# Supplementary Materials
for
A new family of multivariate heavy-tailed distributions
with variable marginal amounts of tailweight:
Application to robust clustering
by

Florence Forbes and Darren Wraith

Introduction

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 clustering examples for several simulated and real data sets.

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.

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 $\boldsymbol{\Delta}_\mathbf{w} = \mathrm{diag}(w_1^{-1},\dots,w_M^{-1})$ is the $M \times M$ diagonal matrix whose diagonal components are the inverse weights $\{w_1^{-1},\dots,w_M^{-1}\}$. When the weights are all one, a standard multivariate Gaussian case is recovered. The generalization we propose is therefore to define:

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) \;\; = \;\; \int_0^\infty \dots \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu},\boldsymbol{D}\boldsymbol{\Delta}_\mathbf{w}\boldsymbol{A}\boldsymbol{D}^T)\; f_\mathbf{w}(w_1 \dots w_M;\boldsymbol{\theta})\,\mathrm{d}w_1 \dots dw_M \tag{2}$$

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 $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_M\}$, $f_\mathbf{w}(w_1 \dots w_M;\boldsymbol{\theta}) = f_{W_1}(w_1;\boldsymbol{\theta}_1)\dots f_{W_M}(w_M;\boldsymbol{\theta}_M)$.

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

$$\mathbf{Y} \;\; = \;\; \boldsymbol{\mu} + DA^{1/2}[X_1/\sqrt{W}_1,\dots,X_M/\sqrt{W}_M]^T \tag{3}$$

follows one of the distributions below depending on the choice of f_{W_m}. For example, 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. Some illustrations in the bivariate case are given in Figure 1 for different parameters values. In particular, 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 next two sections provide two other examples.

Appendix A: A multivariate K distribution

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 can generalize the work of Eltoft et al. [2006] with a distribution denoted by $\mathcal{MK}$ by introducing multiple weights. For f_{W_m} taken as an Inverse Gamma distribution $Inv\mathcal{G}(w_m; \alpha_m, \gamma_m)$ in equation (2), it follows that:

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

where $K_q(\, . \,)$ denotes the modified Bessel function of the second kind and order q.

Appendix B: A multivariate NIG distribution

When W_m^{-1} in equation (2) 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 (variable Z below corresponds now to W^{-1}): $p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) = \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu} + z\boldsymbol{\Sigma}\boldsymbol{\beta}, z\boldsymbol{\Sigma}) \; f_Z(z;\boldsymbol{\theta}) \; \mathrm{d}z$, where $\boldsymbol{\beta}$ is an additional M-dimensional vector parameter for skewness. Using the decomposition of the scale matrix $\boldsymbol{\Sigma}$, our generalized multivariate pdf is then given by

$$p(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\theta}) = \int_0^\infty \ldots \int_0^\infty \mathcal{N}_M(\mathbf{y};\boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{\Delta}_\mathbf{z}\boldsymbol{A}\boldsymbol{D}^T\boldsymbol{\beta}, \boldsymbol{D}\boldsymbol{\Delta}_\mathbf{z}\boldsymbol{A}\boldsymbol{D}^T) \; f_\mathbf{z}(z_1 \ldots z_M;\boldsymbol{\theta}) \, \mathrm{d}z_1 \ldots dz_M \; ,$$

where $\boldsymbol{\Delta}_\mathbf{z} = \mathrm{diag}(z_1, \ldots z_M)$. Using expression (6) in the main paper, this can be equivalently written as

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

Using the parameterization in Karlis and Santourian [2009], if we set $f_{Z_m}(z_m)$ to an Inverse Gaussian distribution $IG(z_m; \gamma_m, \delta_m)$, it follows that our generalization of the multivariate NIG distribution with $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_M\}$ and $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_M\}$ is:

$$\mathcal{MNIG}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta}) = \prod_{m=1}^{M} \delta_m \exp(\delta_m\gamma_m + [\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m \; [\boldsymbol{D}^T\boldsymbol{\beta}]_m) \frac{\alpha_m}{\pi q_m} K_1(\alpha_m q_m) \qquad (4)$$

with $\alpha_m^2 = \gamma_m^2 + A_m[\boldsymbol{D}^T\boldsymbol{\beta}]_m^2$ and $q_m^2 = \delta_m^2 + A_m^{-1}[\boldsymbol{D}^T(\mathbf{y}-\boldsymbol{\mu})]_m^2$. An illustration of such a distribution is given in Figure 2. The difference between our generalized NIG as given in formula (4) and the standard multivariate NIG, where there are only single values of γ and δ [Karlis and Santourian, 2009], is illustrated also in Figure 2.

Appendix C: Mean and covariance matrix

Denoting $\tilde{\boldsymbol{X}} = [X_1/\sqrt{W}_1 \dots X_m/\sqrt{W}_M]^T$, the general expressions are:

$$E[\mathbf{Y}] = \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{A}^{1/2} E[\tilde{\mathbf{X}}] \text{ and } Var[\mathbf{Y}] = \boldsymbol{D}\boldsymbol{A}^{1/2} Var[\tilde{\mathbf{X}}]\boldsymbol{A}^{1/2}\boldsymbol{D}^T .$$

Therefore, the expectation exists and is equal to $\boldsymbol{\mu}$ if $\nu_m > 1$ in the t-distribution case and if $\alpha_m > 1/2$ for the Pearson VII case. The expectation is not defined otherwise. Similarly for the covariance matrix, in the t-distribution case, if all $\nu_m > 2$ then

$$Var[\mathbf{Y}] = \boldsymbol{D}\boldsymbol{A}^{1/2} \text{diag}\left(\nu_1/(\nu_1 - 2), \dots \nu_M/(\nu_M - 2)\right) \boldsymbol{A}^{1/2}\boldsymbol{D}^T .$$

While for the Pearson VII distribution, if all $\alpha_m > 1$, it follows that

$$Var[\mathbf{Y}] = \boldsymbol{D}\boldsymbol{A}^{1/2} \text{diag}\left(\gamma_1/(\alpha_1 - 1), \dots \gamma_M/(\alpha_M - 1)\right) \boldsymbol{A}^{1/2}\boldsymbol{D}^T .$$

Appendix D: Characteristic function

Denote by $\phi_{\mathbf{Y}}$ the characteristic function of a random vector $\mathbf{Y}$. It follows from (3) that, $\forall \mathbf{t} \in \mathbb{R}^M$, $\phi_{\mathbf{Y}}(\mathbf{t}) = E[\exp(i\mathbf{t}^T\mathbf{Y})] = \exp(i\mathbf{t}^T\boldsymbol{\mu}) \prod_{m=1}^{M} \phi_{\tilde{\mathbf{X}}_m}([\boldsymbol{A}^{1/2}\boldsymbol{D}^T\mathbf{t}]_m)$. In the Pearson VII case, $\phi_{\tilde{\mathbf{X}}_m}$ is the characteristic function of a 1D distribution $\mathcal{P}(0, 1, \alpha_m, \gamma_m)$. It can be shown as in [Witkovský, 2001] that $\forall t \in \mathbb{R}$, $\phi_{\tilde{X}_m}(t) = \Gamma(\alpha_m)^{-1} 2^{-\alpha_m+1} K_{\alpha_m}(\sqrt{2\gamma_m}|t|)(\sqrt{2\gamma_m}|t|)^{\alpha_m}$, where $K_q(\ . \ )$ denotes the modified Bessel function of the second kind and order q. The t-distribution case follows easily by replacing α_m and γ_m by $\nu_m/2$.

Appendix E: Local dependence function

When $M = 2$, the *local dependence function* $\mathcal{D}(y_1, y_2)$ of a bivariate distribution introduced by Holland and Wang [1987] is the mixed partial derivative of the log density. It is defined whenever the log density is a mixed differentiable function:

$$\mathcal{D}(y_1, y_2) = \frac{\partial^2 \log f(y_1, y_2)}{\partial y_1 \partial y_2} = \frac{1}{f(y_1, y_2)} \left[\frac{\partial^2 f(y_1, y_2)}{\partial y_1 \partial y_2} - \frac{\frac{\partial \log f(y_1, y_2)}{\partial y_1} \frac{\partial \log f(y_1, y_2)}{\partial y_2}}{f(y_1, y_2)} \right].$$

Holland and Wang [1987] introduced the local dependence function as a continuous analogue of the concept of local cross-product ratios for discrete variables. It has a number of properties and motivations (see also Jones [1996]). In particular Holland and Wang [1987] showed that, under some mild regularity conditions, any bivariate distribution may be specified by marginal distributions and the local dependence function. Jones [1996] also motivated the local dependence function from the point of view of localizing the Pearson correlation coefficient. For our bivariate $\mathcal{MP}$ when $\boldsymbol{\mu} = 0$ the local dependence function is:

$$\begin{aligned}
\mathcal{D}_{\mathcal{MP}}(y_1, y_2) &= -(2\alpha_1 + 1)D_{11}D_{21}\frac{2A_1\gamma_1 - (y_1D_{11} + y_2D_{21})^2}{(2A_1\gamma_1 + (y_1D_{11} + y_2D_{21})^2)^2} \\
&\quad -(2\alpha_2 + 1)D_{22}D_{12}\frac{2A_2\gamma_2 - (y_1D_{12} + y_2D_{22})^2}{(2A_2\gamma_2 + (y_1D_{12} + y_2D_{22})^2)^2} \qquad (5)
\end{aligned}$$

and simplifies with $\alpha_m = \gamma_m = \nu_m/2$ for the $\mathcal{MS}$ case. Figure 3 displays the local dependence functions for examples of the $\mathcal{MS}$ distribution. In the $\mathcal{MS}$ case, it is easily seen from (5) that when the *dofs* ν_1 and ν_2 tend to infinity, $\mathcal{D}(y_1, y_2)$ tends to a constant equal to $\rho(1-\rho^2)^{-1}(\Sigma_{11}\Sigma_{22})^{-1/2}$ where Σ_{11} and Σ_{22} denotes respectively the variance of Y_1 and Y_2 and ρ is the Pearson correlation coefficient. This constant value of $\mathcal{D}$ is the local dependence function of a bivariate Gaussian distribution whose covariance structure is defined by ρ, Σ_{11} and Σ_{22} and this is consistent with the fact that our $\mathcal{MS}$ distribution tends to a bivariate Gaussian when the *dof* parameters increase to infinity. Local dependence is positive in the positive and negative quadrants and negative in the other quadrants. This amounts to a positive association between absolute values of the marginal random variables. In particular, Figure 3 (b) illustrates the effect of one of the *dof* becoming large which implies part of the local dependence function to become flat.

Appendix F: Maximum likelihood estimation of parameters

F.1: Details on the M step

For the updating of $\boldsymbol{\psi} = \{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}, \boldsymbol{\alpha}, \boldsymbol{\gamma}\}$, the M-step consists of two independent steps for $(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})$ and $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ respectively:

$$\begin{aligned}
(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})^{(r+1)} &= \arg\max_{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}} \; E[\log p(\mathbf{y}|\mathbf{W}; \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})|\mathbf{y}, \boldsymbol{\psi}^{(r)}] \qquad (6) \\
&= \arg\max_{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}} \sum_{i=1}^{N} E[\log p(\mathbf{y}_i|\mathbf{W}_i; \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})|\mathbf{y}_i, \boldsymbol{\psi}^{(r)}] \qquad (7) \\
&= \arg\min_{\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}} \sum_{i=1}^{N} \Big((\mathbf{y}_i - \boldsymbol{\mu})^T \boldsymbol{D} \bar{\boldsymbol{\Delta}}_i^{(r)\,-1} \boldsymbol{A}^{-1} \boldsymbol{D}^T (\mathbf{y}_i - \boldsymbol{\mu}) + \log|\boldsymbol{A}| \Big) \;, \qquad (8)
\end{aligned}$$

where $\bar{\boldsymbol{\Delta}}_i^{(r)} = \text{diag}(1/\bar{w}_{i1}^{(r)}, \ldots, 1/\bar{w}_{iM}^{(r)})$. In the second line (7) above, $p(\mathbf{y}_i|\mathbf{W}_i; \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})$ is Gaussian with covariance $\boldsymbol{D}\boldsymbol{\Delta}_{\mathbf{w}_i}\boldsymbol{A}\boldsymbol{D}^T$ so that $\log p(\mathbf{y}_i|\mathbf{W}_i; \boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A})$ is linear in the W_{im}'s which after taking the expectation leads to a linear expression (8) in terms of the $\bar{w}_{im}^{(r)}$'s.

$$
\begin{aligned}
(\boldsymbol{\alpha},\boldsymbol{\gamma})^{(r+1)} &= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \; E[\log p(\mathbf{W};\boldsymbol{\alpha},\boldsymbol{\gamma})|\mathbf{y},\boldsymbol{\psi}^{(r)}] \qquad (9)\\
&= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \sum_{i=1}^{N}\sum_{m=1}^{M} E[\log p(W_{im};\alpha_m,\gamma_m))|\mathbf{y}_i,\boldsymbol{\psi}^{(r)}]\\
&= \arg\max_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \left(\sum_{m=1}^{M} N(\alpha_m\log\gamma_m - \log\Gamma(\alpha_m) + (\alpha_m-1)\Upsilon(\alpha_m^{(r)}+1/2))\right.\\
&\quad \left. -\sum_{m=1}^{M}\sum_{i=1}^{N}\left((\alpha_m-1)\log\left(\gamma_m^{(r)}+\frac{1}{2}\frac{[\boldsymbol{D}^{(r)\,T}(\mathbf{y}_i-\boldsymbol{\mu}^{(r)})]_m^2}{A_m^{(r)}}\right)+\gamma_m\bar{w}_{im}^{(r)}\right)\right)
\end{aligned}
$$

where $\Upsilon(\;.\;)$ is the Digamma function that verifies $E[\log W] = \Upsilon(\alpha) - \log\gamma$ when W follows a Gamma $\mathcal{G}(\alpha,\gamma)$ distribution. The Digamma function also satisfies $\frac{d\log\Gamma(\alpha)}{d\alpha} = \Gamma(\alpha)^{-1}\frac{d\Gamma(\alpha)}{d\alpha} = \Upsilon(\alpha)$.

F.2: Details on the updating of A

The updating of $\boldsymbol{A}$ uses the following corollary.

Corollary: *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 α is a positive real number is $\boldsymbol{A} = \frac{diag(\boldsymbol{S})}{\alpha}$.*

To apply it, we need to show that $\sum_{i=1}^{N}\boldsymbol{M}_i^{(r)}$ is a symmetric positive definite matrix. Indeed, omitting the (r) superscript, we have: $\boldsymbol{M}_i^T = \boldsymbol{\Delta}_i^{-\frac{1}{2}}\boldsymbol{D}^T\boldsymbol{V}_i\boldsymbol{D}\boldsymbol{\Delta}_i^{-\frac{1}{2}} = \boldsymbol{M}_i$ because $\boldsymbol{V}_i = (\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i-\boldsymbol{\mu})^T$ is symmetric and for all $\mathbf{x} > 0, \mathbf{x}^T\sum_{i=1}^{N}\boldsymbol{M}_i\mathbf{x} = \sum_{i=1}^{N}(\boldsymbol{\Delta}_i^{-\frac{1}{2}}\mathbf{x})^T\boldsymbol{D}^T\boldsymbol{V}_i\boldsymbol{D}(\boldsymbol{\Delta}_i^{-\frac{1}{2}}\mathbf{x}) > 0$ because $\boldsymbol{D}^T\boldsymbol{V}_i\boldsymbol{D}$ is positive definite.

F.3: Updating constrained $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

The updating equations can be easily derived when some on the parameters are assumed to be equal for several dimensions. For instance, if we assume that for all m, $\alpha_m = \alpha$ and $\gamma_m = \gamma$, $\alpha^{(r+1)}$ and $\gamma^{(r+1)}$ are solutions of :

$$
\log\left(\frac{NM\alpha}{\sum_{m=1}^{M}\sum_{i=1}^{N}\bar{w}_{im}^{(r)}}\right) - \Upsilon(\alpha) + \Upsilon(\alpha^{(r)}+\frac{1}{2}) - \frac{1}{NM}\sum_{m=1}^{M}\sum_{i=1}^{N}\log\left(\gamma^{(r)}+\frac{1}{2}\frac{[\boldsymbol{D}^{(r)\,T}(\mathbf{y}_i-\boldsymbol{\mu}^{(r)})]_m^2}{A_m^{(r)}}\right) = 0
$$

and $\quad \gamma = \dfrac{NM\alpha}{\sum_{m=1}^{M}\sum_{i=1}^{N}\bar{w}_{im}^{(r)}}$.

Similarly, in the t-distribution case, with $\nu_m = \nu$ for all m, ν can be updated as the solution of the equation:

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

Note that the latter equation is very close to the update equation for a standard t-distribution. The only difference resides in the term $\frac{1}{M}\sum_{m=1}^{M}(\log(\bar{w}_{im}^{(r)}) - \bar{w}_{im}^{(r)})$ which in the t-distribution case appears as an average of the weights across the dimensions. Recall however that the standard t-distribution is not included in the multiple scaled family. It is then easy to extend these equations to the case when only some of the *dof*'s are assumed to be equal.

Appendix G: Algorithm for computing $\boldsymbol{D}^{(r+1)}$

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

$$f(\boldsymbol{D}) \quad = \quad \arg\min_{\boldsymbol{D}} \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}(\bar{\Delta}_i^{(r)}\boldsymbol{A}^{(r)})^{-1}\boldsymbol{D}^T\boldsymbol{V}_i^{(r)} \ ,$$

with $\boldsymbol{V}_i^{(r)} = (\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})(\mathbf{y}_i - \boldsymbol{\mu}^{(r+1)})^T$. 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 $\boldsymbol{D}^0 = [\boldsymbol{d}_1^0, \dots, \boldsymbol{d}_M^0]$ where the $\boldsymbol{d}_m^0$'s are M-dimensional orthonormal vectors.
Step 2. For any couple $(l, m) \in \{1, \dots, M\}^2$ 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

$$\sum_{i=1}^{N}(\frac{\bar{\omega}_{il}^{(r)}}{A_l^{(r)}} - \frac{\bar{\omega}_{im}^{(r)}}{A_m^{(r)}})[d_l, d_m]^T\boldsymbol{V}_i^{(r)}[d_l, d_m] \ .$$

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

Appendix H: Mixtures of multiple scaled distributions

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 kth component of the mixture given the current estimates of the mixture parameters $\boldsymbol{\phi}^{(r)}$,

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)}\, \mathcal{MP}(\mathbf{y}_i; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{D}_k^{(r)}, \boldsymbol{A}_k^{(r)}, \boldsymbol{\alpha}_k^{(r)}, \boldsymbol{\gamma}_k^{(r)})}{p(\mathbf{y}; \boldsymbol{\phi}^{(r)})} \;.$$

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

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

with

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

and regrouping under C (resp. C') the terms not involving $\boldsymbol{\alpha}, \boldsymbol{\gamma}$ (resp. $\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}$),

$$\begin{aligned} Q_2(\boldsymbol{\alpha}, \boldsymbol{\gamma}; \boldsymbol{\phi}^{(r)}) \;&=\; C + \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik}^{(r)} \sum_{m=1}^{M} \Bigg(\alpha_{km} \log \gamma_{km} - \log \Gamma(\alpha_{km}) + (\alpha_{km} - 1) \Upsilon(\alpha_{km}^{(r)} + \frac{1}{2}) \\ &\quad -(\alpha_{km} - 1) \log \left(\gamma_{km}^{(r)} + \frac{1}{2} \frac{[\boldsymbol{D}_k^{(r)\,T} (\mathbf{y}_i - \boldsymbol{\mu}_k^{(r)})]_m^2}{A_{km}^{(r)}} \right) - \gamma_{km} \bar{w}_{kim}^{(r)} \Bigg) \end{aligned}$$

$$Q_3(\boldsymbol{\mu}, \boldsymbol{D}, \boldsymbol{A}; \boldsymbol{\phi}^{(r)}) = C' - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} \tau_{ik}^{(r)} \left((\mathbf{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{D}_k \bar{\boldsymbol{\Delta}}_{ki}^{(r)\,-1} \boldsymbol{A}_k^{-1} \boldsymbol{D}_k^T (\mathbf{y}_i - \boldsymbol{\mu}_k) + \log |\boldsymbol{A}_k| \right).$$

In the above, $\bar{\boldsymbol{\Delta}}_{ki}^{(r)} = \mathrm{diag}(1/\bar{\omega}_{ki1}^{(r)} \ldots 1/\bar{\omega}_{kiM}^{(r)})$, where $\bar{\omega}_{kim}^{(r)}$ is the expectation $E[W_{im} | Z_i = k, \mathbf{y}_i, \boldsymbol{\phi}^{(r)}]$ given by

$$\bar{\omega}_{kim}^{(r)} = \frac{\alpha_{km}^{(r)} + \frac{1}{2}}{\gamma_{km}^{(r)} + \frac{1}{2} \dfrac{[D_k^{(r)\,T} (\mathbf{y}_i - \mu_k^{(r)})]_m^2}{A_{km}^{(r)}}} \;.$$

M step

The sum $\sum\limits_{i=1}^{N} \tau_{ik}^{(r)}$ is denoted by $n_k^{(r)}$.

Updating the π_k's. The update of $\boldsymbol{\pi}$ is standard: for $k \in \{1 \ldots K\}$, $\pi_k^{(r+1)} = n_k^{(r)}/N$.

Updating the $\boldsymbol{\mu}_k$'s. It follows from expression 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

$$\mu_{km}^{(r+1)} \;=\; \frac{\sum\limits_{i=1}^{N} \tau_{ik}^{(r)}\; [\boldsymbol{D}_k^{(r)} \bar{\boldsymbol{\Delta}}_{ki}^{(r)\,-1} \boldsymbol{D}_k^{(r)\,T}\; \mathbf{y}_i]_m}{\sum\limits_{i=1}^{N} \tau_{ik}^{(r)} \bar{\omega}_{kim}^{(r)}} \;.$$

Updating the $\boldsymbol{D}_k$'s. Setting $\boldsymbol{V}_{ik}^{(r)} = \tau_{ik}^{(r)}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})^T$, it follows

$$\boldsymbol{D}_k^{(r+1)} = \arg\min_{\boldsymbol{D}_k} \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}_k(\bar{\boldsymbol{\Delta}}_{ki}^{(r)} \boldsymbol{A}_k^{(r)})^{-1} \boldsymbol{D}_k^T \boldsymbol{V}_{ik}^{(r)}) \ .$$

The parameter $\boldsymbol{D}_k$ can then be updated using the algorithm derived from Flury and Gautschi (see Appendix G).

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

$$\boldsymbol{A}_k^{(r+1)} \quad = \quad \arg\min_{\boldsymbol{A}_k} \sum_{i=1}^{N} \text{trace}(\boldsymbol{D}_k^{(r+1)}(\bar{\boldsymbol{\Delta}}_i^{(r)} \boldsymbol{A}_k)^{-1} \boldsymbol{D}_k^{(r+1)\ T} \boldsymbol{V}_{ik}^{(r)}) + \tau_{ik}^{(r)} \log |\boldsymbol{A}_k|$$

which leads to

$$A_{km}^{(r+1)} \quad = \quad \frac{\sum\limits_{i=1}^{N} \tau_{ik}^{(r)}\ \bar{\omega}_{kim}^{(r)}\ [\boldsymbol{D}_k^{T(r+1)}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r+1)})]_m^2}{n_k^{(r)}} \ .$$

Updating the $\boldsymbol{\alpha}_k$'s and $\boldsymbol{\gamma}_k$'s. As in the individual ML estimation, the estimates do not exist in closed form, but are given as a solution of the equations below:

$$\log \left(\frac{n_k^{(r)} \alpha_{km}}{\sum\limits_{i=1}^{N} \tau_{ik}^{(r)} \bar{w}_{kim}^{(r)}} \right) - \Upsilon(\alpha_{km}) + \Upsilon(\alpha_{km}^{(r)} + \frac{1}{2}) - \frac{1}{n_k^{(r)}} \sum_{i=1}^{N} \tau_{ik}^{(r)} \log \left(\gamma_{km}^{(r)} + \frac{1}{2} \frac{[\boldsymbol{D}_k^{(r)\ T}(\mathbf{y}_i - \boldsymbol{\mu}_k^{(r)})]_m^2}{A_{km}^{(r)}} \right) = 0$$

$$\text{and} \qquad \gamma_{km} = \frac{n_k^{(r)} \alpha_{km}}{\sum\limits_{i=1}^{N} \tau_{ik}^{(r)} \bar{w}_{kim}^{(r)}}.$$

In the t-distribution case, the ν_{km}'s can be updated as the solution of the following equation instead:

$$-\Upsilon(\frac{1}{2}\nu_{km}) + \log(\frac{1}{2}\nu_{km}) + 1 + \frac{1}{n_k^{(r)}} \sum_{i=1}^{N} \tau_{ik}^{(r)} (\log(\bar{w}_{kim}^{(r)}) - \bar{w}_{kim}^{(r)}) + \Upsilon(\frac{\nu_{km}^{(r)} + 1}{2}) - \log(\frac{\nu_{km}^{(r)} + 1}{2}) = 0 \ .$$

Appendix I: Application to clustering

An important application of mixtures of heavy tailed distributions (and in particular t-distributions) is robust clustering. In this section, we illustrate the increased flexibility and modelling capabilities provided by our multiple scaled t-distribution model when applied to clustering.

I.1: Simulated data, Elongated clusters

In a first simulated example, we illustrate the ability of our model to deal with various cluster shapes. In particular, the mixture of multiple scaled t-distributions model, referred to as MMST, is able to recover correctly elongated clusters with a significant amount of data points in one of the clusters' tails (Figure 4 (d)). We consider the case of three clusters slightly separated from one another which are generated from the product of two univariate t-distributions with different degrees of freedom (see Figure 4 (a)). For $k = 1, \dots, 3$, in the first dimension $\nu_{k1} = 1$ and in the second $\nu_{k2} = 30$ (closer to Gaussian). The means for the three clusters are respectively $[-4, 0]^T$ (blue), $[0, 0]^T$ (green) and $[0, -4]^T$ (red), with the same scale matrix equal to the identity matrix. The sample size is 4500 with 1500 observations in each cluster. The data is rotated by 45 degrees to provide a test in terms of finding the correct orientation.

Table 1 provides the classification results for the different models, MMST, mixtures of t-distributions and Gaussian distributions. To ensure that the global maximum was found, several different initial parameter values were used (except for the dof's that were always all initialized to 20) using k-means and various thresholds with trimmed k-means [Cuesta-Albertos et al., 1997]. Convergence of the EM algorithm was monitored using Aitken's acceleration [McLachlan and Krishnan, 2008, Chap. 4.9]. For all approaches considered we report the Brier score [Brier, 1950] and Dice score [Dice, 1945], the former providing a measure which incorporates the uncertainty of the classification while the latter is a more standard measure. The Dice score measures the overlap between a classification result and the ground truth. Denoting by TP the number of true positives, FP the number of false positives and FN the number of false negatives, the Dice score is given by $\frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}}$. The range for the Brier score is between 0 (perfect classification) and 1, with a lower value indicating a better classification. The Dice score also ranges between 0 and 1 but here 1 represents the best classification and a lower value a worse classification.

A graphical display of the classifications for the mixture of t-distributions and for the MMST is provided in Figure 4. For the t-distribution mixture, only the classifications in the fixed dof's cases are shown. The classification obtained for estimated dof's is similar to when the dof's are fixed to 1.

As we can see from Table 1 and Figure 4 the results for the standard t-distribution (overall Brier score 0.127) reflect the difficulty the t-distribution faces in balancing the two very different tail behaviors. The estimated dofs for the t-distributions are respectively 1, 3.79 and 3.00 expressing a preference for a heavy tailed solution. The classification results for the MMST are significantly better and indicate a close agreement with the data (overall Brier score 0.019). The estimated dofs are $\{1.02, 299\}$ for component 1 (blue), $\{1.26, 299\}$ for component 2 (green) and $\{1.16, 299\}$ for component 3 (red) with the disparity between the two dimensions appearing to be well estimated from the data.

Regarding the estimation of the dof's, we show for illustration in Figure 5, the profile log-likelihoods in terms of the dof parameter for the upper most (blue) component in our 2-dimensional example. The profile log-likelihood is the log-likelihood seen as a function of one of the parameters (here the dof), the others being fixed to their optimal values. In

Table 1: Classification results for the 2D elongated clusters (Figure 4). $\mathcal{L}$ denotes the log-likelihood and class$\mathcal{L}$ the classification log-likelihood (see *e.g.* Celeux and Soromenho [1996]).

Mixture Model	Brier Score				Dice Score	$\mathcal{L}$	class$\mathcal{L}$
	Comp 1	Comp 2	Comp 3	Overall	Overall		
Gaussian mixture	0.036	0.345	0.334	0.715	0.512	-31023	-32466
t-mixt. estimated dof	0.063	0.031	0.033	0.127	0.743	-23454	-23745
t-mixt. $\nu_1 = \nu_2 = \nu_3 = 1$	0.020	0.033	0.021	0.074	0.805	-23605	-24165
t-mixt. $\nu_1 = \nu_2 = \nu_3 = 30$	0.118	0.063	0.055	0.237	0.609	-26064	-26312
MMST, estimated dof	**0.004**	**0.009**	**0.006**	**0.019**	**0.950**	**-22948**	**-23015**

practice, the parameter values for the MMST are based on the estimated values. For the standard t-mixture, the same parameter values are taken so that the profile log-likelihoods can be compared. The solution of the standard t-mixture is to take a low dof and this is one of the solutions that is possible (and close to the estimated solution). The maximum for one of the dimensions of the MMST corresponds to a dof greater than 30 (the plot was truncated at this value for visualisation purposes so that we can see the lower dof's more easily as in practice the estimated dof was around 200). Interestingly the likelihood looks quite flat after about a dof of 9. The log-likelihood has been standardised to make it easier to compare (*i.e* |log-likelihood|/max(log-likelihood)).

We note that it is possible to find a solution for the Gaussian case (and similarly for the t-mixture with high dof's fixed to 300) which provides a good classification of the data with a Dice score of 0.91. However this solution is suboptimal with a much lower log-likelihood (and classification log-likelihood) than the results reported in Table 1.

The above results suggest therefore quite a large difference in the obtained classifications for data displaying very different tail behaviors between dimensions. For illustration purposes, the simulated data in this section shows quite a large difference between the two tails. If we simulate other data sets by varying the degrees of freedom parameter in the first dimension, and keep the degrees of freedom for the second dimension the same (30), we find that for a degrees of freedom parameter equal to 4 (in the first dimension) the clustering results between the t-mixture and the MMST start to become similar. This result is due more to the components being rather well separated so that for even such small changes in degrees of freedom, we cannot see a clear difference in terms of classification. This highlights the fact that the quality of the classification alone is not a fully informative criterion to assess the performance of our model. For this reason, in the next section, we focus also on parameter estimation to illustrate the gain of MMST in terms of goodness-of-fit.

I.2: Simulated data, 10-dimensional Gaussian clusters with concentrated outliers

For our second simulated example, we consider a 10-dimensional problem previously analyzed by Cuesta-Albertos et al. [2008, Example 2] in the context of robust clustering. The mixture consists of the product measure of a standard 8-dimensional Gaussian distribution with zero mean and covariance matrix equal to eight times the identity matrix with a mixture of three bivariate Gaussian distributions with parameters $\pi_k = \frac{1}{3}, k = 1, \dots, 3, \boldsymbol{\mu}_1 = [-9, 0]^T, \boldsymbol{\mu}_2 = [1, 5]^T, \boldsymbol{\mu}_3 = [3.5, -3.5]^T$ and $\boldsymbol{\Sigma}_1 = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 8.5 & -7.5 \\ -7.5 & 8.5 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. The sample size is 600 and is slightly contaminated by 10 additional data sampled from a uniform distribution on some parallelepiped. We consider two cases corresponding to two different contaminations. The first case is the example of Cuesta-Albertos et al. [2008] where the parallelepiped is $[-4, 4]^8 \times [6, 10] \times [11, 19]$ (see Figure 6 (a)). In the second case, the parallelepiped is $[-4, 4]^8 \times [-15, -10] \times [20, 25]$ (see Figure 6 (b)). The last two dimensions of the simulated data sets are shown in Figure 6 with the colors blue, green and yellow indicating the three main components and the outliers shown in black in the upper right and left hand corners of the corresponding plot.

The outlying data provides a good example to compare the robustness of the parameter estimates (of the 3 main component distributions) between the multiple scaled t-distribution and standard t-distribution. As the 10 additional data are outliers in only two of the 10 dimensions, we expect the multiple scaled t-distributions to downweight these sample points more via a lower estimated degrees of freedom parameter in either or both of these dimensions (ν_{k9} and ν_{k10}). In contrast, the single estimated degrees of freedom parameter (ν) for the t-distribution is forced to provide an average (in some sense) across all dimensions. In the first case (case a), we observe a slight difference in the classifications obtained with the t-mixture and MMST. The results for both models are quite good and close to the true classification but the MMST shows consistently better Dice and Brier scores (see Table 2). As seen in Figure 7, the t-mixture tends to overestimate the first component (centered at $\boldsymbol{\mu}_1 = [-9, 0]^T$ in blue) and not to capture correctly the border with the second component (centered at $\mu_2 = [1, 5]^T$ in green). In the second case (case b), the classification results for the t-mixture and the MMST are the same and correspond to a classification very close to the true classification (see the Dice and Brier score in Table 2). At issue, though, in both these cases is the degree to which the parameter estimates of at least one of the components is influenced by the outlying observations. For instance, in case b, depending on the initial values given (trimmed k-means with 10% of the points excluded or k-means) the outlying observations are either allocated to component 1 ($\boldsymbol{\mu}_1 = [-9, 0]^T$ (blue)) or component 2 ($\boldsymbol{\mu}_2 = [1, 5]^T$ (green)) and the parameter estimates of these components (notably $\boldsymbol{\mu}_k$) can be affected by the outlying observations. To see the influence of the outlying observations we calculated the absolute and mean squared errors of the true mean parameter values to the parameter estimates in the t-mixture and MMST cases over 30 repeated simulations of the data set. Because the outliers are in dimensions 8 and 9, we compute the absolute and mean square errors both in these two dimensions and in the 8 first ones separately. To allow for sampling variability, the true parameter values are replaced by the parameter estimates obtained by fitting a Gaussian mixture to the same data set without the outlying observations. The results using trimmed k-means to get initial values are reported in Table 2 and the corresponding boxplots are shown in Figure 8. The parameter estimates for the MMST compared to the t-mixture are considerably less distorted by the outlying observations in dimensions 9 and 10. This is consistent with the fact that in case a (resp. case b) the MMST estimated degrees of freedom parameters in the $9th$ and $10th$ dimensions for component 1 are approximately 5

Table 2: Dice and Brier scores and estimation errors for the mean parameters (all three components together) over 30 simulations of the data. For both the t-mixture and MMST, median values of the errors are shown with the range in brackets. Case a: outliers in far right corner. Case b: outliers in far left corner.

	Absolute Error		Squared Error		Classification	
	Dim 1 to 8	Dim 9 & 10	Dim 1 to 8	Dim 9 & 10	Dice Score	Brier Score
t-mixt.	1.12	3.84	0.11	5.21	0.89	0.05
Case a	(0.72,1.90)	(2.94,4.35)	(0.04,0.30)	(3.94,7.00)	(0.80,0.94)	(0.03,0.11)
MMST	**0.77**	**0.97**	**0.04**	**0.33**	**0.97**	**0.02**
Case a	(0.47,1.20)	(0.62,1.54)	(0.01,0.09)	(0.14,0.78)	(0.92,1.00)	(0.01,0.03)
t-mixt.	**0.44**	1.44	**0.02**	1.29	**0.98**	**0.01**
Case b	(0.22,0.92)	(0.97,1.70)	(0.00,0.07)	(0.34,1.73)	(0.94,0.99)	(0.01,0.02)
MMST	0.65	**0.40**	0.03	**0.06**	**0.98**	**0.01**
Case b	(0.44,0.98)	(0.12,0.77)	(0.01,0.08)	(0.00,0.29)	(0.95,1.00)	(0.00,0.02)

and over 60 (resp. 9 and over 60), compared to an overall degrees of freedom parameter over 60 (resp. over 300) for the t-mixture. Thus, in both cases, 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.

I.3: Object detection using a stereoscopic camera pair

We then tested our model on a data set derived from the CAVA database `http://perception.inrialpes.fr/CAVA_Dataset/Site/`. The CAVA database 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 this work, 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. They represent locations in space that have been reconstructed from a stereo camera pair using some stereo-motion reconstruction method (see Arnaud et al. [2008], Khalidov [2010], Khalidov et al. [2011] for more details) based on so called interest points detected in the right and left images. 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 in the main paper.

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. For the components to the left and right, the estimated degrees of freedom are low for the second and third dimensions (approx. 2 and 5 respectively for both components), and high for the first dimension (approx. 200). For the middle component, the estimated degrees of freedom parameter for the second dimension is low (approx. 5), and high for the first and third dimension. The estimated degrees of freedom for the component representing the visual source are all high (approx. 100). Thus, for the MMST representation there is a considerable amount of difference in the estimated degrees of freedom across dimensions for most of the components (3 out of

4). For the t-mixture, the estimated degrees of freedom parameter is low (approx. 1) for the component representing the visual source and background. The estimated degrees of freedom for the middle and right components are high (approx. 100), and low for the left (approx. 6). When the degrees of freedom are fixed to a low value (2) we also see a similar result (Figure 4 (c) in the paper).

In a second stage, we removed the artifacts near the camera pair by considering only the points such that $Y_3 > 1000$ (Figure 11) and re-ran the clustering algorithms with $K = 3$.

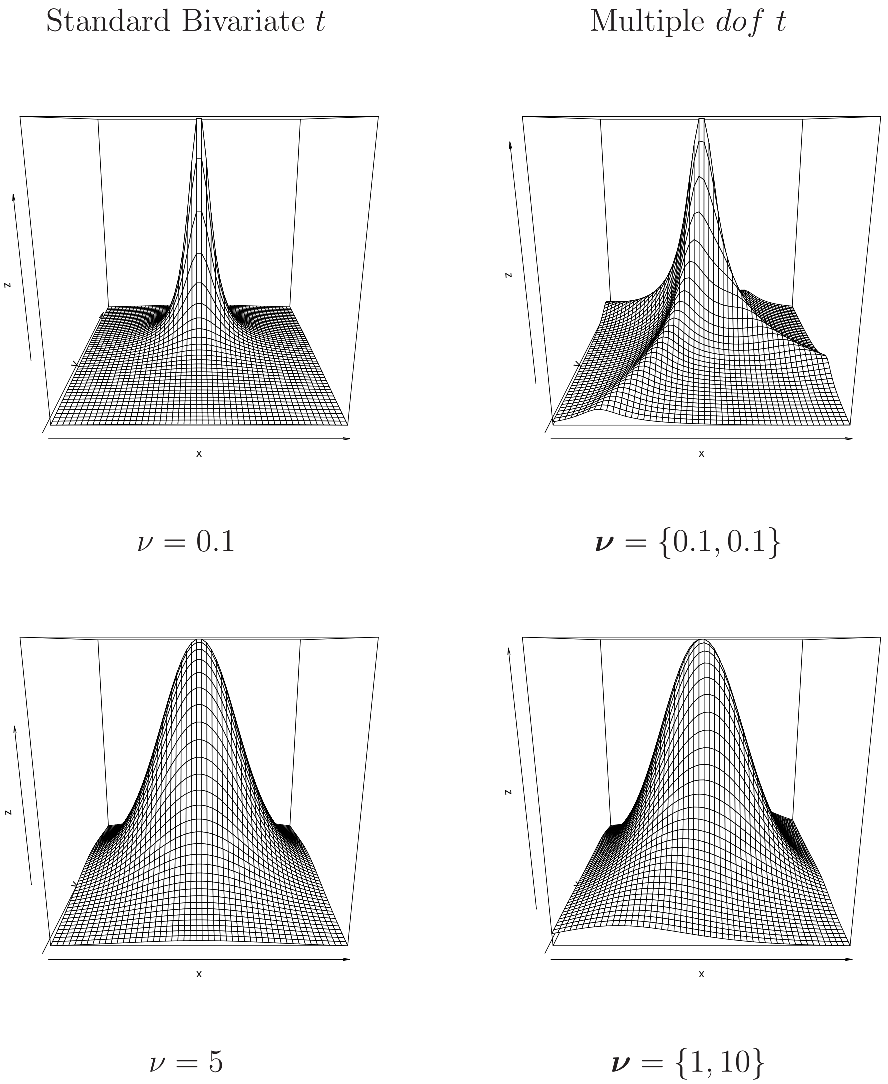
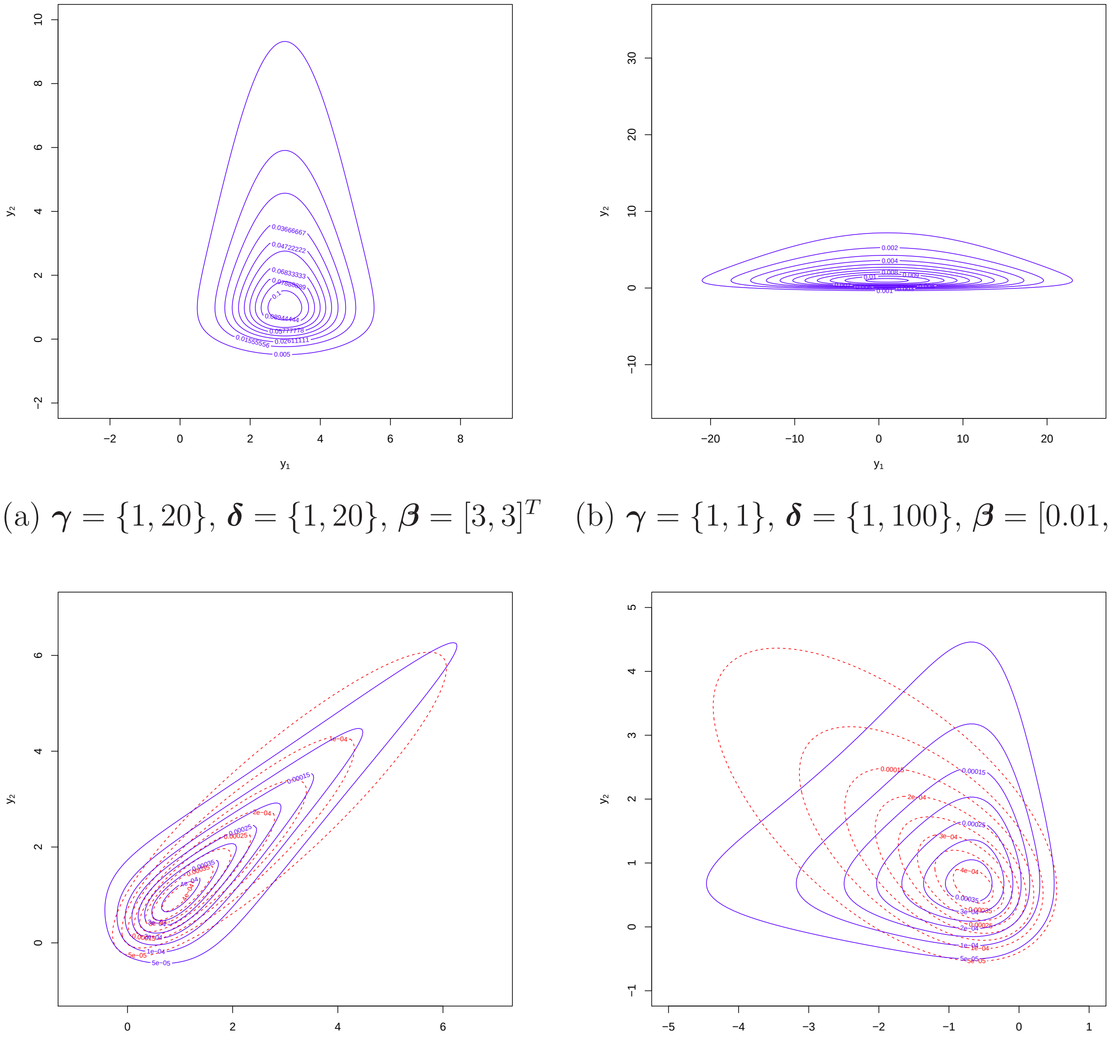


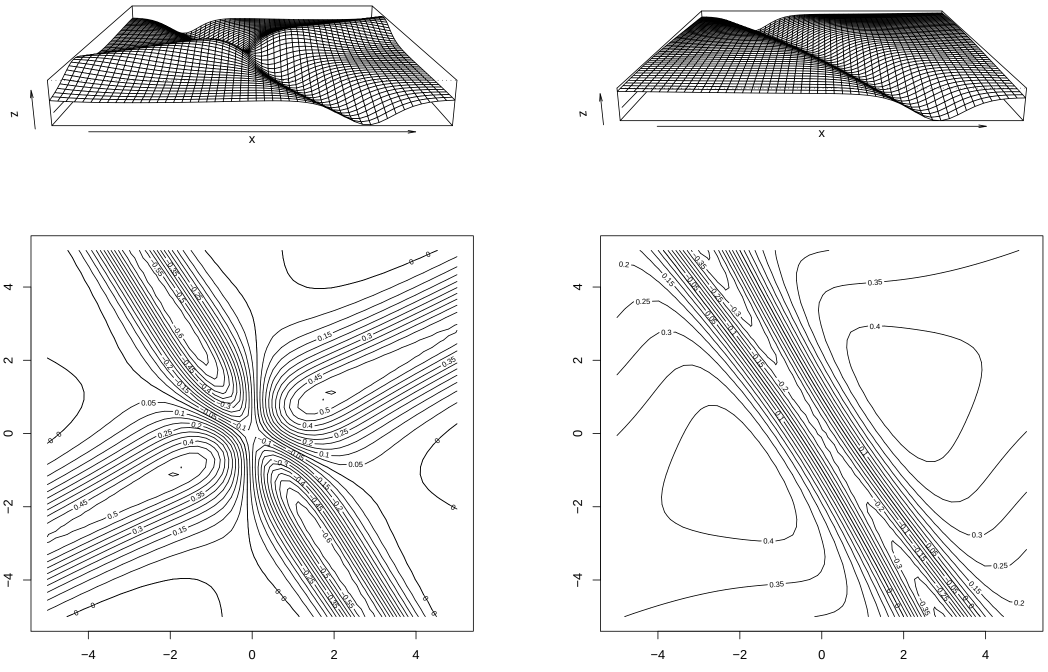
Figure 1: Perspective plots of Bivariate t-distributions with $\boldsymbol{\mu} = [1,2]^T$, $\boldsymbol{A} = \text{diag}(4,4)$. First column: Standard Bivariate t-distributions with (top) $\nu = 0.1$, (bottom) $\nu = 5$. Second column: Multiple dof t-distributions with (top) $\boldsymbol{\nu} = \{0.1, 0.1\}$, $\xi = \frac{\pi}{3}$, (bottom) $\boldsymbol{\nu} = \{1, 10\}$ $\xi = \frac{\pi}{3}$.



(a) $\boldsymbol{\gamma} = \{1, 20\}$, $\boldsymbol{\delta} = \{1, 20\}$, $\boldsymbol{\beta} = [3, 3]^T$ (b) $\boldsymbol{\gamma} = \{1, 1\}$, $\boldsymbol{\delta} = \{1, 100\}$, $\boldsymbol{\beta} = [0.01, 3]^T$

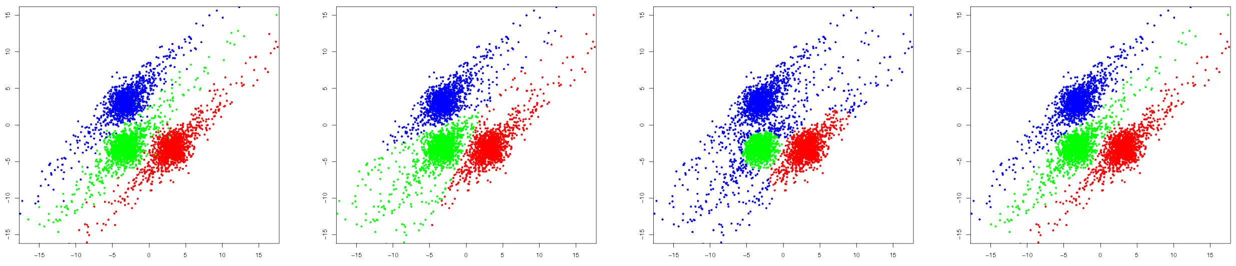
(c) $\boldsymbol{\beta} = [2, 2]^T$ (d) $\boldsymbol{\beta} = [-2, 2]^T$

Figure 2: First row, contour plots of multiple scaled multivariate NIG distributions with $\boldsymbol{\mu} = [0, 0]^T$, $\boldsymbol{\Sigma} = \mathrm{diag}(1, 1)$ and (a) $\boldsymbol{\gamma} = \{1, 20\}$, $\boldsymbol{\delta} = \{1, 20\}$, $\boldsymbol{\beta} = [3, 3]^T$ while in (b) $\boldsymbol{\gamma} = \{1, 1\}$, $\boldsymbol{\delta} = \{1, 100\}$, $\boldsymbol{\beta} = [0.01, 3]^T$. Second row, contour plots of multiple scaled multivariate NIG distributions (solid blue lines) with $\boldsymbol{\mu} = [0, 0]^T$. The difference with the standard multivariate NIG for $\gamma = 1$ and $\delta = 1$ (red dotted lines) is illustrated: $\boldsymbol{\gamma} = \{1, 1\}$, $\boldsymbol{\delta} = \{1, 1\}$ with (c) $\boldsymbol{A} = \mathrm{diag}(1.5, 0.5)$, $\xi = \frac{\pi}{4}$ (equivalently $\boldsymbol{\Sigma}$ has 1 on its diagonal entries and other entries are set to 0.5), $\boldsymbol{\beta} = [2, 2]^T$ while in (d) $\boldsymbol{\Sigma} = \mathrm{diag}(1, 1)$ and $\boldsymbol{\beta} = [-2, 2]^T$.



(a) (b)

Figure 3: Perspective (top) and contour (bottom) plots of the local dependence function for examples of $\mathcal{MS}$ distributions with parameters: (a) $\xi = \pi/3, A_1 = 1.5, A_2 = 1.2, \nu_1 = 1, \nu_2 = 3$; (b) $\xi = \pi/3, A_1 = 1.5, A_2 = 1.2, \nu_1 = 1, \nu_2 = 100$.



(a) True classification (b) t-mixture ν_k=1 (c) t-mixture ν_k=30 (d) MMST

Figure 4: Classification results for a 2-dimensional simulated example of three elongated clusters. (a) True classification; (b) (resp. (c)) Classification for the t-distribution mixture with all dof's fixed to 1 (resp. to 30); (d) Classification for the mixture of multiple scaled t-distributions (MMST) with estimated dofs. The classification for the t-mixture with estimated dof's is closed to (b). The different colours indicate the different components to which observations are assigned to.

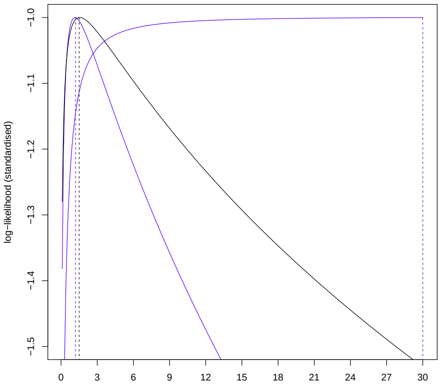
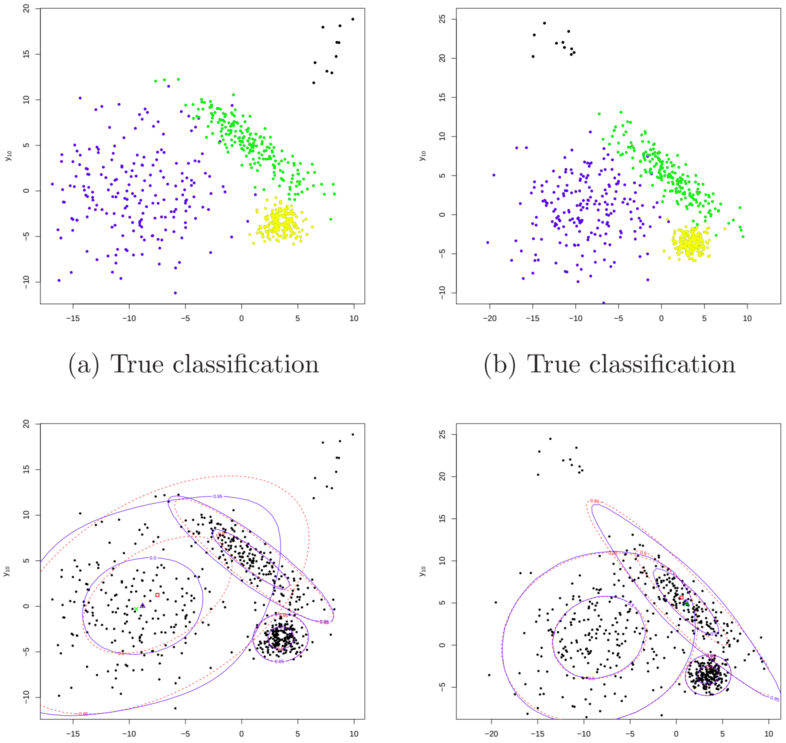


Figure 5: Profile likelihoods in terms of the dof parameter for the upper most (blue) component in Figure 4. The log-likelihoods have been standardized for an easier comparison ($i.e$ |log-likelihood|/max(log-likelihood)). The two blue lines correspond to the two dimensions in the MMST case and the black line to the standard t-mixture. The dashed lines indicate the maximum.



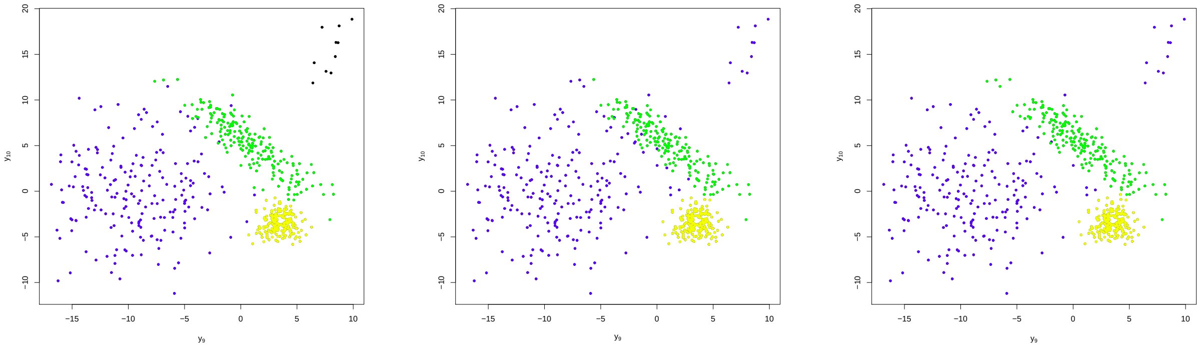
(a) True classification

(b) True classification

(c) Case a

(d) Case b

Figure 6: Last two dimensions of the samples for the 10D simulated data sets. (a) and (b): The colors indicate the three main components of the data. The 10 additional data (outliers) points are shown in black (top right (a) and left (b) hand corners). (c) and (d): Estimation results. The 5% and 95% level contours for each components are shown with dot red lines for the t-mixture and solid blue lines for the MMST. For the most impacted component, the true mean is shown with a green cross while the red square and blue triangle respectively represent the estimated means in the t-mixture and MMST cases.

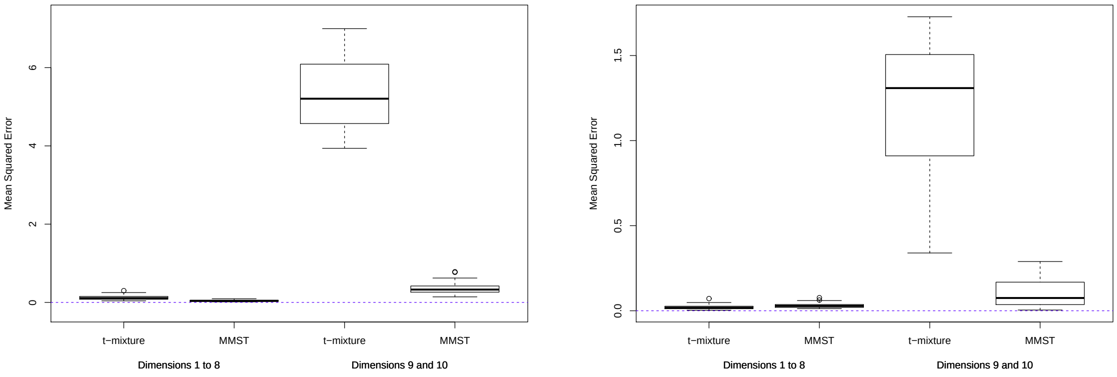


(a) True classification

(b) t-mixture

(c) MMST

Figure 7: Case a: Classification result with the standard t-mixture (b) and the MMST (c). The main difference with the true classification (a) can be seen at the border of components 1 (blue) and 2 (green).



Case a

Case b

Figure 8: Boxplots of the mean squared errors for the mean parameters. The median and variability of the mean squared error in case a and b over 30 simulated datasets for dimensions 1 to 8 are shown on the left hand side and for dimensions 9 and 10 on the right hand side of each plots.

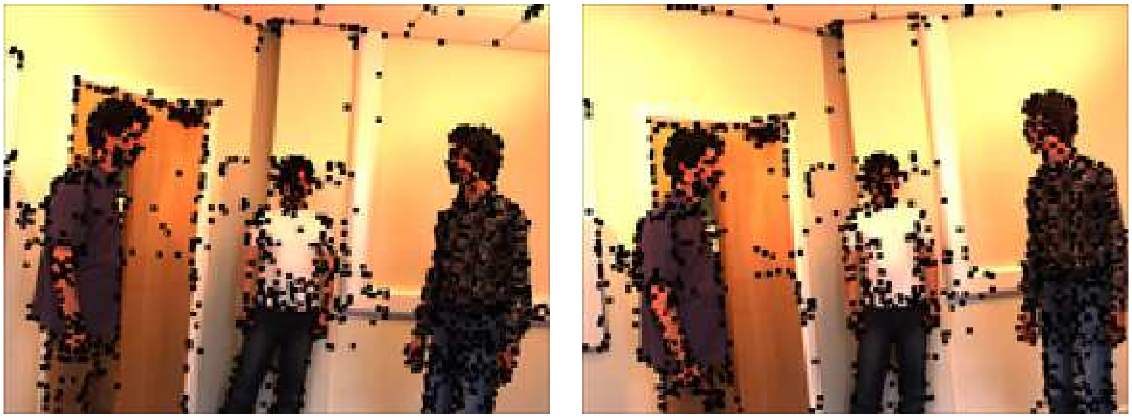


Figure 9: Left and right images corresponding to a frame in a stereo recording with two cameras with the detected interest points in each image. These interest points are then to be matched to produce the 3D data set under consideration.

3D audiovisual recording data

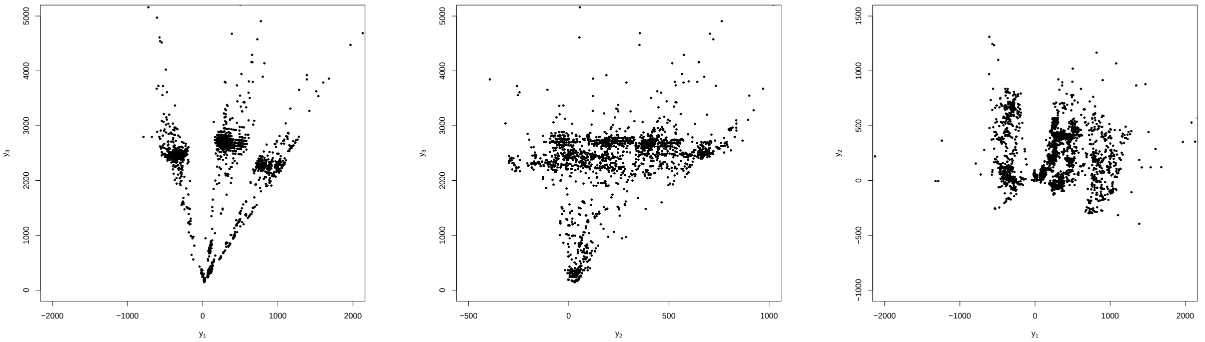


Figure 10: Pairwise scatterplots of the 3D audiovisual recording data with numerous camera artefacts.

3D audiovisual recording data after cutoff

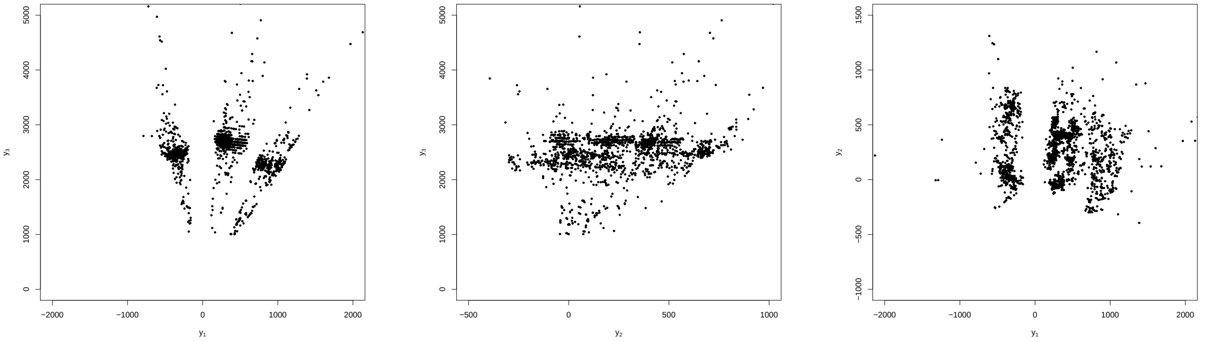


Figure 11: Pairwise scatterplots of the 3D audiovisual recording data. Data after manually removing the artefact points with a cutoff keeping only points such that $Y_3 > 1000$

Objects location estimations

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. However, only two coordinates (Y_1 and Y_3) over three are available. The missing coordinate (Y_2) corresponds to the height and cannot be reliably measured. The reason is that the data is provided by a setting that is efficient at detecting moving textures. In our example, the detected points may then depend on what people wear (*i.e.* clothes with more or less texture), which may significantly impact the determination of the objects location in the vertical dimension. For instance, a person with textured trousers is likely to be located at a lower height than one with uniform ones. In contrast, the projections of the detected points on the ground are much less impacted. The reference coordinates (ground truth) given in Table 3 correspond then to 2D centres of gravity manually determined by an expert from the points assigned to each person and projected on the ground. Table 3 shows the location estimation (in cm) for each detected cluster using the MMST and t-mixture.

Table 3: Estimated person positions on the ground (Y_1 and Y_3 in cm) using the MMST and standard t-mixture with $K = 3$. Locations are estimated as the means of the clusters. The closest to the ground truth estimations are indicated with bold characters.

	Left cluster (green)		Middle cluster (yellow)		Right cluster (blue)	
Estimated model ($K = 3$)	Y_1	Y_3	Y_1	Y_3	Y_1	Y_3
Ground truth	-383.4	2483.4	317.4	2725.0	867.8	2302.3
MMST	**-357.4**	2503.3	342.6	**2699.2**	**894.1**	**2280.2**
t-mixture	-354.8	**2476.0**	**334.9**	2699.1	770.0	2264.8

References

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*, pages 109–116, Chania, Crete, Greece, October 2008. ACM.

G.W. Brier. Verification of forecasts expressed in terms of probability. *Month. Weather Rev.*, 78:13, 1950.

G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

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

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26: 297–302, 1945.

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.

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.

P. W. Holland and Y. J. Wang. Dependence function for continuous bivariate densities. *Communications in Statististics.-Theory and Methods*, 16:863–876, 1987.

M. C. Jones. The local dependence function. *Biometrika*, 83(4):899–904, 1996.

D. Karlis and A. Santourian. Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing*, 19(1):73–83, 2009.

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

V. Khalidov, F. Forbes, and R. Horaud. Conjugate mixture models for clustering multimodal data. *Neural Computation*, 23(2):517–557, 2011.

G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 2008. 2nd edition.

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.