distributions ($\lambda = -1/2$ in [Tortora et al., 2014b]).



(a) $CD4$ v. $ZAP70$ (b) $CD45$ v. $CD4$

Figure 11: Lymphoma data results for the mixture of GH distributions proposed in [Browne and McNicholas, 2013].

measurements from biopsies of 30 DLBCL patients with each sample stained with three antibodies, CD3, CD5, and CD19. The data is a subset from one patient and we compare the classification performance of the different approaches on two of the more difficult groups to identify (See Figure 12).

Table 4: Classification results for Diffuse Large B-cell Lymphoma (DLBCL) data

| Model | Log-likelihood | BIC | ARI[c] | Brier score |
|---|---|---|---|---|
| MSNIG | -51,694 | 103,667 | 0.80 | 0.08 |
| NIG | -51,607 | 103,457 | 0.73 | 0.11 |
| Skew-t (U) | -51,792 | 103,827 | 0.66 | 0.13 |
| Skew-t | -51,655 | 103,544 | 0.72 | NA[a] |
| Skew-normal | -51,688 | 103,577 | 0.72 | NA[a] |
| Coalesced GH | NA[b] | NA | 0.77 | 0.11 |
| MSGH$^{TFBM}$ | NA[b] | NA | 0.80 | 0.10 |

[a] These values were not able to be evaluated from the output of the R package **mixsmsn**

[b] The log-likelihood values from the output of the R package **MixGHD** are based on a rescaled version of the data and thus are not comparable to the results from the other approaches. Attempts to use the original scale of the data for **MixGHD** appeared to cause problems in the algorithm.

[c] Although the above results are based on the best results from 10 different initialisations, all of the approaches displayed some sensitivity to the initial values chosen and more accurate results may be possible.

In this example, both the MSNIG and the MSGH$^{TFBM}$ appear to provide the best classification performance (ARI=0.80), with the MSNIG having a slightly lower Brier score, indicating less uncertainty about the classification (Brier score = 0.08).

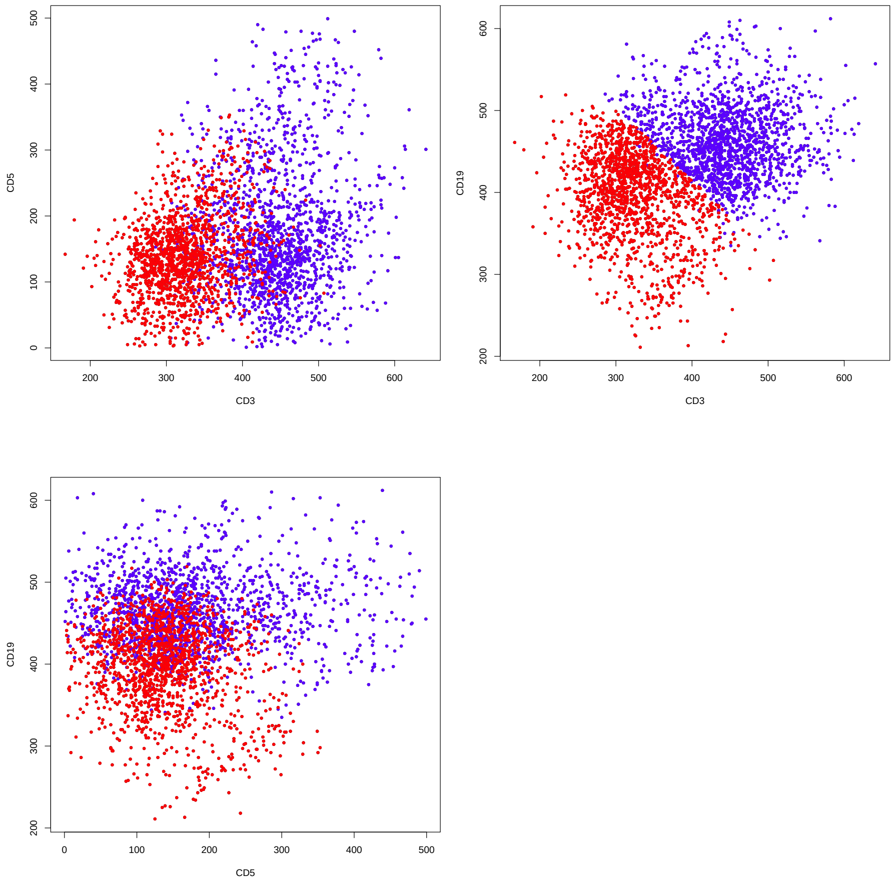


Figure 12: Manual gating of two groups (red (group 1) and blue (group 2)) for Diffuse Large B-cell Lymphoma (DLBCL) data [Aghaeepour et al, 2013]

## References

K. Aas and I. Hobaek Haff. The generalised hyperbolic skew Student's t-distribution. *Journal of Financial Econometrics*, 4(2):275–309, 2006.

K. Aas, I. Hobaek Haff, and X. Dimakos. Risk estimation using the multivariate normal inverse Gaussian distribution. *Journal of Risk*, 8(2):39–60, 2005.

Aghaeepour et al. Critical assessment of automated flow cytometry analysis techniques. *Nature Methods*, 10:228–238, 2013.

O. Barndorff-Nielsen. Exponentially Decreasing Distributions for the Logarithm of Particle Size. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 353 (1674):401–419, 1977.

O. Barndorff-Nielsen, J. Kent, and M. Sorensen. Normal variance-mean mixtures and z Distributions. *International Statistics Review*, 50(2):145–149, 1982.

G.W. Brier. Verification of forecasts expressed in terms of probability. *Month. Weather Rev.*, 78, 1950.

R. Browne and P. McNicholas. A mixture of generalized hyperbolic distributions. arXiv:1305.1036, 2013.

C.S. Cabral, V.H. Lachos, and M.O. Prates. Multivariate mixture modelling using skew-normal independent distributions. *Computational Statistics and Data Analysis*, 56:126–142, 2012.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28: 781–793, 1995.

S. G. Coles, J. Heffernan, and J. A. Tawn. Dependence measures for extreme value analyses. *Extremes*, 2:339–365, 1999.

R.D. Cook and M.E. Johnson. A family of distributions for modeling nonelliptically symmetric multivariate data. *Journal of the Royal Statistical Society, Series B*, 43:210–218, 1981.

B. N. Flury and W. Gautschi. An Algorithm for Simultaneous Orthogonal Transformation of Several Positive Definite Symmetric Matrices to Nearly Diagonal Form. *SIAM Journal on Scientific and Statistical Computing*, 7(1):169–184, 1986.

C. Genest and L.P. Rivest. Statistical Inference Procedures for Bivariate Archimedean Copulas. *Journal of the American Statistical Association*, 88(423):1034–1043, 1993.

J.E. Griffin and P.J. Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

J.L. Jensen. On the hyperboloid distribution. *Scandinavian Journal of Statistics*, 8(4):193–206, 1981.

B. Jorgensen. Statistical Properties of the Generalized Inverse Gaussian Distribution. In *Lecture Notes in Statistics*. Springer, New York, 1982.

D. Karlis and A. Santourian. Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, 19:73–83, 2009.

S.X. Lee and G.J. McLachlan. On the fitting of mixtures of multivariate skew $t$-distributions. 2012.

T-I. Lin. Robust mixture modelling using multivariate skew-$t$ distribution. *Statistics and Computing*, 20:343–356, 2010.

T-I. Lin. Learning from incomplete data via parameterized t mixture models through eigenvalue decomposition. *Computational Statistics and Data Analysis*, 71:183–195, 2014.

S.K. Sahu, D.K. Dey, and M.D. Branco. A new class of multivariate skew distributions with applications to Bayesian regression models. *The Canadian Journal of Statistics*, 31:129–150, 2003.

R. Schmidt, T. Hrycej, and E. Stutzle. Multivariate distribution models with generalized hyperbolic margins. *Computational Statistics and Data Analysis*, 50:2065–2096, 2006.

N. Shephard. From characteristic function to distribution function: a simple framework for the theory. *Econometric theory*, 7(4):519–529, 1991.

R Development Core Team. *R: A language and environment for statistical computing.* ISBN 3-900051-07-0, URL http://www.R-project.org/, 2011.

C. Tortora, R. P. Browne, B. C. Franczak, and P. D. McNicholas. MixGHD: Model based clustering and classification using the mixture of generalized hyperbolic distributions. *Version 1.0*, July 2014a.

C. Tortora, B.C. Franczak, R. Browne, and P. McNicholas. Model-based clustering using mixtures of coalesced generalized hyperbolic distributions. arXiv:1403.2332v3, 2014b.