A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweights: Application to robust clustering

Florence Forbes, Darren Wraith

Noname manuscript No.
(will be inserted by the editor)

A New Family of Multivariate Heavy-tailed Distributions with Variable Marginal Amounts of Tailweight: Application to Robust Clustering

Florence Forbes · Darren Wraith

Abstract We propose a family of multivariate heavy-tailed distributions that allow variable marginal amounts of tailweight. The originality comes from introducing multidimensional instead of univariate scale variables for the mixture of scaled Gaussian family of distributions. In contrast to most existing approaches, the derived distributions can account for a variety of shapes and have a simple tractable form with a closed-form probability density function whatever the dimension. We examine a number of properties of these distributions and illustrate them in the particular case of Pearson type VII and t tails. For these latter cases, we provide maximum likelihood estimation of the parameters and illustrate their modelling flexibility on simulated and real data clustering examples.

Keywords Covariance matrix decomposition · EM algorithm · Gaussian scale mixture · Multivariate generalized t-distribution · Outlier detection

1 Introduction

A popular way to approach clustering tasks is via a parametric finite mixture model. The vast majority of the work on such mixtures has been based on Gaussian mixture models (see *e.g.* Fraley and Raftery [2002]). In comparison, the use of mixtures of multivariate t-distributions for clustering has received considerably less attention. Typically, in some applications the tails of Gaussian distributions are shorter than appropriate or parameter estimations are affected by atypical observations (outliers). In contrast to the Gaussian case, no closed-form solution exists for the t-distribution. However, tractability is maintained, both in the univariate and multivariate case, via the use of the EM algorithm [McLachlan and Peel, 2000b, Bishop and Svensen, 2005, Archambeau and Verleysen, 2007] and thanks to

F. Forbes, D. Wraith
INRIA, Laboratoire Jean Kuntzman, Mistis team
655 avenue de l'Europe, Montbonnot
38334 Saint-Ismier Cedex, France
Tel.: +33-4-76 61 52 50
Fax: +33-4-76 61 52 52
E-mail: firstname.lastname@inria.fr

a useful representation of the t-distribution as a so-called *infinite mixture of scaled Gaussians* or *Gaussian scale mixture*. A Gaussian scale mixture distribution is a distribution of the form:

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) = \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)\ f_W(w;\boldsymbol{\theta})\ \mathrm{d}w \tag{1}$$

where $\mathcal{N}_M(\ .\ ;\boldsymbol{\mu},\boldsymbol{\Sigma}/w)$ denotes the M-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/w$ and f_W is the probability distribution of a univariate positive variable W referred to hereafter as the weight variable.

There exist quite a few forms of the multivariate t-distribution [Kotz and Nadarajah, 2004, Nadarajah and Kotz, 2004, Nadarajah and Dey, 2005] with many of the cited variations focusing on introducing non-centrality or skewness. Among all the possible multivariate presentations, the most common form considered is obtained when f_W is a Gamma distribution $\mathcal{G}(\nu/2,\nu/2)$ where ν denotes the degrees of freedom (we shall denote the Gamma distribution when the variable is X by $\mathcal{G}(x;\alpha,\gamma) = x^{\alpha-1}\Gamma(\alpha)^{-1}\exp(-\gamma x)\gamma^\alpha$ where Γ denotes the Gamma function). Using this form the standard density denoted by $t_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ of the M-dimensional t-distribution with parameters $\boldsymbol{\mu}$ (real location vector), $\boldsymbol{\Sigma}$ ($M\times M$ real positive definite scale matrix) and ν (positive real degrees of freedom parameter) is given by

$$\begin{aligned} t_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) &= \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}/w)\ \mathcal{G}(w;\nu/2,\nu/2)\ \mathrm{d}w \\ &= \frac{\Gamma((\nu+M)/2)}{|\boldsymbol{\Sigma}|^{1/2}\ \Gamma(\nu/2)\ (\pi\nu)^{M/2}}\ [1+\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma})/\nu]^{-(\nu+M)/2} \end{aligned} \tag{2}$$

where $\delta(\mathbf{y},\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$ is the Mahalanobis distance between $\mathbf{y}$ and $\boldsymbol{\mu}$ (^{T} means transpose). Note that $\boldsymbol{\mu}$ is the mean when $\nu>1$ but $\boldsymbol{\Sigma}$ is not strictly speaking the covariance matrix of the t-distribution which is $\nu/(\nu-2)\boldsymbol{\Sigma}$ when $\nu>2$.

A difficulty with the standard representation of the t-distribution is that when $\boldsymbol{\Sigma}$ is diagonal this representation can be shown to have zero correlation but the marginal distributions are not statistically independent. Equivalently the product of independent univariate t-distributions with the same degrees of freedom parameter is not a standard multivariate t-distribution with a diagonal scale matrix. We will see that the multivariate generalization we propose has in contrast this property and contains the product of independent t-distributions as a particular case.

Also, as mentioned by Kotz and Nadarajah [2004], the standard t-distribution belongs to the class of elliptically contoured distributions (see for instance Fang et al. [2002] for a definition of elliptical distributions). We will see in the next section that our generalization allows for a greater variety of shapes and in particular contours that are not necessarily elliptic. Note however that our proposal is different from the meta-elliptical distributions of Fang et al. [2002].

Another difficulty or limitation of the standard representation in (2) is that all marginals are t-distributions with the same degrees of freedom parameter ν and hence the same amount of tailweight. It is not then possible to account for very different tail behaviors across dimensions, such as a Gaussian (infinite *dof*) tail in one dimension and a Cauchy (*dof*=1) tail in an other dimension [Azzalini

and Genton, 2008]. In his work, Jones [2002] proposes a dependent bivariate t-distribution with marginals of different degrees of freedom but the tractability of the extension to the multivariate case is unclear. Additional proposals are reviewed in chapters 4 and 5 of Kotz and Nadarajah [2004] but these formulations tend to be appreciably more complicated, often already in the expression of the probability density function. Increasingly, there has been much research on copula approaches to account for flexible distributional forms but the choice as to which one to use in this case and the applicability to (even) moderate dimensions is also not clear [Giordani et al., 2008]. At least one other thread of work has emerged based on copulas: the grouped t copula of Demarta and McNeil [2005] and Daul et al. [2003]. In general the papers take various approaches whose relationships have been characterized in the bivariate case by Shaw and Lee [2008]. However, most of the existing approaches suffer either from the non-existence of a closed-form pdf or from a difficult generalization to more than two dimensions.

In this paper, we show that the scale mixture representation can be further explored and propose a framework that is considerably simpler than those previously proposed with distributions exhibiting interesting properties. We extend the standard t-distribution to allow for the degrees of freedom parameter to be set or estimated differently in each dimension of the variable space. The key elements of the approach are the introduction of multidimensional weights and a decomposition of the matrix $\boldsymbol{\Sigma}$ in (1) which facilitates the separate estimation and also allows for arbitrary correlation between dimensions. More generally, we define a new family of multivariate heavy-tailed distributions which includes our generalized t-distribution but also other generalizations such as those of the Pearson type VII, the so-called K and the Normal Inverse Gaussian distributions. For illustration, we focus on robust clustering and assess the performance of the new family on simulated and real datasets particularly challenging to the standard t-mixture and also to many alternative clustering approaches.

The paper is outlined as follows. The new family is outlined in Section 2. In Section 3, we examine a number of basic properties and characteristics of these distributions and illustrate them on the t and Pearson type VII cases. For these latter cases, we provide in Section 4, maximum likelihood estimation of the parameters via the EM algorithm. Model selection issues are mentioned and briefly illustrated in Section 5. The performance of the approach is illustrated in Section 6 and Section 7 concludes with a discussion and areas for further research.

2 A family of multivariate heavy-tailed distributions

Most of the work on multivariate scale mixture of Gaussians has focused on studying different choices for the weight distribution f_W (see *e.g.* Eltoft et al. [2006]). Surprisingly, little work to our knowledge has focused on the dimension of the weight variable W which in most cases has been considered as univariate. The difficulty in considering multiple weights is the interpretation of such a multidimensional case. The extension we propose consists then of introducing the parameterization of the scale matrix into $\boldsymbol{\Sigma} = \boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^T$, where $\boldsymbol{D}$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ is a diagonal matrix with the corresponding eigenvalues of $\boldsymbol{\Sigma}$. The matrix $\boldsymbol{D}$ determines the orientation of the Gaussian and $\boldsymbol{A}$ its shape. Such a parameterization has the advantage to allow an intuitive incorporation of the multiple weight parameters. We propose to set the scaled Gaussian part in (1) to $\mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)$, where $\boldsymbol{\Delta}_{\mathbf{w}} = \text{diag}(w_1^{-1}, \ldots, w_M^{-1})$ is the $M \times M$ diag-

onal matrix whose diagonal components are the inverse weights $\{w_1^{-1}, \ldots, w_M^{-1}\}$. When the weights are all one, a standard multivariate Gaussian case is recovered. The generalization we propose is therefore to define:

$$p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) = \int_0^\infty \ldots \int_0^\infty \mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}}\boldsymbol{A}\boldsymbol{D}^T)\, f_{\mathbf{w}}(w_1 \ldots w_M; \boldsymbol{\theta})\, \mathrm{d}w_1 \ldots dw_M \quad (3)$$

where $f_{\mathbf{w}}$ is now a M-variate density function to be further specified. In the following developments, we will consider only independent weights, *i.e.* with $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, $f_{\mathbf{w}}(w_1 \ldots w_M; \boldsymbol{\theta}) = f_{W_1}(w_1; \boldsymbol{\theta}_1) \ldots f_{W_M}(w_M; \boldsymbol{\theta}_M)$. We can use then one of the equivalent expressions below

$$\mathcal{N}_M(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_W\boldsymbol{A}\boldsymbol{D}^T) = \prod_{m=1}^M \mathcal{N}_1([\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m; 0, A_m w_m^{-1}) \quad (4)$$

$$= \prod_{m=1}^M A_m^{-1/2}\, \mathcal{N}_1(\frac{[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m}{A_m^{1/2}}; 0, w_m^{-1}) \quad (5)$$

$$= \prod_{m=1}^M \mathcal{N}_1([\boldsymbol{D}^T\mathbf{y}]_m; [\boldsymbol{D}^T\boldsymbol{\mu})]_m, A_m w_m^{-1})\ , \quad (6)$$

where $[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m$ denotes the mth component of vector $\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})$ and A_m the mth diagonal element of the diagonal matrix $\boldsymbol{A}$ (or equivalently the mth eigenvalue of $\boldsymbol{\Sigma}$). Using (4), it follows that

$$p(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\theta}) = \prod_{m=1}^M \int_0^\infty \mathcal{N}_1([\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m; 0, A_m w_m^{-1})\, f_{W_m}(w_m; \boldsymbol{\theta}_m)\, \mathrm{d}w_m\ . \quad (7)$$

The terms in the product reduce then to standard univariate scale mixtures. Another generative way to see this construction which is useful for simulation consists of simulating an M-dimensional Gaussian variable $\boldsymbol{X} = [X_1 \ldots X_M]^T$ with mean zero and covariance matrix equal to the identity matrix and to consider M independent positive variables $W_1, \ldots, W_M$ with respective distributions $f_{W_m}(w_m; \theta_m)$. Then the vector

$$\mathbf{Y} = \boldsymbol{\mu} + DA^{1/2}[X_1/\sqrt{W}_1, \ldots, X_M/\sqrt{W}_M]^T \quad (8)$$

follows one of the distributions below depending on the choice of f_{W_m}. For example, setting $f_{W_m}(w_m; \theta_m)$ to a Gamma distribution $\mathcal{G}(w_m; \alpha_m, \gamma_m)$ results in a multivariate generalization of a Pearson type VII distribution (see *e.g.* Johnson et al. [1994] vol.2 chap. 28 for a definition of the Pearson type VII distribution) while setting $f_{W_m}(w_m)$ to $\mathcal{G}(w_m; \nu_m/2, \nu_m/2)$ leads to a generalization of the multivariate t-distribution. Strictly speaking, to recover the Pearson type VII distribution of Johnson et al. [1994] which depends on two parameters m and c, we have to set $m = \alpha_m - 1/2$ and $c = \sqrt{2}\gamma_m$. This type of distribution is also referred to as the Arellano-Valle and Bolfarine's Generalized t distribution in Kotz and Nadarajah [2004] p. 94. In both cases, we can use (7) to express easily the respective densities denoted by $\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$ and $\mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ with $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_M\}$, $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}$ and $\boldsymbol{\gamma} = \{\gamma_1 \ldots \gamma_M\}$:

$$\mathcal{MS}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\nu}) = \prod_{m=1}^M \frac{\Gamma((\nu_m+1)/2)}{\Gamma(\nu_m/2)(A_m\nu_m\pi)^{1/2}} \left(1 + \frac{[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m^2}{A_m\nu_m}\right)^{-(\nu_m+1)/2} \quad (9)$$

Similarly,

$$\mathcal{MP}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\alpha},\boldsymbol{\gamma}) = \prod_{m=1}^{M} \frac{\Gamma(\alpha_m+1/2)}{\Gamma(\alpha_m)(2A_m\gamma_m\pi)^{1/2}}\left(1+\frac{[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m^2}{2A_m\gamma_m}\right)^{-(\alpha_m+1/2)} \quad (10)$$

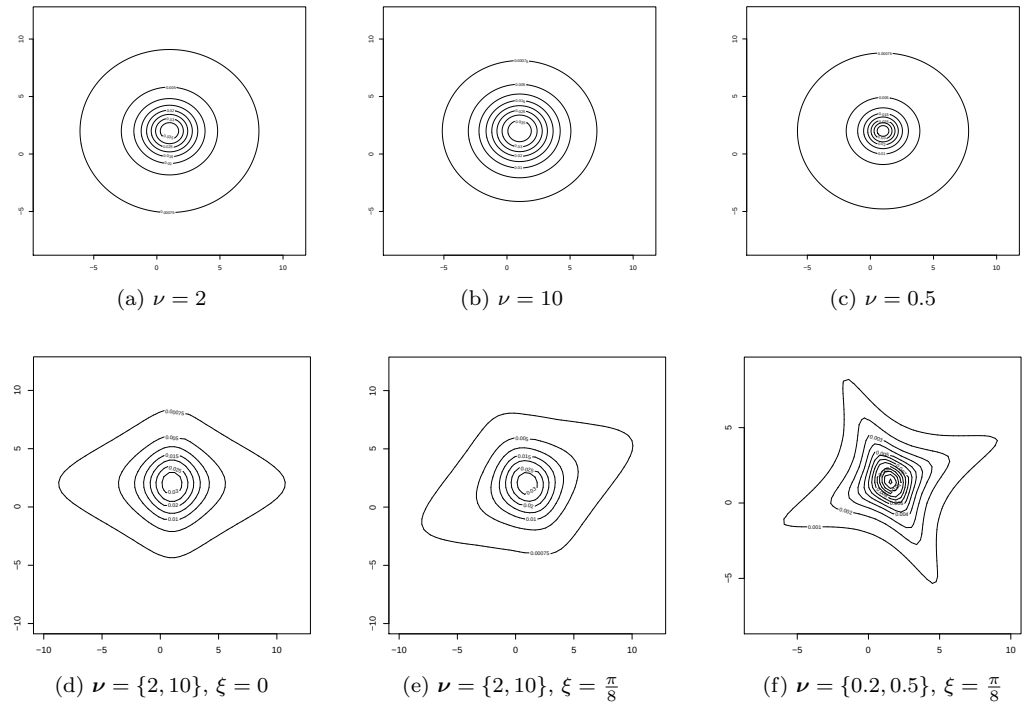
It is clear from (9) and (10) that these distributions are not of the elliptical form (see Fang et al. [2002]). This construction also allows a straightforward comparison with other generalizations based on a similar representation. For instance, Eltoft et al. [2006] consider the case of a single weight variable with an Inverse Gamma distribution $Inv\mathcal{G}(\alpha,\gamma)$ and define the so-called multivariate K model. We provide this generalization in Appendix A of the Supplementary Materials. Other well known distributions are the Normal Inverse Gaussian (NIG) distributions [Barndorff-Nielsen et al., 1982] which can model both skewness and heavy tails. When W_m^{-1} is assumed to follow an Inverse Gaussian distribution we recover the NIG distribution with the skewness parameter set to 0. To recover the more general NIG distribution we have to generalize equation (1) to both a scale and location mixture. Details are given in Appendix B of the Supplementary Materials.

We note that a slightly similar construction to (8) has been proposed in a graphical modelling context by Finegold and Drton [2011] for an alternative multivariate t-distribution by setting: $\mathbf{Y} = \boldsymbol{\mu} + [[\boldsymbol{D}\boldsymbol{A}^{1/2}\mathbf{X}]_1/\sqrt{W}_1, \ldots, [\boldsymbol{D}\boldsymbol{A}^{1/2}\mathbf{X}]_M/\sqrt{W}_M]^T$, which is different from (8) in that the term in $\boldsymbol{D}\boldsymbol{A}^{1/2}$ cannot be factorized. In particular, for this model the pdf is not available explicitly and we suspect it cannot be seen as a Gaussian scale mixture. Our proposal is also different from the *multivariate asymmetric t-distribution* proposed in Fang et al. [2002] which is constructed as a meta-elliptical distribution with given t-marginals with varying *dof*. As we will see in Section 3.2, our marginals are not in general t-distributions.

3 Some properties of the multiple scaled distributions

Using the definition of the multiple scaled t-distribution in (9), a number of different shapes are possible. In Figure 1, we show some of the different shapes in a two-dimensional setting for different values of $\boldsymbol{\nu}$ and $\boldsymbol{D}$, with $\boldsymbol{A}$ fixed to $\mathrm{diag}(4,4)$ and $\boldsymbol{\mu}$ to $[1,2]^T$ (Additional examples are shown in Figure 1 of the Supplementary Materials). In the bivariate case, we use for $\boldsymbol{D}$ a parameterization via an angle ξ so that $D_{11} = D_{22} = \cos\xi$ and $D_{21} = -D_{12} = \sin\xi$, where D_{md} denotes the (m,d) entry of matrix $\boldsymbol{D}$. The comparison with the Standard bivariate t-distribution is also illustrated. Figure 1 shows clearly the non-symmetric shape due to the multiple *dof* (bottom row), the possibility to introduce correlation (plots (e) and (f)) and an interesting star shape for low *dof* (plot (f)). With regards to the Pearson type VII distribution, similar shapes are observed but are not illustrated here.

Although the distributions are not elliptical, some symmetry can be observed. Equation (8) shows that vector $\boldsymbol{Y}-\boldsymbol{\mu}$ can be seen as a rotated version (by rotation matrix $\boldsymbol{D}$) of vector $\boldsymbol{A}^{1/2}[X_1/\sqrt{W}_1, \ldots, X_M/\sqrt{W}_M]^T$ whose components are independent and distributed according to 1D-distributions with symmetric tails. To get more asymmetry, the scale mixture has to be generalized to both location and scale mixture as illustrated for the NIG case in Appendix B of the Supplementary Materials.



(a) $\nu = 2$ (b) $\nu = 10$ (c) $\nu = 0.5$

(d) $\boldsymbol{\nu} = \{2, 10\}$, $\xi = 0$ (e) $\boldsymbol{\nu} = \{2, 10\}$, $\xi = \frac{\pi}{8}$ (f) $\boldsymbol{\nu} = \{0.2, 0.5\}$, $\xi = \frac{\pi}{8}$

Fig. 1 Contour plots of Bivariate t-distributions with $\boldsymbol{\mu} = [1, 2]^T$, $\boldsymbol{A} = \text{diag}(4, 4)$. First row: Standard Bivariate t-distributions. Second row: Multiple *dof* t-distributions.

3.1 Mean and covariance matrix

From equation (8), we can write $\tilde{\mathbf{X}} = [X_1/\sqrt{W}_1, \ldots X_M/\sqrt{W}_M]^T$. $\tilde{\mathbf{X}}$ is a vector of M independent variables $\tilde{X}_m$ whose distributions are given by
$\int_0^\infty \mathcal{N}(x_m \; ; \; 0, 1/w_m) \; f_{W_m}(w_m) \; dw_m$.
In the t-distribution and Pearson VII distribution cases, $\tilde{X}_m$ follows respectively a standard 1D t-distribution $\mathcal{S}(x_m; 0, 1, \nu_m)$ and a standard 1D Pearson VII distribution $\mathcal{P}(x_m; 0, 1, \alpha_m, \gamma_m)$. It follows then from representation (8) that $E[\mathbf{Y}] = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{A}^{1/2}E[\tilde{\mathbf{X}}]$ and $Var[\mathbf{Y}] = \boldsymbol{D}\boldsymbol{A}^{1/2}Var[\tilde{\mathbf{X}}]\boldsymbol{A}^{1/2}\boldsymbol{D}^T$. The expressions for the t-distribution and Pearson VII cases are given in Appendix C of the Supplementary Materials. Other quantities such as higher order moments can then also be derived easily from representation (8).

3.2 Marginals

Using (8), marginals are easy to sample from but computing their pdfs involves numerical integration. In general, 1D marginals correspond to linear combinations of the independent components of $\tilde{\boldsymbol{X}}$. For all $m = 1 \ldots M$, we have $Y_m = \mu_m + [\boldsymbol{D}\boldsymbol{A}^{1/2}\tilde{\boldsymbol{X}}]_m$. For instance, in the bivariate case with $\boldsymbol{\mu} = 0$, using the equivalent

parameterization of $\boldsymbol{\Sigma}$ into a diagonal matrix $\boldsymbol{A} = \text{diag}(A_1, A_2)$ and a matrix $\boldsymbol{D}$ parameterized via an angle ξ, it follows that $Y_1 = \sqrt{A_1}cos(\xi)\tilde{X}_1 - \sqrt{A_2}sin(\xi)\tilde{X}_2$ and $Y_2 = \sqrt{A_1}sin(\xi)\tilde{X}_1 + \sqrt{A_2}cos(\xi)\tilde{X}_2$.

In the t-distribution case, a 1D marginal is then a linear combination of standard 1D t-distributions for which in the general case no closed-form expression is available (see [Kotz and Nadarajah, 2004] for a review of different attempts in various particular cases). However an efficient algorithm to compute such pdfs can be derived according to Witkovský [2001]. The derivation in Witkovský [2001] is based on the inversion formula of the characteristic function which in the univariate case is:

$$f_Y(y) = \frac{1}{2\pi}\int_0^\infty (\exp(ity)\phi_Y(-t)+\exp(-ity)\phi_Y(t))dt = \frac{1}{\pi}\int_0^\infty Re(\exp(-ity)\phi_Y(t))dt \ ,$$

using the hermitian property of characteristic functions $\phi_Y(-t) = \overline{\phi_Y(t)}$ (the over line means the complex conjugate). The characteristic function of $Y_m = \mu_m + [\boldsymbol{D}\boldsymbol{A}^{1/2}\tilde{\boldsymbol{X}}]_m$ is $\phi_{Y_m}(t_m) = \exp(it_m\mu_m) \prod_{d=1}^{M} \phi_{\tilde{X}_d}(t_m[\boldsymbol{D}\boldsymbol{A}^{1/2}]_{md})$, where $[\boldsymbol{D}\boldsymbol{A}^{1/2}]_{md} = A_d^{1/2}D_{md}$ denotes entry (m,d) of matrix $\boldsymbol{D}\boldsymbol{A}^{1/2}$. In the Pearson VII case, $\phi_{\tilde{X}_d}$ is the characteristic function of a 1D distribution $\mathcal{P}(0,1,\alpha_d,\gamma_d)$. It can be shown as in [Witkovský, 2001] that

$$\forall t \in \mathbb{R},\ \phi_{\tilde{X}_d}(t) = \Gamma(\alpha_d)^{-1}2^{-\alpha_d+1}K_{\alpha_d}(\sqrt{2\gamma_d}|t|)(\sqrt{2\gamma_d}|t|)^{\alpha_d},$$

where $K_q(\ .\ )$ denotes the modified Bessel function of the second kind and order q. The t-distribution case follows easily by replacing α_d and γ_d by $\nu_d/2$.

We use the *tdist* R package of V. Witkovsky available at `http://aiolos.um.savba.sk/~viktor/software.html` to plot the pdf of some marginals and compare it with 1D t-distributions (see Figure 2). We also plot the histogram obtained by simulations from equation (8) to illustrate its consistency with the marginal pdf formula. The fact that the marginals are not in general t-distributions is a notable difference with other multivariate t generalizations.

For marginals of dimension greater than 1, we have to deal with linear combinations of the same 1D distributions which are therefore not independent. In this case, we can also easily derive the characteristic function and use a simple multidimensional inversion formula. Let $\mathcal{I}$ be a subset of $\{1,\ldots,M\}$ 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 $\phi_{\boldsymbol{Y}_\mathcal{I}}(\boldsymbol{t}_\mathcal{I}) = \prod_{m\in\mathcal{I}} \exp(it_m\mu_m) \prod_{d=1}^{M} \phi_{\tilde{X}_d}(\sum_{m\in\mathcal{I}} t_m[\boldsymbol{D}\boldsymbol{A}^{1/2}]_{md})$. It follows that the density of $\boldsymbol{Y}_\mathcal{I}$ via the multidimensional inversion formula (see *e.g.* Shephard [1991]) is: $f_{\boldsymbol{Y}_\mathcal{I}}(\boldsymbol{y}_\mathcal{I}) = (2\pi)^{-I} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp(-i\boldsymbol{t}_\mathcal{I}^T\boldsymbol{y}_\mathcal{I})\ \phi_{\boldsymbol{Y}_\mathcal{I}}(\boldsymbol{t}_\mathcal{I})\ d\boldsymbol{t}_\mathcal{I}$. When $I = 2$, and decomposing $\mathbb{R}^2$ into four quadrants,

$$f_{\boldsymbol{Y}_\mathcal{I}}(\boldsymbol{y}_\mathcal{I}) = 2\ (2\pi)^{-2} \int_0^\infty \int_{-\infty}^\infty Re(\exp(-i\boldsymbol{t}_\mathcal{I}^T\boldsymbol{y}_\mathcal{I})\ \phi_{\boldsymbol{Y}_\mathcal{I}}(\boldsymbol{t}_\mathcal{I}))\ d\boldsymbol{t}_\mathcal{I}.$$

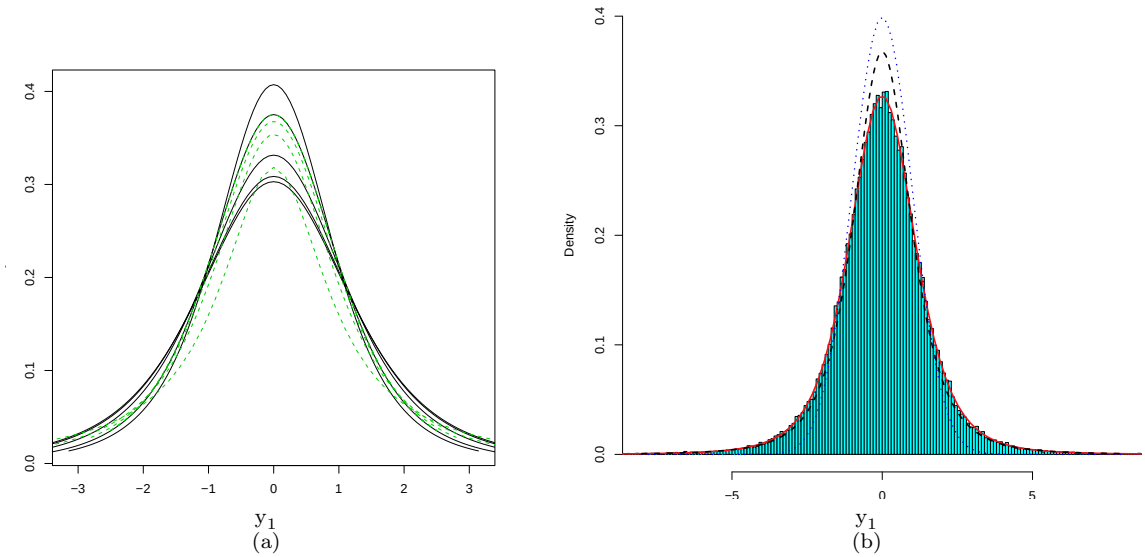


Fig. 2 1D marginals: (a) Y_1 distribution when $\boldsymbol{Y}$ is a bivariate multiple scaled bivariate t-distribution with $\boldsymbol{\mu} = 0$, $\nu_1 = \nu_2 = 3$, $\boldsymbol{A} = \text{diag}(1.5, 0.5)$. Black line curves show the marginal for different values of ξ from $\pi/8$ to $3\pi/8$ (increasing peaks). For comparison dashed line curves show the t-distribution for a *dof* varying from 1 to 4 (increasing peaks); (b) histogram of Y_1 when $\mathbf{Y}$ is a trivariate multiple scaled t-distribution with $\nu_1 = \nu_2 = \nu_3 = 3$ and $\boldsymbol{\Sigma}$ has its diagonal entries set to 1 and the others to 0.5, the Gaussian distribution with $\sigma^2 = 1$ is in blue (dotted line), the t-distribution with $\sigma^2 = 1$ and $\nu = 3$ in black (dashed lines) and the Y_1 distribution in red (solid lines).

This formula also generalizes easily in higher dimensions. For illustration, Figure 3 shows the bivariate marginal (Y_1, Y_2) of a 3 dimensional $\mathcal{MS}$ distributions with $\boldsymbol{\mu} = 0$, $\nu_1 = \nu_2 = \nu_3 = 3$, and $\boldsymbol{\Sigma}$ so that its diagonal entries are 1 and other entries are 0.5.

4 Maximum likelihood estimation of parameters

There are a few approaches to estimation of the standard t-distribution (see McLachlan and Peel [2000a] Section 7.5 and Kotz and Nadarajah [2004]). In this section, we outline for the multiple scaled Pearson VII distribution an EM approach to estimation of the parameters $\boldsymbol{\psi} = \{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$ with $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_M\}$ and $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_M\}$. The t-distribution case can be obtained straightforwardly unless specified otherwise. One difficulty here is to extend the approach to incorporate the decomposition of the scale matrix. The separate estimation of $\boldsymbol{D}$ and $\boldsymbol{A}$ is not straightforward and requires an additional minimization algorithm based on the Flury and Gautschi algorithm [Flury, 1984, Flury and Gautschi, 1986]. Similar difficulties are also encountered in Gaussian model-based clustering [Celeux and Govaert, 1995] for some of the proposed models.

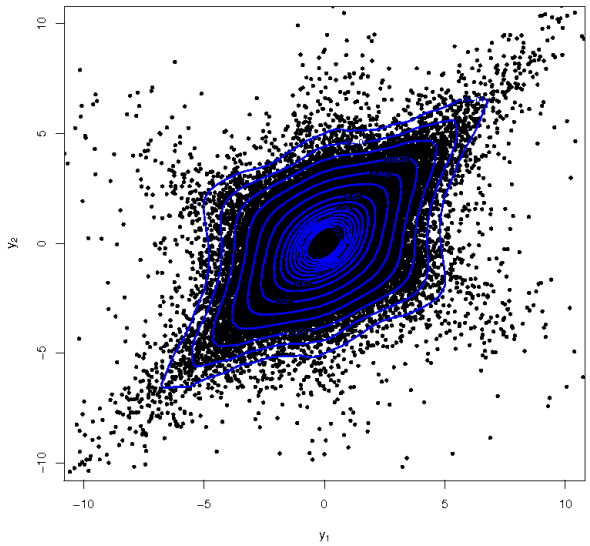


Fig. 3 (Y_1, Y_2) distribution when (Y_1, Y_2, Y_3) is a multiple scale trivariate t-distribution with $\boldsymbol{\mu} = 0$, $\nu_1 = \nu_2 = \nu_3 = 3$ and $\boldsymbol{\Sigma}$ so that its diagonal entries are 1 and other entries are 0.5. Contours are superimposed on points sampled from the distribution using equation (8).

Let us consider an *i.i.d* sample $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ of the multiple scaled Pearson VII distribution defined in (10). As in the standard t-distribution case, a convenient computational advantage of the EM approach is to view the weights as an additional missing variable $\mathbf{W}$. The observed data $\mathbf{y}$ are seen as being incomplete and additional missing weight variables $\mathbf{W}_1 \ldots \mathbf{W}_N$ with for $i \in \{1 \ldots N\}$, $\mathbf{W}_i = [W_{i1} \ldots W_{iM}]^T$ are introduced. These weights are defined so that: $\forall i \in \{1 \ldots N\}$, $\mathbf{Y}_i | \mathbf{W}_i = \mathbf{w}_i \sim \mathcal{N}_M(\boldsymbol{\mu}, \boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{A}\boldsymbol{D}^T)$ and $\mathbf{W}_i \sim \mathcal{G}(\alpha_1, \gamma_1) \otimes \ldots \otimes \mathcal{G}(\alpha_M, \gamma_M)$, where $\boldsymbol{\Delta}_{\mathbf{w}_i} = \text{diag}(w_{i1}^{-1}, \ldots, w_{iM}^{-1})$ and $\otimes$ means that the components of $\mathbf{W}_i$ are independent.

4.1 E step

At iteration (r) with $\boldsymbol{\psi}^{(r)}$ being the current parameter value, the E-step amounts to the computation, for all $i = 1 \ldots N$, of the missing variables posterior distribution $p(\mathbf{w}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(r)})$. The posterior $p(\mathbf{w}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(r)})$ is a product of Gamma distributions, $p(\mathbf{w}_i | \mathbf{y}_i; \boldsymbol{\psi}^{(r)}) = \prod_{m=1}^{M} \mathcal{G}(w_{im}\ ;\ \tilde{\alpha}_m^{(r)}, \tilde{\gamma}_{im}^{(r)})$, with

$$\tilde{\alpha}_m^{(r)} = \alpha_m^{(r)} + 1/2 \tag{11}$$

$$\tilde{\gamma}_{im}^{(r)} = \gamma_m^{(r)} + 1/2 A_m^{(r)-1} [\boldsymbol{D}^{(r)\ T}(\mathbf{y}_i - \boldsymbol{\mu}^{(r)})]_m^2 \ . \tag{12}$$

This is easily derived using the expression of $p(\mathbf{y}_i | \mathbf{w}_i, \boldsymbol{\psi}^{(r)})$ as given by equation (5) and remembering that the Gamma distribution is a conjugate prior for the

precision parameter (w_{im}) when the likelihood is Gaussian. Given the conjugacy, it follows that the posterior for $\mathbf{W}_i$ is of the same form as the prior, *i.e.* a product of Gamma distributions whose parameters, given above, can be deduced from standard Bayesian formula. Then, the expectation $E[W_{im}|\mathbf{y}_i;\boldsymbol{\psi}^{(r)}]$ denoted by $\bar{w}_{im}^{(r)}$ is given by:

$$\bar{w}_{im}^{(r)} = \frac{\tilde{\alpha}_m^{(r)}}{\tilde{\gamma}_{im}^{(r)}} = (\alpha_m^{(r)} + 1/2)(\gamma_m^{(r)} + 1/2A_m^{(r)-1}[\boldsymbol{D}^{(r)\,T}(\mathbf{y}_i - \boldsymbol{\mu}^{(r)})]_m^2)^{-1} \quad (13)$$

Also, the posterior expectation of $\log W_{im}$ follows easily, $E[\log W_{im}|\mathbf{y}_i;\boldsymbol{\psi}^{(r)}] = \Upsilon(\tilde{\alpha}_m^{(r)}) - \log\tilde{\gamma}_{im}^{(r)}$ where $\Upsilon(\,.\,)$ is the Digamma function.

The quantity $A_m^{(r)-1}[\boldsymbol{D}^{(r)\,T}(\mathbf{y}_i - \boldsymbol{\mu}^{(r)})]_m^2$ in equation (13) can also be interpreted as the squared Mahalanobis distance between $[\boldsymbol{D}^{(r)\,T}\mathbf{y}_i]_m$ and $[\boldsymbol{D}^{(r)\,T}\boldsymbol{\mu}^{(r)}]_m$ (when the variance is $A_m^{(r)}$), which is typically large for model outliers. For a given dimension m, expression (13) shows that the expected weight $\bar{w}_{im}^{(r)}$ at sample point i is lower when the distance is greater.

4.2 M step

For the updating of $\boldsymbol{\psi}$, the M-step consists of two independent steps for $(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})$ and $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ respectively. Details are given in Appendix F.1 of the Supplementary Materials. The optimization of these two steps leads to the following update equations.

Updating $\boldsymbol{\mu}$. Fixing $\boldsymbol{D}$ to the current estimation $\boldsymbol{D}^{(r)}$, leads to
$\boldsymbol{\mu}^{(r+1)} = (\sum_{i=1}^{N} \bar{\boldsymbol{\Delta}}_i^{(r)\,-1})^{-1} \sum_{i=1}^{N} \boldsymbol{D}^{(r)}\bar{\boldsymbol{\Delta}}_i^{(r)\,-1}\boldsymbol{D}^{(r)\,T}\mathbf{y}_i$,
where $\bar{\boldsymbol{\Delta}}_i^{(r)} = \mathrm{diag}(1/\bar{w}_{i1}^{(r)}, \ldots, 1/\bar{w}_{iM}^{(r)})$. Equivalently, for all $m = 1 \ldots M$, $\boldsymbol{\mu}_m^{(r+1)} = (\sum_{i=1}^{N} \bar{w}_{im}^{(r)})^{-1} \sum_{i=1}^{N} [\boldsymbol{D}^{(r)}\bar{\boldsymbol{\Delta}}_i^{(r)\,-1}\boldsymbol{D}^{(r)\,T}\mathbf{y}_i]_m$.

Updating $\boldsymbol{D}$. Using the equality $x^T\boldsymbol{S}x = \mathrm{trace}(\boldsymbol{S}xx^T)$ for any matrix $\boldsymbol{S}$, and defining the matrix $\boldsymbol{V}_i = (\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})^T$, it follows that for fixed $\boldsymbol{A}$ and $\boldsymbol{\mu}$, $\boldsymbol{D}$ is obtained by minimizing

$$\boldsymbol{D}^{(r+1)} = \arg\min_{\boldsymbol{D}} \sum_{i=1}^{N} \mathrm{trace}(\boldsymbol{D}(\bar{\boldsymbol{\Delta}}_i^{(r)}\boldsymbol{A})^{-1}\boldsymbol{D}^T\boldsymbol{V}_i).$$

Using current values $\boldsymbol{\mu}^{(r+1)}$ and $\boldsymbol{A}^{(r)}$, the parameter $\boldsymbol{D}$ can be updated using an algorithm derived from Flury and Gautschi (see Celeux and Govaert [1995]) which is outlined in the Appendix G of the Supplementary Materials.

Updating $\boldsymbol{A}$. $\boldsymbol{A}$ is updated as $\boldsymbol{A}^{(r+1)} = \arg\min_{\boldsymbol{A}} \mathrm{trace}(\sum_{i=1}^{N} \boldsymbol{M}_i^{(r)}\boldsymbol{A}^{-1}) + N\log|\boldsymbol{A}|$
where $\boldsymbol{M}_i^{(r)} = \bar{\boldsymbol{\Delta}}_i^{(r)\,-\frac{1}{2}}\boldsymbol{D}^T\boldsymbol{V}_i\boldsymbol{D}\bar{\boldsymbol{\Delta}}_i^{(r)\,-\frac{1}{2}}$ and $\sum_{i=1}^{N} \boldsymbol{M}_i^{(r)}$ is a symmetric positive definite matrix. So we can use Corollary A-2 in Celeux and Govaert [1995] with

$\boldsymbol{S} = \sum_{i=1}^{N} \boldsymbol{M}_i^{(r)}$. More details and the corollary are provided in Appendix F.2. Finally, by setting $\boldsymbol{D}$ and $\boldsymbol{\mu}$ to their current estimations $\boldsymbol{D}^{(r+1)}$ and $\boldsymbol{\mu}^{(r+1)}$,

$$\boldsymbol{A}^{(r+1)} = N^{-1}\text{diag}(\sum_{i=1}^{N} \bar{\boldsymbol{\Delta}}_i^{(r)\,-\frac{1}{2}} \boldsymbol{D}^{(r+1)T} \boldsymbol{V}_i^{(r)} \boldsymbol{D}^{(r+1)} \bar{\boldsymbol{\Delta}}_i^{(r)\,-\frac{1}{2}}) \,.$$

Equivalently, for all m, $A_m^{(r+1)} = N^{-1} \sum_{i=1}^{N} \bar{w}_{im}^{(r)} \, [\boldsymbol{D}^{(r+1)T}(\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})]_m^2 \,.$

Updating $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. The derivation is similar to the standard t-distribution case [McLachlan and Peel, 2000a]. The updated estimates $\alpha_m^{(r+1)}$ and $\gamma_m^{(r+1)}$ do not exist in closed form, but are given as solutions of the following equations: $\gamma_m = N\alpha_m(\sum_{i=1}^{N} \bar{w}_{im}^{(r)})^{-1}$ and $\log\left(\frac{N\alpha_m}{\sum_{i=1}^{N} \bar{w}_{im}^{(r)}}\right) - \Upsilon(\alpha_m) + \Upsilon(\tilde{\alpha}_m^{(r)}) - \frac{1}{N}\sum_{i=1}^{N} \log(\tilde{\gamma}_{im}^{(r)}) = 0$, where $\tilde{\alpha}_m^{(r)}$ and $\tilde{\gamma}_{im}^{(r)}$ are given in (11) and (12). In the t-distribution case, the ν_m's can be updated as the solution of the equation:

$$-\Upsilon(\frac{\nu_m}{2}) + \log(\frac{\nu_m}{2}) + 1 + \frac{1}{N}\sum_{i=1}^{N}(\log(\bar{w}_{im}^{(r)}) - \bar{w}_{im}^{(r)}) + \Upsilon(\frac{\nu_m^{(r)}+1}{2}) - \log(\frac{\nu_m^{(r)}+1}{2}) = 0.$$

In both cases, a solution can be found using a one-dimensional search such as Newton's method. In general, the updating of ν_m can be slow and alternative approaches are also possible [Shoham, 2002]. Alternatively, ν_m could be fixed a priori, and in this context is a form of M estimation [McLachlan and Peel, 2000a]. We note that for small sample sizes it may be also necessary to fix the value of ν_m.

Updating constrained $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. Similar updating equations can be easily derived when some on the parameters are assumed to be equal for several dimensions. We provide in Appendix F.3 of the Supplementary Materials the case where we assume that for all m, $\alpha_m = \alpha$ and $\gamma_m = \gamma$.

4.3 Mixture of multiple scaled distributions

The previous results can be extended to cover the case of K-component mixture of multiple scaled distributions. For multiple scaled Pearson VII distributions, with the usual notation for the proportions $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$ and $\boldsymbol{\psi}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}_k, \boldsymbol{\gamma}_k\}$ for $k = 1 \ldots K$, we consider,

$$p(\mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \pi_k \mathcal{MP}(\mathbf{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\alpha}_k, \boldsymbol{\gamma}_k)$$

where k indicates the kth component of the mixture and $\boldsymbol{\phi} = \{\boldsymbol{\pi}, \boldsymbol{\psi}\}$ with $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \ldots \boldsymbol{\psi}_K\}$ the mixture parameters. In the EM framework, an additional variable Z is introduced to identify the missing class labels, where $\{Z_1, \ldots, Z_N\}$ define the component of origin of the data $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$. In the light of the characterization of multiple scaled distributions, an equivalent modelling is: $\forall i \in$

$\{1 \dots N\}$, $\mathbf{Y}_i|\mathbf{W}_i = \mathbf{w}_i, Z_i = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{D}_k\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{A}_k\boldsymbol{D}_k^T)$ and $\mathbf{W}_i|Z_i = k \sim \mathcal{G}(\alpha_{1k}, \gamma_{1k}) \otimes \ldots \otimes \mathcal{G}(\alpha_{Mk}, \gamma_{Mk})$, where $\boldsymbol{\Delta}_{\mathbf{w}_i} = \text{diag}(w_{i1}^{-1}, \ldots, w_{iM}^{-1})$. Inference using the EM algorithm with two sets of missing variables $\boldsymbol{Z}$ and $\boldsymbol{W}$ to fit such mixtures, is similar to the individual ML estimation (see Appendix H of the Supplementary Materials).

5 Model selection issues

For practical purposes it is worth mentioning that in our framework, as the density is available in closed form, it is straightforward to consider information criteria based on penalized likelihood such a the Bayesian Information Criterion (BIC). In particular, one model selection issue of interest is related to the choice of the scale mixture distribution itself. As mentioned in the last paragraph of subsection 4.2, constraints on the *dof* parameters across dimensions can be easily accounted for and raise then the question of which model to fit to a given multivariate data set. In this section, we illustrate the use of BIC to decide between two models, the standard t-distribution or $\mathcal{MS}$ distribution and then on the number of different marginal tailweights (*i.e.* the number of free *dof* parameters in the $\mathcal{MS}$ case)). We consider three simple two-dimensional examples and three models, namely a standard t-distribution with a single *dof*, a $\mathcal{MS}$ distribution with two different *dof* and one with the same *dof* value in both dimensions. The three simulated data sets are then *i.i.d.* samples $\boldsymbol{y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ with $N = 500$. They are all simulated with $\boldsymbol{\mu} = [0, 0]^T$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ parameters but correspond to different model choices. The first data set is simulated from a standard t-distribution with $\nu = 3$, the second from a $\mathcal{MS}$ distribution with different *dof*'s $\boldsymbol{\nu} = \{1, 30\}$ and the third with equal *dof*'s $\boldsymbol{\nu} = \{3, 3\}$.

For an *i.i.d.* sample $\boldsymbol{y}$, the BIC for a multiple scaled mixture as defined in equation (7), is given by $BIC = -2\sum_{i=1}^{N} \log p(\boldsymbol{y}_i; \boldsymbol{\mu}^{ML}, \boldsymbol{\Sigma}^{ML}, \boldsymbol{\theta}^{ML}) + par \log N$, where the superscript ML indicates that the parameters are replaced by their maximum likelihood estimations and *par* is the number of free parameters in the model. If the dimension is M, $par = M(M+3)/2+1$ for the standard t-distribution and $par = M(M+3)/2+\tilde{M}$ for the $\mathcal{MS}$ with $\tilde{M}$ different *dof*'s, $\tilde{M}$ being between 1 (equal *dof*'s) and M (all different *dof*'s). We ran 100 simulations for each of the three models and computed the BIC for each simulation. In the three examples, the minimum BIC corresponds, as expected, to the model used to simulate the data for each of the 100 simulations. As an illustration, Table 1 shows the BIC values for one simulation. For a single *dof* parameter (first and third lines in the Table), the BIC values are close for the two $\mathcal{MS}$ cases. As one model is included in the other, the likelihood values are close and BIC tends to prefer not surprisingly the simplest model with less parameters (equal *dof*'s). Larger differences are seen for the multiple *dof* simulation (second line in Table 1) where the multiple *dof* model clearly outperforms the single *dof* ones.

Another model selection issue is related to the choice of the number of components in a mixture of $\mathcal{MS}$ distributions and is discussed in Section 6. Maximum likelihood estimation for such a mixture is provided in the Supplementary Materials and in the case of a K-component $\mathcal{MS}$ mixture, the *par* value becomes

Table 1 Two-dimensional (M=2) standard t and multiple scaled distributions simulations: the Bayesian Information Criterion (BIC) is used for selecting the appropriate model for the data. The BIC values for one simulation (over 100) of the three models are given. Bold characters indicate the minimum BIC. The same relative difference between the BIC values is consistently observed for all 100 simulations.

Simulated model	BIC values for different estimated models		
	t-distribution	MS with $\nu_1 \neq \nu_2$	$\mathcal{MS}$ with $\nu_1 = \nu_2$
t-distribution, $\nu = 3$	**3352.1**	3404.3	3398.2
$\mathcal{MS}$, $\boldsymbol{\nu} = \{1, 30\}$	4025.7	**3801.5**	3915.3
$\mathcal{MS}$, $\boldsymbol{\nu} = \{3, 3\}$	3540.0	3519.5	**3513.7**

$par = K - 1 + K(M(M+3)/2 + M)$. An illustration with more discussion is given in the real data example of the next Section.

6 Application to object detection using a stereoscopic camera pair

An important application of mixtures of heavy tailed distributions (and in particular t-distributions) is robust clustering. Prior to addressing the real data set illustrated in this section, we tested the increased flexibility and modelling capabilities provided by our model when applied to clustering of simulated data. We first considered simulated elongated clusters to illustrate the ability of our model to deal with various cluster shapes. Details are available in Appendix I.1 of the Supplementary Materials. For comparison, the results for the standard t-distribution reflect the difficulty the t-distribution faces in balancing the two very different tail behaviors. The classification results for the mixture of multiple scaled t-distributions model, referred to as MMST, are significantly better and indicate a close agreement with the data.

In a second simulated example, we considered a 10-dimensional problem previously analyzed by Cuesta-Albertos et al. [2008, Example 2] in the context of robust clustering. The example consists of 10-dimensional Gaussian clusters with concentrated outliers. The outlying data provides a good example to compare the robustness of the parameter estimates between the multiple scaled t-distribution and standard t-distribution. The parameter estimates for the MMST compared to the t-mixture are considerably less distorted by the outlying observations. Only the MMST is able to deal with a heavy tail in one of the directions or dimensions while the t-distribution is forced to, in some sense, provide an average across all dimensions. Details are available in Appendix I.2 of the Supplementary Materials.

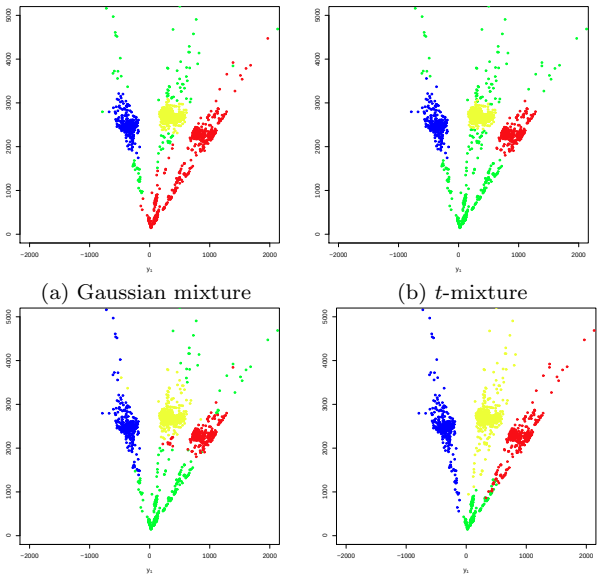
As the results of the EM algorithm can be particularly sensitive to initial values, we used a number of approaches to generate different initial values for parameters, including the use of random partitions, k-means and trimmed k-means [Cuesta-Albertos et al., 1997] with different amount of trimmed data (5%, 25% and 50%). Often the most successful strategy found was by estimating the $\boldsymbol{\mu}_k$'s and $\boldsymbol{\Sigma}_k$'s using the results from a trimmed k-means clustering with all $\nu_k = 20$. This value of the *dof* appears to be a good starting point that allows the subsequently estimated *dofs* to decrease to lower values if necessary in a reasonable number of iterations. For trimmed k-means, we used the *trimcluster* R package of C. Hennig available via the Comprehensive R-Archives Network CRAN of the R-project.

The computational speed of the EM algorithm is comparable to the standard t-distribution case with the exception that the update of $\boldsymbol{D}$ can be slow for high dimensional applications as the Flury and Gautschi algorithm involves sequentially updating every pair of column vectors of $\boldsymbol{D}$. A more global approach to the update of $\boldsymbol{D}$ has been proposed recently by Browne and McNicholas [2012] which has the potential to significantly speed up the computation time.

We then tested our model on a data set derived from the CAVA database `http://perception.inrialpes.fr/CAVA_Dataset/Site/`. The CAVA database [Arnaud et al., 2008, Khalidov, 2010] is a set of audiovisual recordings using binocular and binaural camera/microphone pairs gathered in order to test computational methods for audiovisual scene analysis (Figure 9 in the Supplementary Materials). In this paper, we are only interested in the visual part of the data set for it provides 3D data that show some interesting clustering characteristics to illustrate our approach. The 3D observations are shown in Figure 10 of the Supplementary Materials. The three observed elongated clusters correspond to three moving people in the original audiovisual recording. The fact that the three clusters join at the bottom of Figure 10 (Supplementary Materials, first two plots) is due to a larger number of mis-matched image features (artifacts) as we get closer to the camera pair. There is actually a rather large number of points near the camera pair and they cannot be considered as outliers in contrast to the previous examples. One of the goals is to recover from this data the locations in space of the main audio-visual objects (here the moving and speaking people) in the scene. We used for this data set different mixtures: Gaussian, standard t-distribution, and multiple scaled t-distribution mixtures. Although we know three people are actually present in the scene, we chose first to fit 4 clusters, the extra one being for the camera pair artifacts. The results are shown in Figure 4. There are a few different cluster representations possible for this data set depending on the distributional form assumed. For the MMST (Figure 4 (d)), the cluster representation consists of 3 distinct components for the objects with the relatively longer tails in the 3rd dimension for each component being captured as part of the 3 objects. A fourth component represents the visual source. By comparison, the cluster representation for the t-mixture (Figure 4 (b)) consists of only the mass for each of the objects represented as 3 distinct components with the fourth component (the visual source) capturing the rest of the data including the tails of the objects that were represented in the MMST case. For the Gaussian mixture (Figure 4 (a)), the mass of the objects are assigned to different components, but the background is a mix of different components, one component representing the visual source and an object, another component representing the tails of the other two objects. We can see part of the difficulty with this problem and some reason for the differences by examining the estimated degrees of freedom for the MMST (see details in Appendix I.3 of the Supplementary Materials).

In a second stage, we removed the artifacts near the camera pair by considering only the points such that $Y_3 > 1000$ (Figure 11 in the Supplementary Materials) and re-ran the clustering algorithms with $K = 3$. The resulting classifications shown in Figure 5 illustrate even more striking differences between the MMST and the other mixtures. Also shown are extra plots for dimensions 1 and 2. They all show that the MMST provides better defined groups. Without the possibility to form a fourth cluster with the points in the tails, the standard t-mixture cannot fully adapt to the shape of the clusters with one of them estimated as Gaussian

with a high *dof* of 107 for the middle yellow cluster. The other *dof*'s are estimated at 1.59 (furthermost right blue cluster) and 7.01 (furthermost left green cluster).



(a) Gaussian mixture (b) t-mixture

(c) t-mixture ν_k=2 (d) MMST

Fig. 4 Classification results with $K = 4$ for the audiovisual recording data. Classifications are shown in the first (x-axis) and third (y-axis) dimensions for (a) a Gaussian, (b) standard t, (c) standard t with all $\nu_k = 2$ mixtures and (d) for the MMST. The different colors indicate the 4 different components to which observations are assigned to.

Regarding the objects location, a simple estimator is the mean of the corresponding component. To assess the quality of this estimation, a ground truth is available from manual determination by an experimenter. Table 3 in the Supplementary Materials shows the location estimation (in cm) for each detected cluster using the MMST and t-mixture. The gain in precision provided by the MMST over the standard t-mixture is significant for the far right person who corresponds to the third right cluster for which the two models provide very different estimations (see Figures 5(b) and (d)). For the other coordinates, similar results are obtained for the t-distribution and the MMST with the t-mixture performing slightly better in two cases out of four.

Regarding clustering, it is also of interest to select automatically the number of objects in the scene. As already mentioned, the use of the Bayesian Information Criterion (BIC) is straightforward in our setting. We computed the criterion for $K = 2$ to $K = 8$ for three models: the MMST, the t-mixture and Gaussian mixture. The BIC values are shown in Figure 6 (a). For the MMST, the BIC values (blue plain line) are consistently lower than the values for the other models, the Gaussian case (black dot-dashed line) exhibiting the largest BIC values for all tested K. This is consistent with the better clustering result observed previously for the MMST with $K = 3$ and suggests the MMST would provide a better fit in any case. Regarding the selection of K, all BIC values decrease when K

increases suggesting $K = 8$ as the best choice for all three models. This is a typical behavior of BIC which is known to overestimate the number of clusters in case of model mis-specification. This also suggests that none of the compared models can actually model the data under consideration. However, in contrast to the others, the MMST case shows a local minimum at $K = 4$ suggesting that this value as a good candidate. The clustering obtained in this case is shown in Figure 6. The BIC preference for an additional fourth cluster to the three ideal ones (one per person) is clearly explained in Figure 6 (b) that emphasizes the existence of two rather separated groups within the points detected from the far left person in the scene.

7 Conclusion and future work

We have proposed a simple way to construct multivariate heavy-tailed distributions that can exhibit different marginal amount of tailweights. To our knowledge, this possibility has not previously been reached in a satisfactory manner. The various existing attempts generally suffer from either intractable probability density functions or from the difficulty to generalise to more than 2 dimensions. In contrast, our approach is applicable to high, potentially very high, dimensional spaces and with arbitrary correlation between dimensions. An important by-product of the availability of the density in closed form is that we can easily address model selection issues using information criteria based on penalised likelihoods. We have illustrated this advantage on typical model selection issues using the Bayesian Information Criterion. Estimation of the parameters of the new family is also relatively straightforward using the familiar EM algorithm and properties of the family are well defined with almost all found in analytical form.

Exploring the standard Gaussian scale mixture representation further, it follows that our construction can be used for a variety of distributions that can be seen as scale mixtures and more generally as location and scale mixtures such as the Multivariate Normal Inverse Gaussian (MNIG) distribution. Although out of the scope of this paper, a more complete study of this later extension would be of practical importance as it would allow the handling of skewed data [Karlis and Santourian, 2009]. The extension would thus provide a considerable degree of freedom in modelling data of varying tail behavior and directional shape.

Another interesting feature of the scale mixture representation is the introduction of multidimensional weight variables (W_m) that can be directly exploited in a supervised or informative prior context. For example, the Multivariate Pearson Type VII distribution we proposed could be used to generalize the work by Forbes et al. [2010] where a similar distribution is defined but with a diagonal scale matrix (no correlation between dimensions). In the work of Forbes et al. [2010], the emphasis is on the priors (f_{W_m}) on the weights which in this case are assumed to be in addition data point dependent (W_{im}) to guide the detection of small brain lesions from multimodal MRI data using prior (expert) information provided by neurologists. Our multivariate Pearson Type VII distribution is then a useful generalization of its t-distribution counterpart in that the weights are dependent on two parameters in the Gamma prior whose mean can be adjusted as desired in contrast to the t-distribution case. The advantage of multidimensional weights in this context offers a considerable benefit to account for data points (typically

brain lesion points) that are outliers in some dimensions (typically for some MR modalities) but inliers in others.

In the simpler context of the t-distribution, we provided an extension of this distribution (the $\mathcal{MS}$ distribution) that allows for the degrees of freedom parameters to be estimated differently in each dimension of the variable space. The key advantage of such an approach is the ability to avoid the compromise which can occur for a single degrees of freedom parameter in cases where the tail behaviors are very different in some dimensions.

Considering mixtures of $\mathcal{MS}$ distributions, we tested the approach on clustering examples using simulated and real data. The results suggested that the approach could significantly improve accuracy not only from an improved goodness of fit but also in terms of robustness to contamination by concentrated outliers.

For future research, parsimonious models could be considered using special decompositions of the covariance matrix such as in the model-based clustering approach of Celeux and Govaert [1995] and Fraley and Raftery [2002]. This has been done by Andrews and McNicholas [2012] for the standard t-distribution and would be straightforward to generalize to multiple scale distributions. Similarly, for very high dimensional data, other parsimonious models could also be considered with a special modelling of the covariance matrix such as in the High Dimensional Data Clustering (HDDC) framework of Bouveyron et al. [2007].

8 Supplementary material

Missing Appendices, Tables, and Figures are available in a companion supplemental file.

References

J. Andrews and P. McNicholas. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, pages 1–9, 2012. To appear.

C. Archambeau and M. Verleysen. Robust Bayesian clustering. *Neural Networks*, 20:129–138, 2007.

E. Arnaud, H. Christensen, Y-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant, F. Forbes, and R. Horaud. The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements. In *10th International Conference on Multimodal Interfaces, ICMI 2008, October, 2008*, pages 109–116, Chania, Crete, Grèce, October 2008. ACM.

A. Azzalini and M. G. Genton. Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76:106–129, 2008.

O. Barndorff-Nielsen, J. Kent, and M. Sorensen. Normal variance-mean mixtures and z Distributions. *International Statistics Review*, 50(2):145159, 1982.

C. M. Bishop and M. Svensen. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005.

C. Bouveyron, S. Girard, and C. Schmid. High dimensional data clustering. *Computational Statistics and Data Analysis*, 52:502–519, 2007.

R. Browne and P. McNicholas. Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, Published online, 2012.

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

J. A. Cuesta-Albertos, A. Gordaliza, and C. Matran. Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, 25(2):pp. 553–576, 1997.

J. A. Cuesta-Albertos, C. Matrn, and A. Mayo-Iscar. Robust estimation in the normal mixture model based on robust clustering. *Journal of the Royal Statistical Society Series B*, 70(4):779–802, 2008.

S. Daul, E. DeGiorgi, F. Lindskog, and A. J. McNeil. The grouped t-copula with an application to credit risk. *RISK*, 16:73, 2003.

S. Demarta and A. J. McNeil. The t copula and related copulas. *International Statistics Review*, 73:111, 2005.

T. Eltoft, T. Kim, and T-W. Lee. Multivariate Scale Mixture of Gaussians Modeling. In Justinian Rosca, Deniz Erdogmus, Jose Principe, and Simon Haykin, editors, *Independent Component Analysis and Blind Signal Separation*, volume 3889 of *Lecture Notes in Computer Science*, pages 799–806. Springer Berlin / Heidelberg, 2006.

H-B. Fang, K-T. Fang, and S. Kotz. The meta-elliptical distributions with given marginals. *Journal of Multivariate Analysis*, 82:1–16, July 2002.

M. Finegold and M. Drton. Robust graphical modeling of gene networks using classical and alternative T-distributions. *Annals of Applied Statistics*, 5(2A): 1057–1080, 2011.

B. N. Flury. Common Principal Components in K Groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984.

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

F. Forbes, S. Doyle, D. Garcia-Lorenzo, C. Barillot, and M. Dojat. A Weighted Multi-Sequence Markov Model For Brain Lesion Segmentation. In *13th International Conference on Artificial Intelligence and Statistics (AISTATS10)*, Sardinia, Italy, 13-15 May 2010.

C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97: 611–631, 2002.

R. Giordani, X. Mun, and R. Kohn. Flexible multivariate density estimation with marginal adaptation (extended version). *Unpublished working paper*, 2008.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, vol.2, 2nd edition*. John Wiley & Sons, New York, 1994.

M.C. Jones. A dependent bivariate t distribution with marginals on different degrees of freedom. *Statistics and Probability Letters*, 56(2):163–170, 2002.

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

V. Khalidov. *Conjugate Mixture Models for the Modelling of Visual and Auditory Perception*. PhD thesis, Grenoble University, October 2010.

S. Kotz and S. Nadarajah. *Multivariate t Distributions and their Applications*. Cambridge, 2004.

G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000a.

G.J. McLachlan and D. Peel. Robust mixture modelling using the t distribution. *Statistics and computing*, 10:339–348, 2000b.

S. Nadarajah and D. K. Dey. Multitude of multivariate T distributions. *Statistics*, 39:149, 2005.

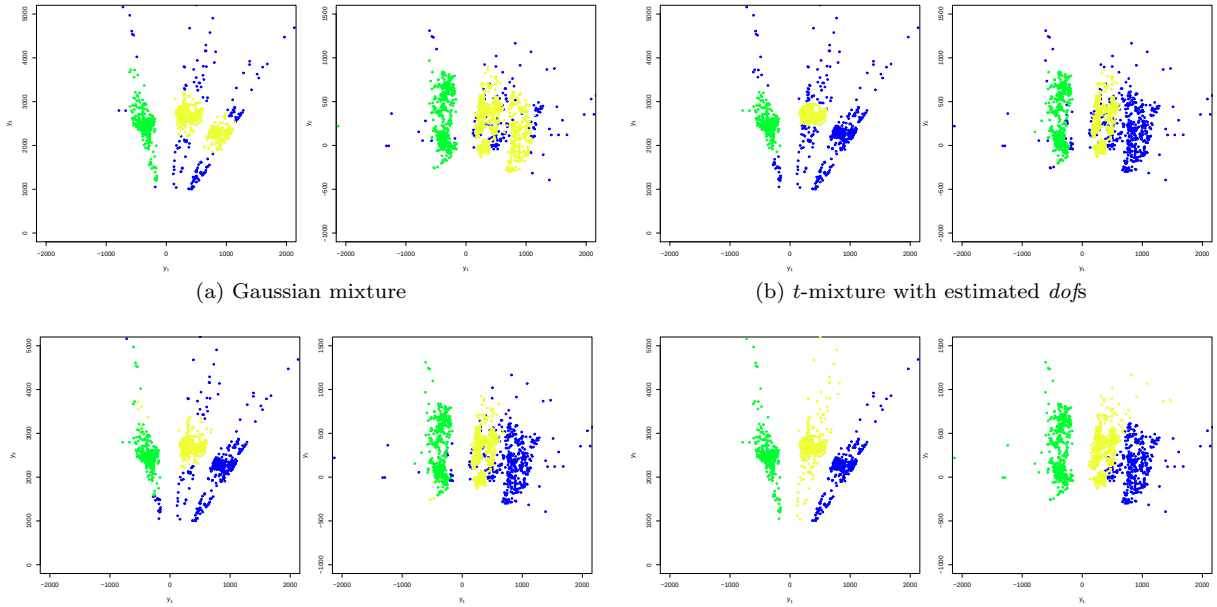
S. Nadarajah and S. Kotz. Multitude of bivariate T distributions. *Statistics*, 38: 527, 2004.

W. T. Shaw and K. T. A. Lee. Bivariate Student distributions with variable marginal degrees of freedom and independence. *Journal of Multivariate Analysis*, 99(6):1276–1287, 2008.

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

S. Shoham. Robust clustering by deterministic agglomeration EM of mixtures of multivariate t distributions. *Pattern Recognition*, 35(5):1127–1142, 2002.

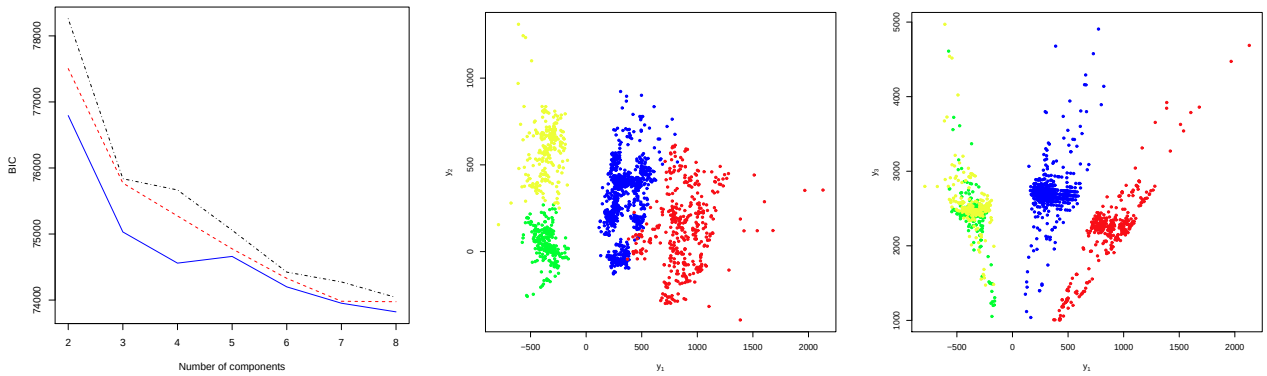
V. Witkovský. On the exact computation of the density and of the quantiles of linear combinations of t and F random variables. *Journal of Statistical Planning and Inference*, 94(1):1–13, 2001.



(a) Gaussian mixture (b) t-mixture with estimated dofs

(c) t-mixture with all dofs fixed to 2 (d) MMST

Fig. 5 Classification results with $K = 3$ for the audiovisual recording data after cutting off artifacts keeping points such that $Y_3 > 1000$. Classifications are shown in the first (x-axis) and third (y-axis) dimensions (left) and in the first (x-axis) and second (y-axis) dimensions (right) for (a) a Gaussian, (b) standard t, (c) standard t with all $\nu_k = 2$ mixtures and (d) for the MMST.



(a) BIC values for $K = 2$ to 8 (b) $K = 4$, (Y_1, Y_2) (c) $K = 4$, (Y_1, Y_3)

Fig. 6 Number of components (K) selection with BIC: (a) BIC values for $K = 2$ to $K = 8$ for the MMST (plain blue line), the t-mixture (dashed red line) and the Gaussian mixture (dot-dashed black line) models; (b) and (c) Clustering results in two subspaces with $K = 4$ for the MMST case.