Journal of Classification (2023) 40:145–167
https://doi.org/10.1007/s00357-022-09427-7

Model-Based Clustering and Classification Using Mixtures of Multivariate Skewed Power Exponential Distributions

Utkarsh J. Dang[1] · **Michael P.B. Gallaugher**[2] · **Ryan P. Browne**[3] · **Paul D. McNicholas**[4]

Accepted: 24 November 2022 / Published online: 15 February 2023

Abstract
Families of mixtures of multivariate power exponential (MPE) distributions have already been introduced and shown to be competitive for cluster analysis in comparison to other mixtures of elliptical distributions, including mixtures of Gaussian distributions. A family of mixtures of multivariate skewed power exponential distributions is proposed that combines the flexibility of the MPE distribution with the ability to model skewness. These mixtures are more robust to variations from normality and can account for skewness, varying tail weight, and peakedness of data. A generalized expectation-maximization approach, which combines minorization-maximization and optimization based on accelerated line search algorithms on the Stiefel manifold, is used for parameter estimation. These mixtures are implemented both in the unsupervised and semi-supervised classification frameworks. Both simulated and real data are used for illustration and comparison to other mixture families.

Keywords Generalized expectation-maximization algorithm · Mixture models · Model-based classification · Model-based clustering · Multivariate skewed power exponential distribution

1 Introduction

Mixture modeling has been firmly established in the literature as a useful method for finding homogeneous groups within heterogeneous data. Using mixture models for cluster analysis has a long history (Hasselblad, 1966; Day, 1969) dating at least to Wolfe (1965), who used a Gaussian mixture model for clustering. When using mixture models for clustering,

✉ Utkarsh J. Dang
utkarshdang@cunet.carleton.ca

[1] Department of Health Sciences, Carleton University, Ottawa, Ontario, Canada

[2] Department of Statistical Science, Baylor University, Waco, TX, USA

[3] Department of Statistics & Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada

[4] Department of Mathematics & Statistics, McMaster University, Hamilton, Ontario, Canada

which is known as model-based clustering, mixture models are used to partition data points to learn group memberships, or labels, of observations with unknown labels. If some observations are a priori labeled, a semi-supervised analogue of model-based clustering is used and this is known as model-based classification. Extensive details on model-based clustering and classification are given by McNicholas (2016a) and recent reviews are provided by Bouveyron and Brunet-Saumard (2014) and McNicholas (2016b).

A G-component finite mixture model assumes that a random vector $\mathbf{X}$ has density of the form

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g),$$

where $g = 1, \ldots, G$, $\pi_g > 0$ are the mixing proportions with $\sum_{g=1}^{G} \pi_g = 1$, and $f_g(\cdot)$ are the component densities. The Gaussian mixture model (see, e.g., Banfield and Raftery, 1993; Celeux and Govaert, 1995; Tipping and Bishop 1999; McNicholas and Murphy, 2008) remains popular due to its mathematical tractability. However, it is inflexible in the presence of cluster skewness and different levels of cluster kurtosis, and has been known to result in an overestimate of the number of clusters and poor density estimation for known clusters (see Franczak et al., 2014; Dang et al., 2015, for examples). Therefore, it has become popular to consider mixtures of more flexible distributions for clustering to deal with such scenarios.

Mixture models that can deal with varying cluster tail-weight, skewness and/or concentration, and kurtosis are increasingly becoming common. A small selection of such models include mixtures using power transformations (Zhu et al., 2022), mixtures of multivariate t-distributions (Peel & McLachlan, 2000; Andrews & McNicholas, 2012), mixtures of normal inverse Gaussian distributions (Karlis & Santourian, 2009; Subedi & McNicholas, 2014; O'Hagan et al., 2016), mixtures of skew-t distributions (Lin, 2010; Murray et al., 2014; Vrbik & McNicholas, 2014; Lee & McLachlan, 2014; 2016), mixtures of shifted asymmetric Laplace distributions (Morris & McNicholas, 2013; Franczak et al., 2014), mixtures of multivariate power exponential distributions (Dang et al., 2015), mixtures of variance-gamma distributions (McNicholas et al., 2017), and mixtures of generalized hyperbolic distributions and variations thereof (Browne & McNicholas, 2015; Murray et al., 2017).

Two common approaches to introducing skewness are by means of a normal variance-mean mixture model and via hidden truncation using an elliptical distribution and a skewing function. The former assumes that a random vector $\mathbf{X}$ can be written in the form

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{V},$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\alpha}$ are location and skewness vectors, respectively, $\mathbf{V} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $W \perp \mathbf{V}$, and $W > 0$ is a positive random variable with density $h(w|\boldsymbol{\Theta})$. Depending on the distribution of W, different skewed distributions can be derived, e.g., the generalized hyperbolic, skew-t, variance-gamma, and normal inverse Gaussian distributions. The hidden truncation approach makes use of a combination of an elliptical distribution and a skewing function. For example, a random vector $\mathbf{X}$ follows a multivariate skew-normal distribution with skewness $\boldsymbol{\alpha}$ if its density can be written as follows:

$$f(\mathbf{x}) = 2\phi_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\alpha}'\mathbf{x}),$$

where $\phi_p(\cdot)$ is the density of the p-dimensional normal distribution and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution (Azzalini & Valle, 1996).

The multivariate power exponential (MPE) distribution (Gómez et al., 1998) has been used in many different applications (e.g., Lindsey, Cho & Bui, and Verdoolaege et al. 1999,

2005, and 2008) and was recently used in the mixture model context by Dang et al. (2015). Depending on the shape parameter β, either a leptokurtic or platykurtic distribution can be obtained. Specifically, if $\beta \in (0, 1)$ then the distribution is leptokurtic, which is characterized by a thinner peak and heavy tails compared to the Gaussian distribution. If $\beta > 1$, a platykurtic distribution is obtained, which is characterized by a flatter peak and thin tails compared to the Gaussian distribution. Other distributions can also be obtained for specific values of the shape parameter, for example, for $\beta = 0.5$, the distribution is a Laplace (double-exponential) distribution and, for $\beta = 1$, it is a Gaussian distribution. Furthermore, when $\beta \to \infty$, the MPE becomes a multivariate uniform distribution.

Dang et al. (2015) derived a family of mixtures of MPE distributions but those mixtures could only account for elliptical clusters. Previously, skew power exponential distributions have been discussed in the univariate case with constrained β (Azzalini, 1986; DiCiccio & Monti, 2004; da Silva Ferreira et al., 2011) or in the multivariate case as scale mixture of skew-normal with constrained β (Branco & Dey, 2001). Herein, we present mixtures based on a novel multivariate skewed power exponential (MSPE) distribution. As compared to earlier proposals, this distribution is more suitable for clustering and classification purposes and can be used for a wide range of β (heavy, Gaussian, and light tails). Using an eigen-decomposition of the component scale distributions (à la Celeux & Govaert, 1995), we construct a family of 16 MSPE mixture models for use in both clustering and semi-supervised classification. These models can account for varying tail weight (heavy, Gaussian, or light), peakedness (thinner or thicker than Gaussian), and skewness of mixture components.

2 Background

Using the parametrization given by Gómez et al. (1998), a random vector $\mathbf{X}$ follows a p-dimensional power exponential distribution if the density is

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta) = \frac{p\Gamma\left(\frac{p}{2}\right)}{\pi^{p/2}\Gamma\left(1+\frac{p}{2\beta}\right)2^{1+\frac{p}{2\beta}}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\delta(\mathbf{x})^{\beta}\right\}, \tag{1}$$

where $\boldsymbol{\mu}$ is the location parameter, $\boldsymbol{\Sigma}$ is a scale matrix, β determines the kurtosis, and

$$\delta(\mathbf{x}) := \delta(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}).$$

In a similar manner to Azzalini and Valle (1996), Lin et al. (2014) derived the multivariate skew t-normal distribution by using an elliptical multivariate t-distribution and the cumulative distribution function of the standard normal distribution as the skewing function.

Herein, the skewness function is still the $N(0, 1)$ cumulative distribution function while the elliptical distribution is now the MPE distribution. Specifically, a random vector $\mathbf{X}$ follows a p-dimensional skew power exponential distribution if the density is of the form

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, \boldsymbol{\psi}) &= 2\, g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta)\, \Phi(\boldsymbol{\psi}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})), \\ &= \frac{2p\Gamma\left(\frac{p}{2}\right)}{\pi^{p/2}\Gamma\left(1+\frac{p}{2\beta}\right)2^{1+\frac{p}{2\beta}}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\delta(\mathbf{x})^{\beta}\right\}\Phi\left(\boldsymbol{\psi}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})\right) \end{aligned} \tag{2}$$

with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$, shape parameter β, and skewness vector $\boldsymbol{\psi}$. Some special cases of this distribution include the skew-normal distribution ($\beta = 1$), a variant of a skew Laplace distribution ($\beta = 0.5$), the power exponential distribution ($\psi = 0$), and

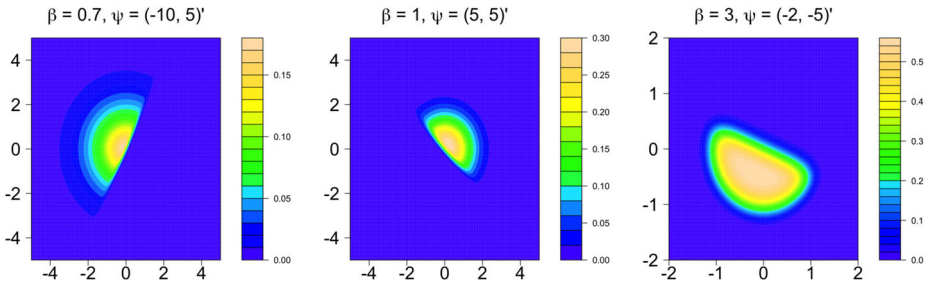

Fig. 1 Contours of the multivariate skew power exponential distribution for different values of the shape and skewness parameters with $\boldsymbol{\mu} = (0, 0)'$ and an identity scale matrix. The middle panel, with $\beta = 1$, is a multivariate skew-normal distribution

a generalization of the multivariate uniform distribution ($\beta \to \infty, \psi = 0$). Examples of contours of the MSPE distribution are given in Fig. 1.

3 Mixtures of MSPE Distributions

3.1 Inference

An iterative procedure is used for parameter estimation; specifically, a generalized expectation-maximization (GEM) algorithm (Dempster et al., 1977) with conditional maximization steps. The expectation-maximization (EM) algorithm (Dempster et al., 1977) is an iterative procedure in which the conditional expected value of the complete-data log-likelihood is maximized on each iteration to yield parameter updates. As opposed to the EM algorithm, the conditional maximization steps increase, rather than maximize, the conditional expected value of the complete-data log-likelihood in each iteration of a GEM algorithm. Consider a random sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a p-dimensional MSPE mixture distribution from a population with G subgroups. If we define

$$z_{ig} = \begin{cases} 1 & \text{if } \mathbf{x}_i \text{is from group} g, \\ 0 & \text{otherwise,} \end{cases}$$

then the complete-data log-likelihood can be written as follows:

$$\mathcal{L}_c(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig} \log\left[2\pi_g \frac{p\Gamma\left(\frac{p}{2}\right)}{\Gamma\left(1+\frac{p}{2\beta_g}\right)2^{1+\frac{p}{2\beta_g}}\pi^{p/2}}|\boldsymbol{\Sigma}_g|^{-\frac{1}{2}} \times \exp\left\{-\frac{\delta_{ig}(\mathbf{x}_i)^{\beta_g}}{2}\right\}\Phi\left(\boldsymbol{\psi}_g'\boldsymbol{\Sigma}_g^{-1/2}(\mathbf{x}_i-\boldsymbol{\mu}_g)\right)\right].$$

For parsimony, an eigen-decomposition is commonly imposed on component scale matrices using the re-parameterization $\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g'$, where $\boldsymbol{\Delta}_g$ is a diagonal matrix with entries proportional to the eigenvalues of $\boldsymbol{\Sigma}_g$ (with $|\boldsymbol{\Delta}_g| = 1$), λ_g is the associated constant of proportionality, and $\boldsymbol{\Gamma}_g$ is a $p \times p$ orthogonal matrix of the eigenvectors of $\boldsymbol{\Sigma}_g$ with entries ordered according to the eigenvalues (Banfield & Raftery, 1993; Celeux & Govaert, 1995). A subset of eight models was considered in Dang et al. (2015) including the most parsimonious (EII) and the fully unconstrained (VVV) models, along with a possible

Table 1 Nomenclature, scale matrix structure, and the number of free scale parameters for the eigen-decomposed family of models

Model	λ_g	$\mathbf{\Delta}_g$	$\mathbf{\Gamma}_g$	$\mathbf{\Sigma}_g$	Free parameters
EII	Equal	Spherical	–	$\lambda \boldsymbol{I}$	1
VII	Variable	Spherical	–	$\lambda_g \boldsymbol{I}$	G
EEI	Equal	Equal	Axis-aligned	$\lambda \mathbf{\Delta}$	p
VVI	Variable	Variable	Axis-aligned	$\lambda_g \mathbf{\Delta}_g$	Gp
EEE	Equal	Equal	Equal	$\lambda \mathbf{\Gamma} \mathbf{\Delta} \mathbf{\Gamma}'$	$p\,(p+1)\,/2$
EEV	Equal	Equal	Variable	$\lambda \mathbf{\Gamma}_g \mathbf{\Delta} \mathbf{\Gamma}_g'$	$Gp(p+1)/2-(G-1)p$
VVE	Variable	Variable	Equal	$\lambda_g \mathbf{\Gamma} \mathbf{\Delta}_g \mathbf{\Gamma}'$	$p(p+1)/2+(G-1)p$
VVV	Variable	Variable	Variable	$\lambda_g \mathbf{\Gamma}_g \mathbf{\Delta}_g \mathbf{\Gamma}_g'$	$Gp\,(p+1)\,/2$

constraint on β_g, for their family of mixture models using elliptical power exponential distributions (Table 1). Herein, we consider the same eight models to form a family of mixtures of skewed power exponential distributions.

After initialization (Section 3.2), the algorithm proceeds as follows.

E-Step: In the E-step, the group membership estimates $\hat{z}_{ig}$ are updated using

$$\hat{z}_{ig} := \mathbb{E}_{\widehat{\mathbf{\Theta}}}[Z_{ig}|\mathbf{x}_i] = \frac{\hat{\pi}_g f\left(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_g, \hat{\mathbf{\Sigma}}_g, \hat{\beta}_g, \hat{\boldsymbol{\psi}}_g\right)}{\sum_{j=1}^{G}\hat{\pi}_j f\left(\mathbf{x}_i|\hat{\boldsymbol{\mu}}_j, \hat{\mathbf{\Sigma}}_j, \hat{\beta}_j, \hat{\boldsymbol{\psi}}_j\right)},$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, G$.

M-Step: The update for π_g is $\hat{\pi}_g = n_g/n$, where $n_g = \sum_{i=1}^{n}\hat{z}_{ig}$. However, the updates for $\boldsymbol{\mu}_g$, $\mathbf{\Sigma}_g$, β_g, and $\boldsymbol{\psi}_g$ are not available in closed form. For estimating β_g, either a Newton-Raphson method or a root finding algorithm may be used and is identical to the estimate in Dang et al. (2015). In our implemented code, we constrain β_g to be less than 20 for numerical stability. Let

$$\mathcal{Q} := \mathbb{E}_{\mathbf{\Theta}}[\mathcal{L}_c(\mathbf{\Theta}|\mathbf{x})].$$

Then, a Newton-Raphson update is used for the location parameter $\hat{\boldsymbol{\mu}}_g$ with the following:

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\mu}_g} &= \hat{\beta}_g \sum_{i=1}^{n}\hat{z}_{ig}\delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g-1}\hat{\mathbf{\Sigma}}_g^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g) - \sum_{i=1}^{n}\hat{z}_{ig}\frac{\phi(\boldsymbol{\psi}_g'\mathbf{\Sigma}_g^{-1/2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g))}{\Phi(\boldsymbol{\psi}_g'\mathbf{\Sigma}_g^{-1/2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g))}\mathbf{\Sigma}_g^{-1/2}\boldsymbol{\psi}_g, \quad (3)\\
\frac{\partial^2 \mathcal{Q}}{\partial \boldsymbol{\mu}_g \boldsymbol{\mu}_g'} &= \hat{\beta}_g \sum_{i=1}^{n}\hat{z}_{ig}\left[-\delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g-1}\hat{\mathbf{\Sigma}}_g^{-1} - 2(\hat{\beta}_g-1)\delta_{ig}(\mathbf{x}_i)^{\hat{\beta}_g-2}\hat{\mathbf{\Sigma}}_g^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)'\hat{\mathbf{\Sigma}}_g^{-1}\right]\\
&\quad - \sum_{i=1}^{n}\hat{z}_{ig}\boldsymbol{\psi}_g'\mathbf{\Sigma}_g^{-1/2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g)\frac{\phi(\boldsymbol{\psi}_g'\mathbf{\Sigma}_g^{-1/2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g))}{\Phi(\boldsymbol{\psi}_g'\mathbf{\Sigma}_g^{-1/2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g))}\hat{\mathbf{\Sigma}}_g^{-1}\boldsymbol{\psi}_g\boldsymbol{\psi}_g'\hat{\mathbf{\Sigma}}_g^{-1}\\
&\quad - \sum_{i=1}^{n}\hat{z}_{ig}\left[\frac{\phi(\boldsymbol{\psi}_g'\mathbf{\Sigma}_g^{-1/2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g))}{\Phi(\boldsymbol{\psi}_g'\mathbf{\Sigma}_g^{-1/2}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g))}\right]^2\hat{\mathbf{\Sigma}}_g^{-1}\boldsymbol{\psi}_g\boldsymbol{\psi}_g'\hat{\mathbf{\Sigma}}_g^{-1},
\end{aligned}$$

where $\delta_{ig}(\mathbf{x}_i) := \left(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g\right)'\hat{\mathbf{\Sigma}}_g^{-1}\left(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_g\right)$.

For estimating the skewness parameter $\boldsymbol{\psi}$, the density is first re-parameterized as follows:

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta, \boldsymbol{\psi}) &= 2\, g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta)\, \Phi(\boldsymbol{\psi}'\boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu})) \\ &= 2\, g(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta)\, \Phi(\boldsymbol{\eta}'(\mathbf{x}-\boldsymbol{\mu})), \end{aligned}$$

where $\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\psi}$. A quadratic lower-bound principle (Böhning & Lindsay, 1988; Hunter & Lange, 2004) on the relevant part of the complete-data log-likelihood using the re-parameterized density uses the following property to construct a quadratic minorizer:

$$\log(\Phi(s)) \geq \log(\Phi(s_0)) + \frac{\phi(s_0)}{\Phi(s_0)}(s - s_0) + \frac{1}{2}(-1)(s - s_0)^2,$$

where -1 is the lower bound of the second derivative in the Taylor series around s_0. Then, an estimate for $\boldsymbol{\eta}_g$ yields

$$\begin{aligned} \boldsymbol{\eta}_g = \left[\sum_{i=1}^{n} \hat{z}_{ig}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)'\right]^{-1} \Bigg[\sum_{i=1}^{n} \hat{z}_{ig} \frac{\phi(\boldsymbol{\eta}_{g0}'(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g))}{\Phi(\boldsymbol{\eta}_{g0}'(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g))}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g) \\ + \sum_{i=1}^{n} \hat{z}_{ig}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g)'\boldsymbol{\eta}_{g0}\Bigg] \end{aligned}$$

and we can back-transform to obtain $\boldsymbol{\psi}_g = \boldsymbol{\Sigma}_g^{1/2}\boldsymbol{\eta}_g$.

For the scale matrices $\boldsymbol{\Sigma}_g$, the estimation makes use of minorization-maximization algorithms (Hunter & Lange, 2000; 2004) by exploiting the concavity of the functions containing $\boldsymbol{\Sigma}_g$ (or parts of its decomposition) and accelerated line search algorithms on the Stiefel manifold (Absil et al., 2009; Browne & McNicholas, 2014), with different updates depending on whether the latest estimate for β_g is less than 1 or is greater than or equal to 1. For more details, see Dang et al. (2015). Combining the constraints of the eigen-decomposition in Table 1, with constraining β_g to be equal or different between groups, results in a family of 16 models. For example, a VVIE model represents a VVI scale structure (as in Table 1) and the shape parameter constrained to be equal between groups ($\beta_g = \beta$).

3.2 Initialization

It is well known that the performance of the EM algorithm depends heavily on the starting values. The following strategy is adopted. The group memberships are initialized using a combination of emEM approach (Biernacki et al., 2003), k-means algorithm (Hartigan & Wong, 1979), and random soft starts. Specifically, the most superior run (highest log-likelihood) from 10 different k-means-based *short EM* runs is chosen for a *long EM* run. This process is repeated with random soft starts. The fits of the two *long EM* runs are compared based on a model selection criterion (see Section 3.3) to choose a best model. Once these initial memberships are set, the $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ are initialized using a constrained model. The kurtosis parameters β_g are initialized to 0.5, and the skewness is initialized as a zero vector.

3.3 Convergence, Model Selection, and Performance Assessment

Following Lindsay (1995) and McNicholas et al. (2010), the iterative GEM algorithm is stopped based on the Aitken's acceleration (Aitken 1926). Specifically, an asymptotic estimate of the log-likelihood at iteration $k + 1$ is compared with the current log-likelihood

value and considered converged when the difference is less than some positive ϵ. We use $\epsilon = 0.005$ for the analyses herein.

In a general clustering scenario, the number of groups is generally not known a priori and the covariance model is not known. Therefore, a model selection criterion is required. The most common criterion for model selection is the Bayesian information criterion (BIC; Schwarz 1978). The BIC can be written as follows:

$$\text{BIC} = 2l(\hat{\boldsymbol{\vartheta}}) - m \log n, \tag{4}$$

where m is the number of free parameters, n is the sample size, and $l(\hat{\boldsymbol{\vartheta}})$ is the maximized log-likelihood. When written as in Eq. 4, a greater BIC represents a superior model fit. The integrated complete likelihood (ICL; Biernacki et al. 2000) was also considered for model selection; however, in initial testing, the ICL did not consistently outperform the BIC in simulations and thus, for the remainder of this manuscript, we use only the BIC.

To evaluate classification performance, we use the adjusted Rand index (ARI; Hubert and Arabie 1985). The ARI compares two different partitions; specifically, in our case, the estimated classification and the (known) true classifications. The ARI takes a value of 1 when there is perfect classification and has expected value 0 under random classification (see Steinley (2004) for extensive details on the ARI).

4 Analyses

4.1 Overview

The performance of the MSPE mixture models is compared with mixture model implementations based on the MPE distribution (Dang et al., 2015), as well as implementations from the `mixture` package (Pocuca et al., 2022) of the generalized hyperbolic distribution (`ghpcm`) and the Gaussian distribution (`gpcm`). We chose these mixtures for comparison as Gaussian mixtures remain widely used and the generalized hyperbolic distribution has special cases that include some parameterizations of inverse Gaussian, variance-gamma, skew-t (note there are formulations of the skew-t distribution that cannot be obtained from the generalized hyperbolic), multivariate normal-inverse Gaussian, and asymmetric Laplace distribution. Using these comparators, we obtain comparisons to mixtures based on purely elliptical (Gaussian), elliptical with flexible kurtosis modeling (MPE), and skewed (generalized hyperbolic) distributions. For a fair comparison, we restrict the models in the other implementations to those in Table 1. In addition, we use BIC as the model selection criterion, we run $G = 1, \ldots, 4$ for all simulations and real data analyses, and we use the same starting soft memberships for all comparator models. Data from the MSPE distribution is simulated using a Metropolis-Hastings rule.

4.2 Simulations

4.2.1 Simulation 1: Heavy and Light-Tailed Skewed Clusters

A three-component mixture is simulated with 500 observations in total. Group sample sizes are sampled from a multinomial distribution with mixing proportions $(0.35, 0.25, 0.4)'$. The first component is simulated from a heavy-tailed three-dimensional MSPE distribution with $\boldsymbol{\mu}_1 = (3, 3, 0)'$, $\beta_1 = 0.85$, and $\boldsymbol{\psi}_1 = (-5, -10, 0)'$. The second component is simulated with $\boldsymbol{\mu}_2 = (3, 6, 0)'$, $\beta_2 = 0.9$, and $\boldsymbol{\psi}_2 = (15, 10, 0)'$. The third component is simulated

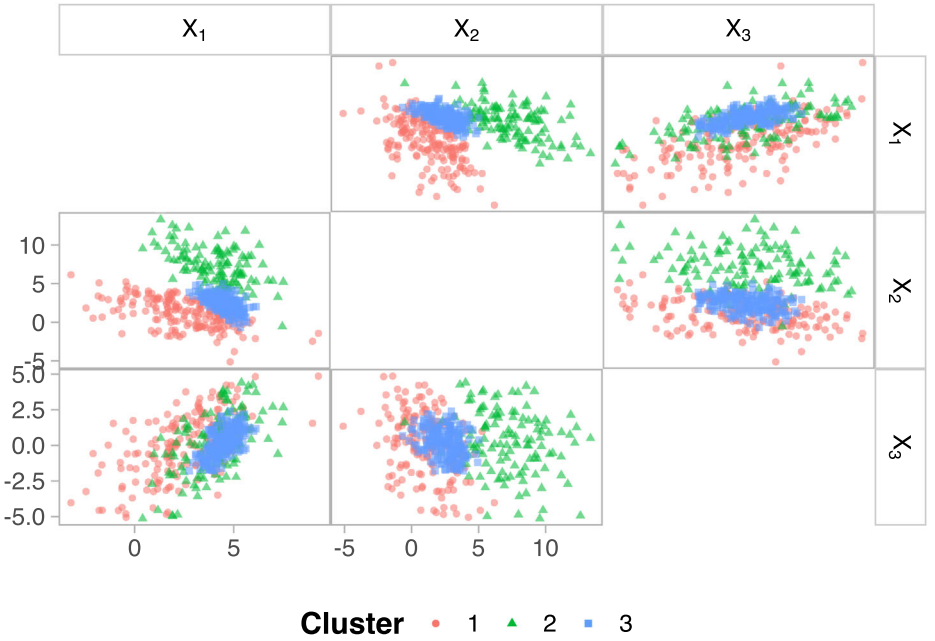

Fig. 2 An example scatterplot matrix of the three-component mixture of simulation 1

with light tails with $\boldsymbol{\mu}_3 = (4, 2, 0)'$, $\beta_3 = 2$, and $\boldsymbol{\psi}_3 = \boldsymbol{\psi}_2$. Lastly, the scale matrices were common to all three components with diag $(\boldsymbol{\Delta}_g) = (4, 3, 1)'$ and

$$\boldsymbol{\Gamma}_g = \begin{pmatrix} 0.36 & 0.48 & -0.80 \\ -0.80 & 0.60 & 0.0 \\ 0.48 & 0.64 & 0.6 \end{pmatrix},$$

for $g = 1, 2, 3$. The simulated components are not well separated (an example scatterplot matrix is given in Fig. 2). All four mixture implementations are run on 100 such datasets for $G = 1, \ldots, 4$.

For the MSPE family, a three-component (four-component) model is selected by the BIC 98 (2) times. For the MPE family, the BIC selects a three-component (four-component) model 87 (13) times. When the four-component models are selected, typically, this is because the model chosen has split up the heavy-tailed cluster into two separate components. The exception is that for two chosen four-component MPE solutions, the light-tailed cluster was split into two separate components. For the `gpcm` family, the BIC selects a three-component (four-component) model 99 (1) times. On the other hand, for the `ghpcm` algorithm, one-component, two-component, and three-component models are selected 3, 46, and 51 times, respectively. This under-fitting—the heavy-tailed and light-tailed components are merged—may be due to the use of the BIC.

The ARI values for the selected MSPE models range from 0.71 to 0.91, with a median (mean) ARI value of 0.83 (0.83). The selected MPE models yield ARI values ranging between 0.67 and 0.91, with a median (mean) value of 0.80 (0.79). The selected `gpcm` models yield ARI values ranging between 0.71 and 0.91, with a median (mean) value of 0.80 (0.80). Similarly, the `ghpcm` algorithm yields ARI values ranging between 0 and 0.82, with a median (mean) value of 0.41 (0.47). For the MSPE models, an EEEV model was

selected 68 out of the 100 times, with a less parsimonious model selected the other times. To demonstrate parameter recovery, a similar simulation setup was carried out and findings are provided in in the Appendix.

4.2.2 Simulation 2: Heavy-Tailed and Gaussian Skewed Clusters

A three-component mixture is simulated with 500 observations in total. Group sample sizes are sampled from a multinomial distribution with mixing proportions $(0.3, 0.45, 0.25)'$. The first component is simulated from a three-dimensional skewed normal distribution (i.e., $\beta_1 = 1$) with $\boldsymbol{\mu}_1 = (0, 1, 2)'$ and $\boldsymbol{\psi}_1 = (3, 5, 10)'$. The second component is simulated from a heavy-tailed three-dimensional MSPE distribution with $\boldsymbol{\mu}_2 = (0, 4, 2)'$, $\beta_2 = 0.8$, and $\boldsymbol{\psi}_2 = (-3, 5, -5)'$. The third component is simulated with $\boldsymbol{\mu}_3 = (-2, -3, 0)'$, $\beta_3 = 0.9$, and $\boldsymbol{\psi}_3 = (5, 10, -5)'$. Lastly, the scale matrices are

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1.00 & 0.50 & 0.40 \\ 0.50 & 1.50 & 0.35 \\ 0.40 & 0.35 & 1.20 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1.00 & 0.30 & 0.20 \\ 0.30 & 1.50 & 0.30 \\ 0.20 & 0.30 & 1.20 \end{pmatrix}.$$

Again, the simulated components are not well separated and all four mixture implementations are run on 100 such datasets for $G = 1, \ldots, 4$.

For the MSPE family, a three-component (four-component) model is selected by the BIC 95 (5) times. For the MPE mixture, the BIC selected a three-component (four-component) model 99 (1) times. Interestingly, for the `ghpcm` mixtures, the BIC selects a three-component model all 100 times. On the other hand, for the `gpcm` family, the BIC selects a three-component (four-component) model 94 (6) times. In all cases when the four-component models are selected, this is because the model chosen has split up one of the heavy-tailed clusters into two components.

The ARI values for the selected MSPE models range from 0.74 to 0.98, with a median (mean) ARI value of 0.94 (0.93). The selected MPE models yield ARI values ranging between 0.72 and 0.96, with a median (mean) value of 0.90 (0.89). Similarly, the `ghpcm` algorithm yields ARI values ranging between 0.70 and 0.94, with a median (mean) value of 0.86 (0.85). The selected `gpcm` models yield ARI values ranging between 0.68 and 0.95, with a median (mean) value of 0.90 (0.88). For the MSPE models, an EEEV model is selected 76 out of 100 times.

4.2.3 Simulation 3: Two Light-Tailed Elliptical Clusters

A simulation from Dang et al. (2015) is replicated, where a two-component EIIV model is simulated with 450 observations with the sample sizes for each group sampled from a binomial distribution with success probability 0.45. Both components had identity scale matrices and zero skewness. The first component is simulated from a two-dimensional MPE distribution with $\boldsymbol{\mu}_1 = (0, 0)'$ and $\beta_1 = 2$ while the second component is simulated using $\boldsymbol{\mu}_2 = (2, 0)'$ and $\beta_2 = 5$. Again, the simulated components are not well separated. All four algorithms are run on 100 such datasets. For the MSPE and MPE families, a two-component model is selected by the BIC for 100 and 99 datasets, respectively. The dataset where a three-component model is selected for the MPE models involves a cluster of four observations that are tightly clustered with tiny eigenvalues. On the other hand, for the `gpcm` family, the BIC selects a two-, three-, and four-component model 80, 12, and 8 times, respectively. Similarly, for the `ghpcm` algorithm, one-, two-, and three-component models are selected

Table 2 Performance comparison of four mixture model families on simulations

		Simulation 1	Simulation 2	Simulation 3
MSPE	Frequencies	3 (98); 4 (2)	3 (95); 4 (5)	2 (100)
	ARI	**0.83 (0.81, 0.86)**	**0.94 (0.93, 0.95)**	**0.88 (0.85, 0.90)**
MPE	Frequencies	3 (87); 4 (13)	3 (99); 4 (1)	2 (99); 3 (1)
	ARI	0.80 (0.76, 0.83)	0.90 (0.88, 0.91)	**0.88 (0.86, 0.90)**
`ghpcm`	Frequencies	1 (3); 2 (46); 3 (51)	3 (100)	1 (1); 2 (95); 3 (4)
	ARI	0.41 (0.30, 0.66)	0.86 (0.82, 0.88)	0.81 (0.79, 0.85)
`gpcm`	Frequencies	3 (99); 4 (1)	3 (94); 4 (6)	2 (80); 3 (12); 4 (8)
	ARI	0.80 (0.77, 0.82)	0.90 (0.87, 0.91)	0.85 (0.77, 0.89)

For each simulation and implementation, a frequency table of number of groups of the best selected model according to the BIC is provided. Median ARI is provided as well as the first and third quartiles for ARI across 100 datasets in each simulation in parentheses as table entries. The ARI for the comparator(s) with superior performance is bolded

1, 95, and 4 times, respectively. Here, more components are being fitted to deal with the light-tailed nature of the data.

For the MPE family (which was used to simulate the data), the ARI values for the selected models range from 0.81 to 0.95, with a median (mean) ARI value of 0.88 (0.88). The MSPE family performed similarly, as expected—the MPE is a special case of the MSPE family—with the ARI values for the selected models ranging from 0.81 to 0.94, with a median (mean) ARI value of 0.88 (0.88). The selected `gpcm` models yield ARI values ranging between 0.35 and 0.96, with a median (mean) value of 0.85 (0.80). Similarly, the `ghpcm` algorithm yields ARI values ranging between 0 and 0.91, with a median (mean) value of 0.81 (0.80). For the MSPE and MPE models, an EIIV model was selected 89 and 95 times out of 100, respectively. Because the ranges of ARI can be more reflective of one poor or great fit, the median, first and third quartile of the ARIs for the selected models are summarized in Table 2 for all simulations.

4.3 Dataset Descriptions

For assessment of performance on real data, we considered the following "benchmark" datasets, i.e., datasets often used for comparison of clustering algorithms, of various sizes and dimensionalities available through various R packages:

- **Body Dataset**: The `body` data from the `gclus` package (Hurley, 2012) has 24 measurements on body dimension, age, etc., for 507 individuals (247 men and 260 women).
- **Coffee Dataset**: The `coffee` dataset (Streuli, 1973), obtained from the `pgmm` package (McNicholas et al., 2022), has 12 chemical measurements on two types of coffee (*arabica* and *robusta*).
- **Female Voles Dataset**: The `fvoles` dataset (Flury, 2012) has measurements on six size variables and age of 86 female voles from two different species (*californicus* and *ochrogaster*).

- **Olive Dataset**: The `olive` dataset (Forina & Tiscornia, 1982), obtained from the `pgmm` package, has percentage composition of eight fatty acids of olive oils from three regions of Italy.
- **Penguins Dataset**: The `penguins` dataset (Horst et al., 2020) has measurements on bill length, bill depth, flipper length, and body mass of 342 penguins from three different species (*Adélie*, *Chinstrap*, and *Gentoo*).
- **Wine Dataset**: The thirteen variable `wine` dataset, obtained from (Hurley, 2012), has 13 different measurements of chemical aspects of 178 Italian wines from three different types of wine.

Several other datasets were also considered, on which performance of the different algorithms is summarized in the Appendix.

4.4 Unsupervised Classification

Unsupervised classification is performed on (scaled) datasets mentioned above using the same comparison distributions as on the simulated data. The ARI and the number of groups chosen by the BIC are shown in Table 3 along with the classification tables in Table 4.

In the case of the `body` dataset, heavy-tailed components with some skewness are fit with the best selected MSPE model (an EEEV model). For the other three comparators, the BIC selected a four-component model, with the fewest misclassifications for the MPE model. The `coffee` dataset is a small dataset ($n = 43$), on which the selected MSPE model (heavy-tailed) had perfect classification. The other three comparators overfit the number of components, with MPE and `gpcm` obtaining the same classification and the `ghpcm` model fitting two small clusters containing six and seven observations, respectively. In the case of the `fvoles` dataset, skewed yet light-tailed clusters are fit. However, the `gpcm` model performed equally well here while the `ghpcm` model misclassified two more observations in comparison. On the other hand, the selected MPE model fit three components (with one component containing only six observations). In the case of the `olive` data, the BIC selected a similarly performing, four-component, solution for all four comparators ($G = 3$) with olive oils from Southern Italy split into two components. The MSPE model fit heavy-tailed components. On the `penguins` data, the two skewed mixtures MSPE and `ghpcm` performed best, fitting the "correct" number of components. The MSPE model fit heavy (with some skewness) clusters. The selected models from the MPE and `gpcm` split different species into two different components with the best fit `gpcm` model having the largest number of misclassifications of all the comparators. On the other hand, in the case of the `wine` data, the MSPE mixture fit heavy-tailed components with five misclassifications. The BIC selected a four-component solution for the other three comparators.

Note that, for the `body` and `wine` datasets, the findings above differ from those previously reported in Dang et al. (2015) for the MPE mixtures; this is likely due to different starting values. For example, in Dang et al. (2015), a solution was obtained using MPE mixtures which fit three components and misclassified only one observation; however, here the selected MPE mixture based on the BIC was a four-component model, likely due to different starting values (one of the MPE mixture fits also only had one misclassification; this was the third best model according to the BIC). Note that in the semi-supervised case as well (Table 5), the MSPE mixtures perform the best followed by the MPE mixtures. The estimates of the β_g parameters from the MSPE fit indicate heavy tails while `ghpcm` overfits the number of components.

Table 3 Performance comparison of four mixture model families on real data for the unsupervised scenarios. Sample size, dimensionality, and the number of known groups (i.e., classes) are in parentheses following each dataset name

Data	MSPE	MPE	`ghpcm`	`gpcm`
Body ($n = 507$, $p = 24$, $G = 2$)	**0.94** (2; −18681)	0.59 (4; −18670)	0.42 (4; −17340)	0.48 (4; −18842)
Coffee ($n = 43$, $p = 12$, $G = 2$)	**1** (2; −1373)	0.38 (3; −1297)	0.27 (4; −497)	0.38 (3; −1311)
Female voles ($n = 86$, $p = 7$, $G = 2$)	**0.91** (2; −1327)	0.8 (3; −1304)	0.82 (2; −1350)	**0.91** (2; −1317)
Olive ($n = 572$, $p = 8$, $G = 3$)	**0.7** (4; −5271)	0.69 (4; −5297)	**0.7** (4; −5230)	0.67 (4; −5437)
Penguins ($n = 342$, $p = 4$, $G = 3$)	**0.96** (3; −2532)	0.81 (4; −2502)	0.95 (3; −2449)	0.76 (4; −2497)
Wine ($n = 178$, $p = 13$, $G = 3$)	**0.92** (3; −5474)	0.68 (4; −5438)	0.8 (4; −4833)	0.8 (4; −5376)

For each implementation, the ARI and the number of selected components as well as the BIC are provided in parentheses.
The ARI for the comparator(s) with superior performance is bolded

Table 4 Classification tables for the cluster analysis of each model and dataset

Dataset	MSPE					MPE				ghpcm				gpcm			
Body		**1**	**2**			**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
	Women	256	4			200	0	4	56	133	9	118	0	148	0	109	3
	Men	4	243			1	167	77	2	10	117	2	118	0	142	5	100
Coffee		**1**	**2**			**1**	**2**	**3**		**1**	**2**	**3**	4	**1**	**2**	**3**	
	Arabica	36	0			22	14	0		19	11	0	6	22	14	0	
	Robusta	0	7			0	0	7		0	0	7	0	0	0	7	
Fvoles		**1**	**2**			**1**	**2**	**3**		**1**	**2**			**1**	**2**		
	californicus	41	0			35	0	6		38	3			41	0		
	ochrogaster	2	43			2	43	0		1	44			2	43		
Olive		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
	Southern Italy	217	0	106	0	217	0	106	0	222	0	101	0	196	0	127	0
	Sardinia	0	0	0	98	0	0	0	98	0	0	0	98	0	0	0	98
	Northern Italy	0	151	0	0	0	150	0	1	0	149	0	2	0	151	0	0
Penguins		**1**	**2**	**3**		**1**	**2**	**3**	**4**	**1**	**2**	**3**		**1**	**2**	**3**	**4**
	Adelie	149	0	**2**		150	1	0	0	151	0	0		0	96	1	54
	Chinstrap	3	0	65		4	64	0	0	6	0	62		0	5	63	0
	Gentoo	0	123	0		0	0	63	60	0	123	0		123	0	0	0
Wine		**1**	**2**	**3**		**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
	Barolo	0	59	0		0	50	0	9	57	0	0	2	0	54	0	5
	Grignolino	66	0	5		5	0	46	20	0	1	47	23	57	0	1	13
	Barbera	0	0	48		48	0	0	0	0	48	0	0	0	0	46	2

Table 5 Median ARI values along with first and third quartiles in parentheses for the four different models for each dataset for the semi-supervised runs

Data	MSPE	MPE	ghpcm	gpcm
Body ($n = 507$, $p = 24$, $G = 2$)	**0.94 (0.93, 0.95)**	**0.94 (0.93, 0.95)**	0 (0, 0.68)	0.92 (0.9, 0.94)
Coffee ($n = 43$, $p = 12$, $G = 2$)	**1 (1, 1)**	**1 (1, 1)**	1 (0, 1)	1 (0, 1)
Female voles ($n = 86$, $p = 7$, $G = 2$)	**0.94 (0.88, 0.94)**	**0.94 (0.88, 0.94)**	**0.94 (0.88, 0.94)**	**0.94 (0.88, 0.94)**
Olive ($n = 572$, $p = 8$, $G = 3$)	1 (0.99, 1)	**1 (1, 1)**	0.99 (0.98, 0.99)	**1 (1, 1)**
Penguins ($n = 342$, $p = 4$, $G = 3$)	**0.96 (0.95, 0.97)**	**0.96 (0.95, 0.97)**	**0.96 (0.95, 0.97)**	**0.96 (0.95, 0.97)**
Wine ($n = 178$, $p = 13$, $G = 3$)	**1 (1, 1)**	0.98 (0.98, 0.98)	0.8 (0, 0.85)	0.91 (0.83, 0.91)

Size, dimensionality, and the number of known groups (i.e., classes) are in parentheses following each dataset name.
The ARI for the comparator(s) with superior performance is bolded

It is interesting to note that while the BIC is used to select one model from a family of models within a family, there is no guarantee that this model, whether within a family or between families, provides the most superior clustering performance.

4.5 Semi-supervised Classification

Using the same datasets as in the unsupervised classification case, semi-supervised classification is now considered. Following McNicholas (2010), the (observed) likelihood in the model-based classification framework can be written

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\pi_g f(\mathbf{x}_i|\boldsymbol{\theta}_g)]^{z_{ig}} \prod_{j=k+1}^{n} \sum_{h=1}^{G} [\pi_h f(\mathbf{x}_j|\boldsymbol{\theta}_h)],$$

where the first k observations have known component memberships (i.e., labels), $f(\mathbf{x}_i|\boldsymbol{\theta}_g)$ is the gth component density, and π_g and z_{ig} have the usual interpretations.

For each dataset previously considered, we take 25 labelled/unlabelled splits with 25% supervision. In Table 5, we display the median ARI values along with the first and third quartiles over the 25 splits. Generally, and as one would expect, performance in the semi-supervised scenarios was found to be better than in the fully unsupervised scenarios. In the case of the `wine` data, as mentioned in Section 4.4, the MSPE mixtures clearly exhibit the best performance (perfect classification), followed by the MPE mixtures. On the `fvoles`, `olive`, and `penguins` datasets, all comparators performed similarly. For the `penguins` dataset, the highest ARI was achieved by the two power exponential based mixtures (only one misclassification based on 25% supervision). On the other hand, on the `coffee` dataset, all models performed well although `gpcm` and `ghpcm` had some runs with poor fits. On the `body` dataset, the two power exponential-based mixtures performed best, with the highest ARI achieved with three misclassifications overall (75% unlabelled), followed by `gpcm`.

5 Discussion

A multivariate skewed power exponential distribution is introduced that is well suited for density estimation purposes for a wide range of data with non-Gaussian clusters. The family of 16 MSPE mixtures presented herein allow for robust mixture models for model-based clustering on skewed as well as symmetric components. These models can model components with varying levels of peakedness and tail-weight (light, heavy, Gaussian) simultaneously with skewness. As a result, these models are well suited to model heterogeneous data with non-Gaussian components.

In addition to these properties, special cases of the MSPE distribution include the skew-normal distribution among others. The performance of such mixtures for clustering is investigated on a wide range of simulated scenarios—on heavy-tailed, light-tailed, Gaussian, and skewed components, and combinations thereof—and on real data of varying dimensionalities and sizes commonly used for illustrating clustering and classification. At present, model selection is performed using the BIC and, although this performs well in most cases, it is by no means perfect and alternative criteria could be considered in more detail. Standard errors of parameter estimates were not considered herein; this could be implemented via the standard information-based

method to obtain the asymptotic covariance matrix of the estimates or via bootstrapping (Basford et al., 1997; Lin et al., 2014).

Through simulations, we showed scenarios where such skewed mixtures are comparative to or better than widely used elliptical mixture models (mixtures of Gaussians) or skewed mixture models gaining increasing attention (mixtures of generalized hyperbolic distributions). When looking at real datasets, we compared in the context of both unsupervised classification (i.e., clustering) and semi-supervised classification. On these, the MSPE model performed just as well or better on most of the investigated datasets compared to three other mixture model families/algorithms. The analysis on the real datasets in the unsupervised case displayed some possible weaknesses, which may be related to initializations or the choice of BIC as the model selection criterion. Performance improved substantially for some of the datasets when a small level of supervision was introduced.

There are numerous areas of possible future work. One such area would be to consider a mixture of factor analyzers with the MSPE distribution for high-dimensional data. A matrix variate extension, in a similar manner to Gallaugher and McNicholas (2018), might also be interesting for modeling three-way data.

Appendix. Supporting Information

Parameter Recovery

A three-component mixture is simulated with 500 observations in total. Group sample sizes are sampled from a multinomial distribution with mixing proportions $(0.2, 0.34, 0.46)'$. The first component is simulated from a heavy-tailed three-dimensional MSPE distribution with $\boldsymbol{\mu}_1 = (12, 14, 0)'$, $\beta_1 = 0.85$, and $\boldsymbol{\psi}_1 = (-5, -10, 0)'$. The second component is simulated with $\boldsymbol{\mu}_2 = (-3, -10, 0)'$, $\beta_2 = 0.9$, and $\boldsymbol{\psi}_2 = (15, 10, 0)'$. The third component is simulated with light tails with $\boldsymbol{\mu}_3 = (3, 1, 0)'$, $\beta_3 = 2$, and $\boldsymbol{\psi}_3 = \boldsymbol{\psi}_2$. Lastly, the scale matrices

Table 6 Parameter recovery of cluster-specific means and covariance matrices (rounded to two decimals)

Parameter	Average generated	Average estimated	Mean squared error
Mean_1	$(11.46, 12.14, -0.06)'$	$(11.45, 12.08, -0.03)'$	$(0, 0.02, 0.01)'$
Mean_2	$(-1.96, -8.99, 0.27)'$	$(-1.96, -8.97, 0.22)'$	$(0, 0.01, 0.01)'$
Mean_3	$(3.50, 1.47, 0.14)'$	$(3.55, 1.51, 0.14)'$	$(0, 0.00, 0.00)'$
Covariance_1	$\begin{bmatrix} 3.35 & -1.67 & 2.21 \\ -1.67 & 3.69 & -0.98 \\ 2.21 & -0.98 & 4.87 \end{bmatrix}$	$\begin{bmatrix} 3.90 & -2.41 & 2.76 \\ -2.41 & 5.00 & -1.59 \\ 2.76 & -1.59 & 6.90 \end{bmatrix}$	$\begin{bmatrix} 0.47 & 0.75 & 0.47 \\ 0.75 & 2.10 & 0.52 \\ 0.47 & 0.52 & 4.62 \end{bmatrix}$
Covariance_2	$\begin{bmatrix} 1.86 & -1.52 & 1.54 \\ -1.52 & 4.92 & -0.90 \\ 1.54 & -0.90 & 3.99 \end{bmatrix}$	$\begin{bmatrix} 2.31 & -2.22 & 2.06 \\ -2.22 & 6.40 & -1.35 \\ 2.06 & -1.35 & 5.46 \end{bmatrix}$	$\begin{bmatrix} 0.25 & 0.62 & 0.35 \\ 0.62 & 2.62 & 0.33 \\ 0.35 & 0.33 & 2.38 \end{bmatrix}$
Covariance_3	$\begin{bmatrix} 0.42 & -0.36 & 0.35 \\ -0.36 & 1.11 & -0.21 \\ 0.35 & -0.21 & 0.91 \end{bmatrix}$	$\begin{bmatrix} 0.48 & -0.36 & 0.38 \\ -0.36 & 1.19 & -0.22 \\ 0.38 & -0.22 & 0.97 \end{bmatrix}$	$\begin{bmatrix} 0 & 0.00 & 0.00 \\ 0 & 0.01 & 0.00 \\ 0 & 0.00 & 0.01 \end{bmatrix}$

were common to all three components with diag $(\mathbf{\Delta}_g) = (4, 3, 1)'$ and

$$\mathbf{\Gamma}_g = \begin{pmatrix} 0.36 & 0.48 & -0.80 \\ -0.80 & 0.60 & 0.0 \\ 0.48 & 0.64 & 0.6 \end{pmatrix},$$

for $g = 1, 2, 3$. The simulated components were well separated to show parameter recovery. MSPE was run on 100 such datasets and perfect classification obtained on this well separated data each time. Note that there is an identifiability issue with individual parameter estimates (different combinations of individual parameter estimates yield the same fit) and closed form equations for overall mean and variance are not available. Hence, we demonstrate parameter recovery of overall cluster-specific mean and covariances in Table 6 by comparing estimates from data simulated (via a Metropolis-Hastings rule) using individual parameter estimates from the GEM fit. Clearly, the estimates are overall close to the generated values.

Performance on Additional Datasets

In addition to those considered in the main body of the manuscript, we also considered the following data available through various R packages:

Wine Dataset The expanded twenty seven variable `wine` dataset, obtained from `pgmm`, has 27 different measurements of chemical aspects of 178 Italian wines (three types of wine).

Iris Dataset The `iris` dataset (included with R) consists of 150 observations, 50 each of 3 different species of iris. There are four different variables that were measured, namely the petal length and width and the sepal length and width.

Swiss Banknote Dataset The Swiss `banknote` dataset, obtained from `MixGHD` (Tortora et al., 2018) looked at 6 different measurements from 100 genuine and 100 counterfeit banknotes. The measurements were length, length of the diagonal, width of the right and left edges, and the top and bottom margin widths.

Crabs Dataset The `crabs` dataset, obtained from `MASS` (Venables & Ripley, 2002), contains 200 observations with 5 different variables that measure characteristics of crabs. There were 100 males and 100 females, and two different species of crabs, orange, and blue. This creates four different groups of crabs based on gender/species combinations.

Bankruptcy Dataset The `bankruptcy` dataset, obtained from `MixGHD`, looked at the ratio of retained earnings to total assets, and the ratio of earnings before interests and taxes to total assets of 33 financially sound and 33 bankrupt American firms.

Yeast Dataset A subset of the `yeast` dataset from Nakai and Kanehisa (1991, 1992) sourced through the `MixSAL` package (Franczak et al., 2018) is also used. There are measurements on three variables: McGeoch's method for signal sequence recognition, the score of the ALOM membrane spanning region prediction program, and the score of discriminant

Table 7 Performance comparison of four mixture model families on real data for the unsupervised scenarios

Data	MSPE	MPE	ghpcm	gpcm
Banknote ($n=200$, $p=6$, $G=2$)	0.86 (3; −2681)	0.68 (4; −2651)	**0.98** (2; −2700)	0.68 (4; −2652)
Bankruptcy ($n=66$, $p=2$, $G=2$)	0.56 (4; −235)	0.53 (3; −228)	0 (3; −236)	**0.58** (3; −263)
Crabs ($n=200$, $p=5$, $G=4$)	0.67 (3; 72)	**0.86** (4; 92)	0.69 (3; 93)	0.61 (3; 63)
Diabetes ($n=145$, $p=3$, $G=3$)	0.44 (2; −479)	**0.66** (3; −487)	0.4 (2; −465)	**0.66** (3; −483)
Iris ($n=150$, $p=4$, $G=3$)	0.57 (2; −799)	**0.92** (3; −795)	0.7 (4; −582)	0.57 (2; −791)
Wine ($n=178$, $p=27$, $G=3$)	0.83 (3; −12179)	0.83 (3; −11982)	0.81 (4; −8665)	**1** (3; −11897)
Yeast ($n=626$, $p=3$, $G=2$)	**0.49** (3; −5042)	0.4 (4; −5028)	−0.04(2; −4863)	0.4 (4; −5053)

Sample size, dimensionality, and the number of known groups (i.e., classes) are in parentheses following each dataset name. For each implementation, the ARI and the number of selected components as well as the BIC are provided in parentheses

analysis of the amino acid content of vacuolar and extracellular proteins along with the possible two cellular localization sites, CYT (cytosolic or cytoskeletal) and ME3 (membrane protein, no N-terminal signal) for the proteins.

Diabetes Dataset The `diabetes` dataset, obtained from `mclust` (Fraley et al., 2012), considered 145 non-obese adult patients with different types of diabetes classified as normal, overt, and chemical. There were three measurements, the area under the plasma glucose curve, the area under the plasma insulin curve, and the steady-state plasma glucose.

Unsupervised Classification Unsupervised classification, i.e., clustering, is performed on the (scaled) datasets mentioned above using the same comparison distributions as on the simulated data. The ARI and the number of groups chosen by the BIC are shown in Table 7.

The `banknote` data is interesting in that while the two elliptical mixtures, MPE and Gaussian mixture models, split the counterfeit and genuine banknotes into four different groups with the same classification overall (Table 8), the selected MSPE model fits three components splitting the counterfeit banknotes into a larger and a smaller component, while the `ghpcm` model splits the observations into two groups only. We see that, for the `crabs`, the MPE distribution exhibits the best performance (with eight "blue males" being misclassified into a different component) while the other three methods choose three components; however, the clusters found are a little different (Table 8). For example, the MSPE model perfectly separates one species of crab from the other; however, for the "blue" species, it does not differentiate between the sexes. For the second species, there are only four misclassifications for differentiating the sexes. The `ghpcm` model has a similar fit to the MSPE model (species separated nicely but not sexes), but with two fewer misclassifications. The `gpcm` had a poorer fit compared to known labels, clustering sex better than species. The `bankruptcy` data shows some interesting results. The `ghpcm` model fits three clusters to the data, with two small clusters of two and three observations, respectively, with poor performance compared to known labels (Table 8). The other three comparators performed somewhat similarly with MSPE fitting four components (including one with eight tightly clustered points). For the `diabetes` data, the two selected skewed mixtures under-fit the number of components (seem to combine the normal and chemical classes but able to differentiate from the overt class) as compared to the elliptical mixtures, which fare better and

Table 8 Classification tables for the cluster analysis of each model and dataset

Dataset	MSPE					MPE				ghpcm				gpcm			
Banknote		1	2	3		1	2	3	4	1	2			1	2	3	4
	Counterfeit	0	85	15		85	0	0	15	100	0			85	0	0	15
	Genuine	99	0	1		0	75	24	1	1	99			0	75	24	1
Bankruptcy		1	2	3	4	1	2	3		1	2	3		1	2	3	
	Bankruptcy	30	0	3	0	0	21	12		31	0	2		0	22	11	
	Financially sound	0	15	10	8	28	5	0		30	3	0			29	4	0
Crabs		1	2	3		1	2	3	4	1	2	3		1	2	3	
	Blue males	50	0	0		8	0	0	42	50	0	0		38	12	0	
	Blue females	50	0	0		49	0	1	0	50	0	0		0	49	1	
	Orange males	0	50	0		0	50	0	0	0	50	0		50	0	0	
	Orange females	0	4	46		0	2	48	0	0	2	48		2	0	48	
Diabetes		1	2			1	2	3		1	2			1	2	3	
	Chemical	34	2			9	26	1		34	2			9	26	1	
	Normal	76	0			72	4	0		76	0			72	4	0	
	Overt	3	30			0	6	27		5	28			0	6	27	
Iris		1	2			1	2	3		1	2	3	4	1	2		
	setosa	0	50			0	50	0		0	0	29	21	0	50		
	versicolor	50	0			4	0	46		46	4	0	0	50	0		
	virginica	50	0			50	0	0		5	45	0	0	50	0		
Wine		1	2	3		1	2	3		1	2	3	4	1	2	3	
	Barolo	58	1	0		58	1	0		0	59	0	0	0	59	0	
	Grignolino	7	62	2		7	62	2		63	1	0	7	71	0	0	
	Barbera	0	0	48		0	0	48		0	0	29	19	0	0	48	
Yeast		1	2	3		1	2	3	4	1	2			1	2	3	4
	CYT	354	33	76		319	3	100	41	291	172			307	5	128	23
	ME3	10	152	1		23	135	0	5	155	8			20	135	0	8

similarly. Moreover, the estimates of skewness were not trivial, and both groups had heavier tails in the MSPE fit. On the other hand, for the MPE fit, the tails were approximately Gaussian (common $\beta_g \approx 1$). For the `yeast` dataset, apart from the `ghpcm` mixtures, the other three mixtures overfit the number of components; however, the `ghpcm` mixture clustering was not meaningful compared to known labels. For the `iris` data, only the MPE mixtures' selected model had three components. While for the expanded 27-dimensional `wine` dataset, the `gpcm` mixtures perform best with perfect classification. Interestingly, the `gpcm` mixtures have poorer performance relatively in the semi-supervised fits on these data. Note that this phenomenon, whereby cluster analysis can obtain better results compared to using semi-supervised classification, has been noted before, e.g., by Vrbik and McNicholas (2015) and Gallaugher and McNicholas (2019).

The relative performance of the MPE versus MSPE mixtures in Table 9 suggests that there are cases in which using these skewed mixtures might not be ideal and could be a possible subject of future work.

Semi-supervised Classification For each dataset, we take 25 labelled/unlabelled splits with 25% supervision. In Table 9, we display the median ARI values along with the first and third quartiles over the 25 splits. For the most part, as found in the main text of this manuscript, performance in the semi-supervised scenarios was better than in the fully unsupervised scenarios. Performance across the four comparators was also quite comparable with few exceptions. For the `yeast` data, both in the unsupervised and semi-supervised context, the MSPE mixtures performed the best. Similarly, for the `diabetes` data, the MSPE mixtures perform the best, with `gpcm` mixtures having the most inferior classification performance. On the 27 dimensional `wine` data, the MPE mixtures performed well with MSPE mixtures having more variability in ARI across the runs. For the `iris` dataset, the MPE and `ghpcm` models showed the best overall performance while, for the `banknote` dataset, all algorithms exhibit similar performance. For the bankruptcy data, the `gpcm` algorithm performed the poorest while, for `crabs`, the `gpcm` algorithm performed the best along with `ghpcm` and MPE mixtures which had similar performance. For the `crabs` dataset, although the MSPE models had poorer classification compared to the other three mixtures, the performance was still close to the other mixture distributions.

A reviewer noted that using mixtures of canonical fundamental skew t (CFUST) distributions (Lee & McLachlan, 2016), one can obtain an ARI close to 1 with there being only one

Table 9 Median ARI values along with first and third quartiles in parentheses for the four different models for each dataset for the semi-supervised runs

Data	MSPE	MPE	ghpcm	gpcm
Banknote ($n=200$, $p=6$, $G=2$)	0.97 (0.97, 0.97)	**0.97 (0.97, 1)**	0.97 (0.97, 0.97)	0.97 (0.97, 0.97)
Bankruptcy ($n=66$, $p=2$, $G=2$)	0.77 (0.63, 0.84)	**0.77 (0.7, 0.84)**	0.77 (0.7, 0.77)	0.51 (0.4, 0.77)
Crabs ($n=200$, $p=5$, $G=4$)	0.8 (0.78, 0.83)	0.85 (0.83, 0.87)	0.85 (0.82, 0.86)	**0.86 (0.83, 0.88)**
Diabetes ($n=145$, $p=3$, $G=3$)	**0.74 (0.7, 0.79)**	0.73 (0.68, 0.79)	0.73 (0.7, 0.79)	0.68 (0.67, 0.73)
Iris ($n=150$, $p=4$, $G=3$)	0.9 (0.89, 0.92)	0.92 (0.9, 0.93)	**0.92 (0.9, 0.95)**	0.9 (0.87, 0.93)
Wine ($n=178$, $p=27$, $G=3$)	0.91 (0.87, 0.95)	**0.95 (0.95, 0.96)**	0.93 (0, 0.95)	0.49 (0.44, 0.62)
Yeast ($n=626$, $p=3$, $G=2$)	**0.84 (0.83, 0.86)**	0.78 (0.75, 0.81)	0.74 (0.71, 0.76)	0.81 (0.78, 0.83)

Size, dimensionality, and the number of known groups (i.e., classes) are in parentheses following each dataset name.
The ARI for the comparator(s) with superior performance is bolded

misclassification. However, we were unable to obtain this solution from CFUST; perhaps due to different initialization. For the semi-supervised runs, this solution could be obtained for all four comparator distributions (for the best out of 25 runs).

Acknowledgements This research was supported by the Natural Sciences and Engineering Research Council of Canada through their Discovery Grants program for Dang, Browne, and McNicholas, the Banting Postdoctoral Fellowship for Gallaugher, as well as the Canada Research Chairs program for McNicholas.

Code and Data Accessibility All data used here are publicly available; references have been provided within the bibliography. An implementation of mixtures of skewed power exponential distributions and mixtures of power exponential distributions is available as the R package `mixSPE` (Dang et al., 2021).

Declarations

Conflict of Interest The authors declare no competing interests.

References

Absil, P.-A., Mahony, R., & Sepulchre, R (2009). Optimization algorithms on matrix manifolds. Princeton University Press.

Aitken, A. C. (1926). On Bernoulli's numerical solution of algebraic equations. In *Proceedings of the royal society of edinburgh* (pp. 289–305).

Andrews, J. L., & McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing*, *22*(5), 1021–1029.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, *46*(2), 199–208.

Azzalini, A., & Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, *83*, 715–726.

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821.

Basford, K., Greenway, D., McLachlan, G., & Peel, D. (1997). Standard errors of fitted component means of normal mixtures. *Computational Statistics*, *12*(1), 1–18.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, *41*(3–4), 561–575.

Böhning, D., & Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, *40*(4), 641–663.

Bouveyron, C., & Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, *71*, 52–78.

Branco, M. D., & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, *79*(1), 99–113.

Browne, R. P., & McNicholas, P. D. (2014). Orthogonal Stiefel manifold optimization for eigen-decomposed covariance parameter estimation in mixture models. *Statistics and Computing*, *24*(2), 203–210.

Browne, R. P., & McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, *43*(2), 176–198.

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

Cho, D., & Bui, T. D. (2005). Multivariate statistical modeling for image denoising using wavelet transforms. *Signal Processing: Image Communication*, *20*(1), 77–89.

da Silva Ferreira, C., Bolfarine, H., & Lachos, V. H. (2011). Skew scale mixtures of normal distributions: Properties and estimation. *Statistical Methodology*, *8*(2), 154–171.

Dang, U. J., Browne, R. P., & McNicholas, P. D. (2015). Mixtures of multivariate power exponential distributions. *Biometrics*, *71*(4), 1081–1089.

Dang, U. J., Browne, R. P., Gallaugher, M. P., & Band McNicholas, P.D. (2021). mixSPE: Mixtures of power exponential and skew power exponential distributions for use in model-based clustering and classification. R package version 0.9.1.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, *56*(3), 463–474.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, *39*(1), 1–38.

DiCiccio, T. J., & Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *Journal of the American Statistical Association*, *99*(466), 439–450.

Flury, B. (2012). Flury: data sets from Flury, 1997. R package version 0.1–3.

Forina, M., & Tiscornia, E. (1982). Pattern-recognition methods in the prediction of italian olive oil origin by their fatty-acid content. *Annali di Chimica*, *72*(3-4), 143–155.

Fraley, C., Raftery, A. E., Murphy, T. B., & Scrucca, L. (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical report 597, Department of statistics, university of Washington, Seattle, Washington.

Franczak, B. C., Browne, R. P., & McNicholas, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(6), 1149–1157.

Franczak, B. C., Browne, R. P., McNicholas, P. D., & Burak, K. L. (2018). MixSAL: Mixtures of multivariate shifted asymmetric Laplace (SAL) distributions. *R package version*, *1*, 0.

Gallaugher, M. P. B., & McNicholas, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, *80*, 83–93.

Gallaugher, M. P. B., & McNicholas, P. D. (2019). On fractionally-supervised classification: Weight selection and extension to the multivariate t-distribution. *Journal of Classification*, *36*(2), 232–265.

Gómez, E., Gomez-Viilegas, M. A., & Marin, J. M. (1998). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, *27*(3), 589–600.

Hartigan, J. A., & Wong, M. A. (1979). A *k*-means clustering algorithm. *Journal of the Royal Statistical Society: Series C*, *28*(1), 100–108.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, *8*(3), 431–444.

Horst, A. M., Hill, A. P., & Gorman, K.B. (2020). Palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Hunter, D. R., & Lange, K. (2000). Rejoinder to discussion of optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, *9*(1), 52–59.

Hunter, D. R., & Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, *58*(1), 30–37.

Hurley, C. (2012). gclus: clustering Graphics. R package version 1.3.1.

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

Lee, S., & McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, *24*(2), 181–202.

Lee, S. X., & McLachlan, G. J. (2016). Finite mixtures of canonical fundamental skew *t*-distributions: The unification of the restricted and unrestricted skew t-mixture models. *Statistics and Computing*, *26*(3), 573–589.

Lin, T.-I. (2010). Robust mixture modeling using multivariate skew *t* distributions. *Statistics and Computing*, *20*(3), 343–356.

Lin, T.-I., Ho, H. J., & Lee, C -R. (2014). Flexible mixture modelling using the multivariate skew-*t*-normal distribution. *Statistics and Computing*, *24*(4), 531–546.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics* (pp. 1–163).

Lindsey, J. K. (1999). Multivariate elliptically contoured distributions for repeated measurements. *Biometrics*, *55*(4), 1277–1280.

McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference*, *140*(5), 1175–1181.

McNicholas, P. D. (2016a). *Mixture model-based classification*. Boca Raton: Chapman & Hall/CRC Press.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, *33*(3), 331–373.

McNicholas, P. D., & Murphy, T. B. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, *18*(3), 285–296.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., & Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. *Computational Statistics & Data Analysis*, *54*(3), 711–723.

McNicholas, P. D., ElSherbiny, A., McDaid, A. F., & Murphy, T. B. (2022). pgmm: Parsimonious Gaussian Mixture Models. R package version 1.2.6-2.

McNicholas, S. M., McNicholas, P. D., & Browne, R.P. (2017). A mixture of variance-gamma factor analyzers. In *Big and complex data analysis* (pp 369–385*).* Springer international publishing, Cham.

Morris, K., & McNicholas, P. D. (2013). Dimension reduction for model-based clustering via mixtures of shifted asymmetric Laplace distributions. *Statistics and Probability Letters*, *83*(9), 2088–2093.

Murray, P. M., Browne, R. B., & McNicholas, P. D. (2014). Mixtures of skew-t factor analyzers. *Computational Statistics and Data Analysis*, *77*, 326–335.

Murray, P. M., Browne, R. B., & McNicholas, P. D. (2017). Hidden truncation hyperbolic distributions, finite mixtures thereof, and their application for clustering. *Journal of Multivariate Analysis*, *161*, 141–156.

Nakai, K., & Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. Proteins: Structure. *Function, and Bioinformatics*, *11*(2), 95–110.

Nakai, K., & Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, *14*(4), 897–911.

O'Hagan, A., Murphy, T. B., Gormley, I. C., McNicholas, P. D., & Karlis, D. (2016). Clustering with the multivariate normal inverse gaussian distribution. *Computational Statistics and Data Analysis*, *93*, 18–30.

Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the *t* distribution. *Statistics and Computing*, *10*(4), 339–348.

Pocuca, N., Browne, R. P., & McNicholas, P.D. (2022). Mixture: Mixture models for clustering and classification. R package version 2.0.5.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, *9*, 386–396.

Streuli, H. (1973). Der heutige stand der kaffeechemie. In *6th international colloquium on coffee chemisrty* (pp. 61–72).

Subedi, S., & McNicholas, P. D. (2014). Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions. *Advances in Data Analysis and Classification*, *8*(2), 167–193.

Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analysers. *Neural Computation*, *11*(2), 443–482.

Tortora, C., ElSherbiny, A., Browne, R. P., Franczak, B. C., McNicholas, P. D., & Amos, D.D (2018). MixGHD: Model based clustering, classification and discriminant analysis using the mixture of generalized hyperbolic distributions. R package version 2.2.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S, fourth edn.* New York: Springer. ISBN 0-387-95457-0.

Verdoolaege, G., De Backer, S., & Scheunders, P. (2008). Multiscale colour texture retrieval using the geodesic distance between multivariate generalized Gaussian models. In *2008 15th IEEE international conference on image processing* (pp. 169–172).

Vrbik, I., & McNicholas, P. D. (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics and Data Analysis*, *71*, 196–210.

Vrbik, I., & McNicholas, P. D. (2015). Fractionally-supervised classification. *Journal of Classification*, *32*(3), 359–381.

Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. U.S. Naval personnel research activity, technical bulletin:65-15.

Zhu, X., Sarkar, S., & Melnykov, V. (2022). Mattransmix: An r package for matrix model-based clustering and parsimonious mixture modeling. *Journal of Classification*, *39*(1), 147–170.