Clustering with the Multivariate Normal Inverse Gaussian Distribution.

Adrian O'Hagan[1,*], Thomas Brendan Murphy[2], Isobel Claire Gormley[2], Paul McNicholas[3], Dimitris Karlis[4]

Abstract

Many model-based clustering methods are based on a finite Gaussian mixture model. The Gaussian mixture model implies that the data scatter within each group is elliptically shaped. Hence non-elliptical groups are often modeled by more than one component, resulting in model over-fitting. An alternative is to use a mean-variance mixture of multivariate normal distributions with an inverse Gaussian mixing distribution (MNIG) in place of the Gaussian distribution, to yield a more flexible family of distributions. Under this model the component distributions may be skewed and have fatter tails than the Gaussian distribution. The MNIG based approach is extended to include a broad range of eigen-decomposed covariance structures. Furthermore, MNIG models where the other distributional parameters are constrained is considered. The Bayesian Information Criterion is used to identify the optimal model and number of mixture components. The method is demonstrated on three sample data sets and a novel variation on the univariate Kolmogorov-Smirnov test is used to assess goodness of fit.

Keywords:
Model-based clustering, multivariate normal inverse Gaussian distribution, **mclust**, information metrics, Kolmogorov-Smirnov goodness of fit

[☆]Supplementary material for this paper has been added to the online version of the manuscript.

[*]Corresponding author. adrian.ohagan@ucd.ie

[1]School of Mathematical Sciences, University College Dublin

[2]School of Mathematical Sciences and INSIGHT: The National Centre for Big Data Analytics, University College Dublin

[3]Department of Mathematics and Statistics, University of Guelph

[4]Department of Statistics, Athens University of Economics and Business

1. Introduction

Mixture models are a commonly employed tool in statistical modeling, in particular the mixture of multivariate Gaussian distributions that forms the basis of the model-based clustering package **mclust** (Fraley and Raftery, 1998, 1999) in **R** (R Development Core Team, 2012). The Gaussian mixture model implies that the data, within each group, have an elliptical scatter. Hence non-elliptical groups are often modeled by more than one component, resulting in over-fitting. This may render the clustering rule ambiguous, meaning the correct (lower) number of groups is not identified. Ultimately this can result in higher misclassification rates. Also, the Gaussian mixture model can struggle to accommodate clusters with heavy tails or outliers.

One solution to this problem is to apply mixtures of t distributions, whose heavier tails can guard against the influence of outliers (McLachlan and Peel, 2000; Andrews and McNicholas, 2011). However this approach still implies that the data are elliptically contoured within each group (Banfield and Raftery, 1993). To address this issue, mixtures of skew-normal or skew-t distributions can be used (Lin et al., 2007b,a; Cabral et al., 2012; Prates et al., 2013a; Vrbik and McNicholas, 2014). However, these distributions can prove numerically unstable in high-dimensional settings (Fruhwirth-Schnatter and Pyne, 2009).

Alternatively, data transformations prior to modeling can be used to reduce skew as much as possible (MacLean et al., 1976). Gottardo and Lo (2011) use the Box-Cox transformation as a precursor to fitting a mixture of t distributions to flow cytometry data. Unfortunately such a process is data and variable dependent, making automatic model fitting and selection difficult or impossible. Interpretability of results also suffers, since the data are no longer modeled in their original units.

A potential antidote to fitting a more complex model is to retain the Gaussian mixture model but merge mixture components from the initial model fit to produce an updated clustering solution (Hennig, 2010). This exploits the Gaussian mixture model's ability to provide a robust density estimate for the data while addressing its propensity for over-fitting.

The multivariate normal inverse Gaussian (MNIG) is a mean-variance mixture of multivariate Gaussians and is a special case of the generalised hyperbolic mixture (McNicholas et al., 2013). This yields a more flexible family of mixture distributions, which may be skewed and have fatter tails than a Gaussian distribution (Karlis and Santourian, 2008). The MNIG based approach is extended to the full range of eigenvalue decomposed covariance structures, as considered in **mclust**. Furthermore, the family of MNIG models where distributional parameters are constrained, is considered. This can improve model parsimony in cases where clusters have similar shape properties. The Bayesian Information Criterion (BIC) (Schwarz, 1978) is used to

identify the optimal model and number of components. Disparities in clustering solutions under the mixture of MNIG and mixture of Gaussian (**mclust**) approaches are highlighted.

Section 2 presents the data sets used as motivating examples: the *Old Faithful* eruptions data, the *FLAME* flow cytometry data and the *Iris* data. Section 3 gives an account of the EM algorithm for fitting the mixture of MNIG distributions by maximum likelihood. Section 4 details a range of model diagnostics used to compare the competing clustering methods. BIC is used for model selection, metrics are used to compare clustering solutions and a goodness-of-fit test based on the Kolmogorov-Smirnov statistic is also detailed. Section 5 presents the results obtained for the motivating data sets, highlighting improvements in the clustering capability of a mixture of MNIG distributions over other mixture distributions. Section 6 summarizes the main findings and explores further avenues of investigation.

2. Illustrative data sets

Three data sets are used as motivating examples; these are described below.

Old Faithful data

The data are comprised of bivariate observations for 272 eruptions of the Old Faithful geyser in Yellowstone National Park (Azzalini and Bowman, 1990). Each observation records the eruption duration and the waiting duration until the next eruption, both measured in minutes. This is a classic test case for any clustering methodology because the data are multimodal. However, there are no "true" group labels available - different numbers of groups can be identified depending on the clustering rule applied (Hunter et al., 2007); see Figure 1.

Flow cytometry FLAME data

The *FLAME* flow cytometry data is the *090806A0minLymphocytes* sample from the T cell phosphorylation data set as analyzed in Pyne et al. (2009). The data comprise 4669 observations and 4 cell surface marker variables: SLP76, ZAP70, CD4 and CD45RA. The goal of collecting and analyzing these data is to identify sub-populations denoting contrasting cell surface regions in terms of response to fluorophore-conjugated reagents, since well-separated regions may have potential to act as distinct disease indicators. Figure 2 presents the pairs plots for the data across all variables. True group labels are not known, but the clustering labels identified by Pyne et al. (2009) using a mixture of multivariate skew distributions are available for comparative purposes.

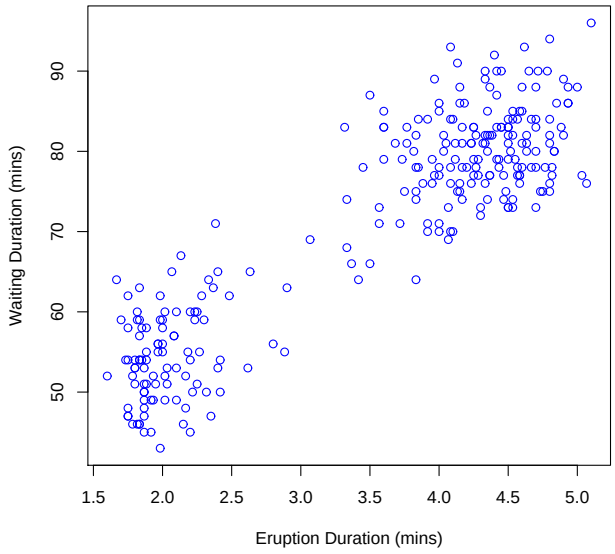


Figure 1: Scatter plot of the *Old Faithful* data.

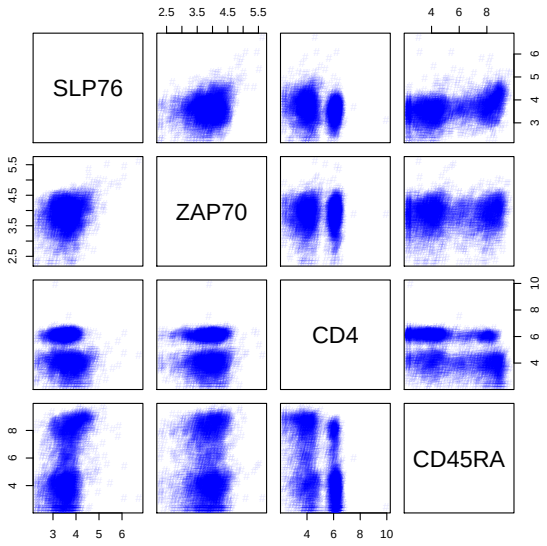


Figure 2: Scatter plot of the flow cytometry *FLAME* data.

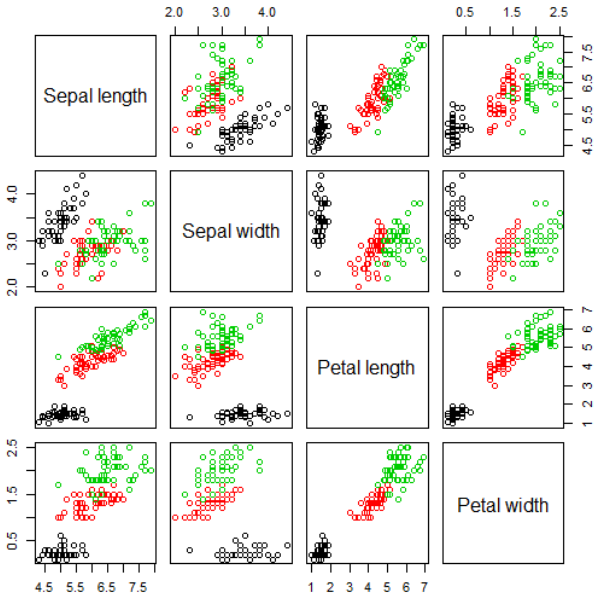


Figure 3: Pairs plot for Fisher's *Iris* data. *Setosa* irises are plotted in red, *Versicolor* in green and *Virginicas* in black.

Iris data

Fisher's *Iris* data (Fisher, 1936) contains the measurements of the variables sepal length, sepal width, petal length and petal width in centimeters for each of 50 *Setosa*, *Versicolor* and *Virginica* irises. Figure 3 provides the pairs plot for the data. In general the *Setosa* irises are strongly separated from the other species, but there is scope for misclassification error among the *Versicolor* and *Virginica* irises. A mixture of normal distributions is generally deemed a good choice for this data set. Hence it permits investigation of the behavior of the more complex mixture of MNIGs model when a simpler model is sufficient.

3. Model-based clustering

Model-based clustering, based on finite mixture models, provides a principled, statistical approach to determining the number of clusters present and how obser-

vations should be allocated to the available clusters (Fraley and Raftery, 2002). In the context of a mixture of G Gaussian distributions the covariance matrix can be expressed as $\Sigma_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$ (Banfield and Raftery, 1993; Bensmail and Celeux, 1996). This expression uses an eigenvalue decomposition into components controlling the volume (the scalar λ_g), orientation (the orthogonal matrix $\mathbf{D}_g$) and shape (the diagonal matrix $\mathbf{A}_g$) of the data cloud. Varying the status of the components of $\mathbf{\Sigma}_g$ between group-dependent and group-independent produces a range of covariance structures; this approach is implemented in **mclust**. Maximization steps for the eigen-decomposed covariance structures are given in Celeux and Govaert (1995).

An alternative is to use a multivariate normal inverse Gaussian distribution (Karlis and Santourian, 2008). The benefits are that the mixture component can be skewed and that it has fatter tails than the multivariate Gaussian distribution (Yu, 2011). The approach is extended to incorporate the range of eigen-decomposed covariance structures utilized in **mclust**. Throughout, the focus is on mixture components following the same form of the distribution function, though this is not the most general case.

3.1. Multivariate normal inverse Gaussian distribution: parametrization

Let a scalar quantity u_{ig}, termed the *mixing component*, be inverse Gaussian distributed $u_{ig} \sim IG(\tilde{\delta}_g, \tilde{\gamma}_g)$ where in the standard case $\tilde{\delta}_g = 1$ for all $g = 1, 2, \ldots G$. Hence

$$f(u_{ig}) = \frac{1}{\sqrt{2\pi}} \exp(\tilde{\gamma}_g) u_{ig}^{-\frac{3}{2}} \exp\left(-\frac{1}{2}\left[\frac{1}{u_{ig}} + \tilde{\gamma}_g^2 u_{ig}\right]\right). \tag{1}$$

Let the p-vector $\mathbf{x}_i$ be multivariate normal distributed such that $\mathbf{x}_i | u_{ig} \sim N(\tilde{\boldsymbol{\mu}}_g + u_{ig}\tilde{\boldsymbol{\beta}}_g, u_{ig}\tilde{\mathbf{\Sigma}}_g)$. The role of u_{ig}'s interaction with the scale matrix $\tilde{\mathbf{\Sigma}}_g$ is to control thickness of the tails of the distribution. This corresponds to the concept of variability being partitioned into intrinsic ($\tilde{\mathbf{\Sigma}}_g$) and mixing ($u_{ig}$) components. For data sets with heavy tails u_{ig} will take a value greater than 1. The role of $\tilde{\boldsymbol{\beta}}$ is to capture the skew of the data through the shift effect it has on the mean $\tilde{\boldsymbol{\mu}}$, compensating for the effects of the value taken by u_{ig} if required. It is important to note that, in the mixture of MNIG distributions the eigenvalue decomposition applies to the covariance conditional on the latent variable u. The conditional density of observation i, where the matrix of mixing components for all observations and groups is $\mathbf{U}$, is

$$f(\mathbf{x}_i | u_{ig}, z_{ig} = 1) = \frac{1}{(2\pi)^{\frac{p}{2}} |u_{ig}\tilde{\mathbf{\Sigma}}_g|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left[\mathbf{x}_i - (\tilde{\boldsymbol{\mu}}_g + u_{ig}\tilde{\boldsymbol{\beta}}_g)\right]' \left[u_{ig}\tilde{\mathbf{\Sigma}}_g\right]^{-1} \left[\mathbf{x}_i - (\tilde{\boldsymbol{\mu}}_g + u_{ig}\tilde{\boldsymbol{\beta}}_g)\right]\right\}. \tag{2}$$

This leads to the complete-data log-likelihood, $l_c(\boldsymbol{\theta}, \boldsymbol{\tau}, \mathbf{Z}, \mathbf{U}|\mathbf{X})$ where A is a constant:

$$l_c = A + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left\{\left[\log\,\tau_g + \tilde{\gamma}_g - \frac{1}{2}\left(\frac{1}{u_{ig}} + \tilde{\gamma}_g^2 u_{ig}\right)\right] - \left[\frac{1}{2}\log|u_{ig}\tilde{\boldsymbol{\Sigma}}_g|\right]\right\}$$

$$-\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\left\{\frac{1}{2u_{ig}}\left[(\mathbf{x}_i - (\tilde{\boldsymbol{\mu}}_g + u_{ig}\tilde{\boldsymbol{\beta}}_g))'(\tilde{\boldsymbol{\Sigma}}_g)^{-1}(\mathbf{x}_i - (\tilde{\boldsymbol{\mu}}_g + u_{ig}\tilde{\boldsymbol{\beta}}_g))\right]\right\}. \quad (3)$$

3.2. EM Algorithm for a mixture of MNIG distributions: the E-step

The complete-data log-likelihood (3) can be factorized in two parts, one of which involves the multivariate normal distribution parameters and the other for the inverse Gaussian parameters. The expected value of (3) is computed with respect to the distribution $P(\mathbf{U}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\tau}^{(t)})$. Defining the weight $w_{ig} = E\left[z_{ig}^{(t+1)}|\mathbf{x}_i\right]$ leads to the standard mixture models result:

$$w_{ig} = \tau_g f(\mathbf{x}_i|\hat{\tilde{\boldsymbol{\mu}}}_g, \hat{\tilde{\boldsymbol{\Sigma}}}_g, u_{ig}, \hat{\tilde{\boldsymbol{\beta}}}_g)/\sum_{g'=1}^{G}\tau_{g'} f(\mathbf{x}_i|\hat{\tilde{\boldsymbol{\mu}}}_{g'}, \hat{\tilde{\boldsymbol{\Sigma}}}_{g'}, u_{ig'}, \hat{\tilde{\boldsymbol{\beta}}}_{g'}). \quad (4)$$

where u_{ig} denotes the inverse Gaussian mixing variable for observation i when it belongs to component g. The expected values of u_{ig} and its inverse are also required. Initially it may appear that the expected values of $u_{ig}z_{ig}$ and $u_{ig}^{-1}z_{ig}$ are actually required. However, since z_{ig} can take only the values 0 or 1, then $u_{ig}z_{ig}$ and $u_{ig}^{-1}z_{ig}$ will be equal to u_{ig} and u_{ig}^{-1} if $z_{ig} = 1$ and equal to 0 if $z_{ig} = 0$. This greatly simplifies the calculations required in the E-step. Letting K denote a modified Bessel function of the second kind (Mechel, 1966), then Karlis and Santourian (2008) show that:

$$\begin{aligned}
s_{ig} &= E(u_{ig}|\mathbf{x}_i, \boldsymbol{\theta}_g) = \frac{\hat{\phi}_{ig}K_{\lambda*+1}(\hat{\phi}_{ig}\hat{\alpha}_g)}{\hat{\alpha}_g K_{\lambda*}(\hat{\phi}_{ig}\hat{\alpha}_g)},\\
t_{ig} &= E(u_{ig}^{-1}|\mathbf{x}_i, \boldsymbol{\theta}_g) = \frac{\hat{\alpha}_g K_{\lambda*-1}(\hat{\phi}_{ig}\,\hat{\alpha}_g)}{\hat{\phi}_{ig}K_{\lambda*}(\hat{\phi}_{ig}\hat{\alpha}_g)},\\
\hat{\phi}_{ig} &= \sqrt{1 + (\mathbf{x}_i - \hat{\tilde{\boldsymbol{\mu}}}_g)'\hat{\tilde{\boldsymbol{\Sigma}}}_g^{-1}(\,\mathbf{x}_i - \hat{\tilde{\boldsymbol{\mu}}}_g)},\\
\hat{\alpha}_g &= \sqrt{\hat{\tilde{\gamma}}^2 + \hat{\tilde{\boldsymbol{\beta}}}_g'\,\hat{\tilde{\boldsymbol{\Sigma}}}_g^{-1}\,\hat{\tilde{\boldsymbol{\beta}}}_g},
\end{aligned}$$

and $\lambda* = 0.5p + 0.5$

3.3. EM Algorithm for a mixture of MNIG distributions: the M-step

The maximum likelihood estimators (MLEs) for the parameters $\boldsymbol{\tau}_g$, $\tilde{\gamma}_g$, $\tilde{\boldsymbol{\mu}}_g$ and $\tilde{\boldsymbol{\beta}}_g$, are available in closed form, where $\sum_{i=1}^{n} w_{ig} = n_{w_g}$.

$$\hat{\boldsymbol{\tau}}_g = \sum_{i=1}^{n} w_{ig} / \sum_{i=1}^{n} \sum_{g=1}^{G} w_{ig} = n_{w_g}/n$$

$$\hat{\tilde{\gamma}}_g = n_{w_g} / \sum_{i=1}^{n} w_{ig} s_{ig}$$

$$\hat{\tilde{\boldsymbol{\mu}}}_g = \left(\frac{\sum_{i=1}^{n} w_{ig} t_{ig} \mathbf{x}_i}{\sum_{i=1}^{n} w_{ig}} - \frac{\sum_{i=1}^{n} w_{ig} \mathbf{x}_i}{\sum_{i=1}^{n} w_{ig} s_{ig}} \right) \left(\frac{\sum_{i=1}^{n} w_{ig} t_{ig}}{\sum_{i=1}^{n} w_{ig}} - \frac{\sum_{i=1}^{n} w_{ig}}{\sum_{i=1}^{n} w_{ig} s_{ig}} \right)^{-1}$$

$$\text{and} \qquad \hat{\tilde{\boldsymbol{\beta}}}_g = \sum_{i=1}^{n} w_{ig} (\mathbf{x}_i - \hat{\tilde{\boldsymbol{\mu}}}_g) / \sum_{i=1}^{n} w_{ig} s_{ig}.$$

Isolating the terms in the complete-data log-likelihood function (3) relevant to the scale matrix $\tilde{\boldsymbol{\Sigma}}_g$, denoted $l_c^{\tilde{\boldsymbol{\Sigma}}_g}$, the function to be maximized can be expressed as:

$$E(l_c^{\tilde{\boldsymbol{\Sigma}}_g}) = -\sum_{i=1}^{n} \operatorname{tr}\left(\mathbf{M}_{ig} \tilde{\boldsymbol{\Sigma}}_g^{-1} \right) + \sum_{i=1}^{n} w_{ig} \log |\tilde{\boldsymbol{\Sigma}}_g^{-1}|. \tag{5}$$

where

$$\mathbf{M}_{ig} = w_{ig} t_{ig} \left[(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_g)' \right] + w_{ig} s_{ig} \tilde{\boldsymbol{\beta}}_g \tilde{\boldsymbol{\beta}}_g' - w_{ig} \left[(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_g) \tilde{\boldsymbol{\beta}}_g' + \tilde{\boldsymbol{\beta}}_g (\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_g)' \right].$$

This allows $E(l_c^{\tilde{\boldsymbol{\Sigma}}_g})$ to be interpreted as a standard function that can be maximized using a fixed form for the scale matrix $\tilde{\boldsymbol{\Sigma}}_g$:

"The $p \times p$ symmetric matrix $\mathbf{C}$ minimizing $f(C) = \operatorname{tr}(\mathbf{B}\mathbf{C}^{-1}) + \alpha \log |\mathbf{C}|$, where $\mathbf{B}$ is a symmetric positive definite matrix and α is a positive real number, is $\mathbf{C} = \mathbf{B}/\alpha$." (Celeux and Govaert, 1995).

The first and second terms of the expression above for $\mathbf{M}_{ig}$ are clearly symmetric, and the remaining terms combine to make a further symmetric contribution. In addition, w_{ig} is guaranteed to be a positive real number. Hence, the regularity

conditions are met. Defining $f^*(C) = -\text{tr}(\mathbf{B}\mathbf{C}^{-1}) + \alpha \log|\mathbf{C}^{-1}|$, then $\mathbf{C}$ is the matrix that maximizes $f^*(C)$. Recognizing that $f^*(C)$ has identical structure to (5), with $\sum_{i=1}^{n} \mathbf{M}_{ig} \to \mathbf{B}$, $\sum_{i=1}^{n} w_{ig} \to \alpha$ and $\tilde{\boldsymbol{\Sigma}}_g \to \mathbf{C}$ then:

$$\hat{\tilde{\boldsymbol{\Sigma}}}_g = \left(\sum_{i=1}^{n} \mathbf{M}_{ig}\right)\left(\sum_{i=1}^{n} w_{ig}\right)^{-1} \tag{6}$$

Further details of the derivation of the result are provided in Celeux and Govaert (1995). Since the structural form of (5) does not depend on the specific scale matrix structure used, the result can be readily extended. Hence the MLE of $\tilde{\boldsymbol{\Sigma}}_g$ for a mixture of MNIG distributions under any of these scale matrix structures is available.

The parameter initialization procedure used follows Karlis and Santourian (2008) i.e. the parameters $\tilde{\beta}_g$ and $\tilde{\gamma}_g$ are set equal to 0 and 1 respectively for all g. A mixture of G multivariate Gaussian distributions is fitted with the covariance structure $\boldsymbol{\Sigma}_g$ matching the scale matrix structure $\tilde{\boldsymbol{\Sigma}}_g$, using **mclust**. The parameters $\tilde{\boldsymbol{\mu}}_g$, $\tilde{\boldsymbol{\Sigma}}_g$ are initialised to the **mclust** MLEs of $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$, as are the τ_g parameters.

3.4. Mixtures of MNIG distributions with parameter constraints

Two types of parameter constraint on the mixture of MNIGs are available. The parameter controlling skew, $\tilde{\boldsymbol{\beta}}_g$ can be set equal across groups ($\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}\ \forall\, g$) or set to 0 across groups ($\tilde{\boldsymbol{\beta}}_g = \mathbf{0}\ \forall\, g$) if skew is largely absent among the fitted clusters. The parameter $\tilde{\gamma}_g$ can also be set equal across groups ($\tilde{\gamma}_g = \tilde{\gamma}\ \forall\, g$). This leads to six possible formulations of MNIG distributions. All six variants are operable across the ten available scale matrix structures. The adapted M-step updates required by each constraint are detailed below.

$\underline{\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}}$: The M-step update for $\tilde{\boldsymbol{\beta}}$ depends on whether the scale matrix structure used varies or does not vary across groups, respectively denoted by $\tilde{\boldsymbol{\Sigma}}_g$ and $\tilde{\boldsymbol{\Sigma}}$:

$$\begin{aligned}
\hat{\tilde{\boldsymbol{\beta}}}^{\tilde{\boldsymbol{\Sigma}}_g} &= \left[\sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}\hat{\tilde{\boldsymbol{\Sigma}}}_g^{-1}(\mathbf{x}_i - \hat{\tilde{\boldsymbol{\mu}}}_g)\right]\left[\sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}s_{ig}\hat{\tilde{\boldsymbol{\Sigma}}}_g^{-1}\right]^{-1}, \\
\hat{\tilde{\boldsymbol{\beta}}}^{\tilde{\boldsymbol{\Sigma}}} &= \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}(\mathbf{x}_i - \hat{\tilde{\boldsymbol{\mu}}}_g)/\sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}s_{ig}.
\end{aligned}$$

In both cases, the M-step update for $\tilde{\boldsymbol{\mu}}_g$ is calculated as:

$$\hat{\tilde{\boldsymbol{\mu}}}_g = \sum_{i=1}^{n}(w_{ig}t_{ig}\mathbf{x}_i - w_{ig}\hat{\tilde{\boldsymbol{\beta}}})/\sum_{i=1}^{n} w_{ig}t_{ig}.$$

$\underline{\tilde{\boldsymbol{\beta}}_g = \mathbf{0}}$: This constraint simplifies the M-step updates required for $\tilde{\boldsymbol{\mu}}_g$ and $\tilde{\boldsymbol{\Sigma}}_g$:

$$\begin{aligned}
\hat{\tilde{\boldsymbol{\mu}}}_g &= \sum_{i=1}^{n} w_{ig}t_{ig}\mathbf{x}_i/\sum_{i=1}^{n} w_{ig}t_{ig}, \\
\hat{\tilde{\boldsymbol{\Sigma}}}_g &= \sum_{i=1}^{n} \mathbf{M}*_{ig}/\sum_{i=1}^{n} w_{ig} \qquad \text{where} \qquad \mathbf{M}*_{ig} = w_{ig}t_{ig}\left[(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_g)(\mathbf{x}_i - \tilde{\boldsymbol{\mu}}_g)'\right].
\end{aligned}$$

$\underline{\tilde{\gamma}_g = \tilde{\gamma}}$: This constraint simplifies the M-step to:

$$\hat{\tilde{\gamma}} = \sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}/\sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}s_{ig} = n/\sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}s_{ig}.$$

4. Model diagnostics

A range of model diagnostics are applied in order to perform model selection and to assess goodness of fit of the fitted models. If a "hard" clustering is required, observations are clustered into the group for which they have the maximum *a posteriori* membership probability.

4.1. The Bayesian information criterion (BIC)

The BIC is a long-established tool for model selection (Schwarz, 1978). Kass and Raftery (1995) and Wasserman (1995) document the relationship between Bayes factors and the BIC in model selection. Dasgupta and Raftery (1998) illustrate the use of BIC in mixture model selection specifically.

4.2. Information metrics

Since group information is unknown for the *Old Faithful* and *FLAME* data sets, none of the clustering methods considered can be assessed relative to the true data groupings. However a range of information metrics may still be used to assess the level of agreement between the clustering solutions. The metrics utilized are: the adjusted Rand index (Hubert and Arabie, 1985); the Mirkin metric (Mirkin and Cherny, 1970); the Van Dongen criterion (Van Dongen, 2000); and the "variation of information" (VI) (Meila, 2007).

The VI metric utilizes a "hard" mapping of group membership probabilities to labels. It is adapted to a "soft" version, VI_{soft} for use with the *Old Faithful* and *FLAME* data sets, where the group membership probabilities are used. This bestows greater reward on agreement between clusterings when both have high posterior probabilities of membership for the observations and groups in question. For instance, envisage two separate observations whose probabilities of belonging to a particular group under two competing clustering methods are $(0.51, 0.91)$ and $(0.9, 0.91)$. Under VI, both observations would undergo hard classification to the group in question under both clustering rules, and contribute the same value to the metric. Under VI_{soft}, the second observation would contribute significantly less to the metric. For two competing clustering solutions C_1 and C_2, $VI_{soft}(C_1, C_2)$ is calculated as:

$$VI_{soft}(C_1, C_2) = H(C_1) + H(C_2) - 2I(C_1, C_2)$$

The entropy of clustering solution C_j is denoted $H(C_j)$, $I(C_1, C_2)$ denotes the mutual information between clustering solutions C_1 and C_2 with corresponding group membership probability matrices $\mathbf{Z_1}$ and $\mathbf{Z_2}$, z_{ig}^j denotes the probability of observation i belonging to group g in clustering solution j and:

$$\begin{aligned} H(C_j) &= -\sum_{g=1}^{G}\left(\sum_{i=1}^{n} z_{ig}^j/n\right)\log\left(\sum_{i=1}^{n} z_{ig}^j/n\right), \\ \text{and} \qquad I(C_1, C_2) &= \sum_{g=1}^{G}\sum_{g*=1}^{G*}\frac{(\mathbf{Z_1}'\mathbf{Z_2})_{g,g*}}{n}\log\left(\frac{(\mathbf{Z_1}'\mathbf{Z_2})_{g,g*}}{n}\left(\sum_{i=1}^{n} z_{ig}^1/n\right)\left(\sum_{i=1}^{n} z_{ig}^2/n\right)\right). \end{aligned}$$

The suite of metrics that are computed on the "hard" mapping of observations to groups (adjusted Rand, Mirkin, Van Dongen and variation of information) will generally take values closely consistent with the soft version of variation of information for cases where the fit of the model to the data is good. However the soft version of variation of information is capable of providing superior insight to the quality of the clustering in cases where the model provides poor fit to the data. This has been illustrated for simulated data in the Supplementary Materials.

4.3. Multivariate Kolmogorov-Smirnov goodness of fit test

A Kolmogorov-Smirnov goodness of fit test is developed to assess model fit for multivariate data. The standard univariate Kolmogorov-Smirnov test statistic takes the form $D = \sup_x |F_n(x) - \hat{F}(x)|$ where $F_n(x)$ is the empirical cumulative distribution function (ECDF) and $\hat{F}(x)$ is the fitted model's CDF (Massey, 1951).

Calculation of the test statistic is straightforward in the univariate case since observations can be ordered directly from smallest to largest by magnitude. The ECDF for multivariate data can be expressed as:

$$F_n(x_1, x_2, \ldots, x_p) = P(X_1 < x_1, X_2 < x_2, \ldots, X_p < x_p).$$

That is to say, it is the proportion of observations that satisfy all p inequalities. There is no issue in principle in constructing such a multivariate ECDF. However difficulty arises in finding the maximum difference in CDFs to calculate a multivariate KS statistic; and in deriving its null distribution.

An alternative approach is to use log density values to measure fit. A model can be verified as providing good fit if its log density is concentrated in the same areas as the underlying data. This avoids the need to engage in high-dimensional numerical integration or Monte Carlo integration, since the dimensionality of the problem is reduced to the univariate density values. The procedure followed is:

1. Select the optimal model $\hat{F}$ by BIC for the original data $\mathbf{X}$.
2. Rank observations in $\mathbf{X}$ according to the magnitude of their log density under $\hat{F}$.
3. Plot the ECDF of the log density values of the original data $\mathbf{X}$ under $\hat{F}$.
4. Simulate new data $\mathbf{Y}$ from $\hat{F}$.
5. Rank the observations in $\mathbf{Y}$ according to the magnitude of their log density under $\hat{F}$.
6. Plot the ECDF of the log density values of the simulated data $\mathbf{Y}$ under $\hat{F}$ on the same axes as 3.
7. Repeat steps 4-6 for a large number of simulations (1000 for the motivating data sets in this paper).

The fundamental intuition is that if the model truly provides good fit, then the ECDF from the simulated values should closely resemble the ECDF derived from the original data. Consequently the ECDFs of the densities of the original and simulated data should be similar to one another. This provides a useful visual diagnostic of model fit.

5. Results

Results are presented for the motivating data sets. In all cases a mixture of MNIG distributions with parameter constraints produces the optimum value of BIC among all competing models.

5.1. Old Faithful data

In the mixture of Gaussians setting, the optimal model has $G = 3$ components and a common covariance structure across groups, $\boldsymbol{\Sigma}_g = \lambda \mathbf{DAD}'$, with BIC $= -2314.4$. The **mixsmsn** package in **R** (Cabral et al., 2014; Prates et al., 2013b) considers mixtures of Gaussian, t, skew-t, skew normal, skew contaminated normal and skew slash contaminated normal models. The optimal candidate among the **mixsmsn** models, as determined by BIC, is a $G = 2$ mixture of skew-t distributions with unequal covariance structure across groups and BIC $= -2313.0$.

Using a mixture of MNIG distributions yields an optimum BIC of -2302.5. The results are attained for a $G = 2$ model with scale matrix $\tilde{\boldsymbol{\Sigma}}_g = \lambda \mathbf{A}_g$ under the constraints that $\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}_g$ (variable levels of skew across groups) and $\tilde{\gamma}_g = \tilde{\gamma}$. The direct impact of the more flexible component shapes in the mixture of MNIGs and mixture of skew-t distributions models is that only two components are required to cluster the data. The lower component depicted in each of the plots in Figure 4 exhibits strong positive skew, permitting a close fit to the tight concentration of points with the smallest values of eruption duration and waiting duration. The upper component exhibits strong negative skew, accounting for cluster membership of observations originally partitioned into two separate (symmetric) components in the mixture of Gaussians model. The level of skew present in the data is marked and varies across groups, so the potential constraints that $\tilde{\boldsymbol{\beta}}_g = 0$ or $\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}$ are not appropriate for the mixture of MNIGs model.

Table 1 compares the classifications resulting from the optimal **mclust** and mixture of MNIGs models. There is complete classification agreement between the optimal mixture of MNIGs and **mixsmsn** models. Table 2 contains the values of the information metrics that can be used to compare the two models. Since no "true" labels are available for the data, the metrics simply reflect the level of harmony between the competing clustering solutions, which is moderate. However the Van Dongen criterion value shows that imposing an optimal matching where the solutions are constrained to have the same number of groups largely eliminates disagreement. The high value of VI_{soft} between the MNIGs and **mixsmsn** solutions corresponds to the exact match in their hard classifications of observations.

Figures 4(c) and 4(d) highlight the nature of the superior BIC attained by the mixture of MNIGs approach. The additional modeling flexibility afforded by the

MNIG distribution permits component densities to have steep contours for skewed, asymmetric groups. This phenomenon is particularly noticeable in the lower left portion of Figure 4(d). In contrast, **mclust** fits an additional cluster to cope with the asymmetry of the data (Figure 4(a)). The overall mixture densities produced by the competing clustering methodologies are depicted in Figures 4(e) and 4(f). The corresponding plots for the mixture of skew-t distributions fitted by the **mixsmsn** package closely resemble those for the mixture of MNIGs.

Figure 5(a) illustrates that the optimal mixture of MNIGs provides a fairly good fit to the data, with a slight tendency to place too much probability mass in the regions of the data with medium log density. Figure 5(b) presents the equivalent plot for the optimal **mclust** solution, which provides marginally better fit than the optimal mixture of MNIGs model, but only at the expense of incorporating an additional cluster.

An alternative to fitting skewed and/or heavy tailed distributions for the purpose of model-based clustering is to consider approaches based on merging mixture components. The function **clustCombi** in **mclust** combines mixture components hierarchically, based on an entropy criterion (Baudry et al., 2010). The starting number of clusters is first established using BIC. Two mixture components, among all possible pairs of components, are then sequentially merged at each step, based on minimizing the entropy of the resulting solution. Judgment as to the optimal overall number of clusters is made by applying an elbow rule to the graph plotting entropy variation versus number of clusters.

On this basis it suggests that merging the upper two components of the *Old Faithful* data is appropriate, producing the same clustering solution as provided by the mixture of MNIGs and **mixsmsn** approaches. However it is worth noting that, although the clustering solutions match, the single Gaussian component fitted to the lower left portion of Figure 4 remains unaffected, and is not able to capture the skew and asymmetry present in that subset of the data.

Table 1: Comparison of classifications of the *Old Faithful* data using optimal **mclust** and mixture of MNIGs models.

	MNIG group 1	*MNIG group* 2
mclust *group* 1	130	0
mclust *group* 2	0	97
mclust *group* 3	45	0

Table 2: Information metric results for the *Old Faithful* data.

Metric	*Optimum*	**mclust** *versus MNIG*	*MNIG versus* **mixsmsn**
Mirkin	0	0.1581	0
Van Dongen	0	0.0827	0
VI_{soft}	0	0.4443	0.926
Adjusted Rand	1	0.6884	1

5.2. Flow cytometry FLAME data

The optimal mixture of Gaussians comprises $G = 7$ components and covariance structure $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A} \mathbf{D}_g$ with BIC $= -38146.7$. The optimal model fitted by the FLAME software (Pyne et al., 2009) comprises $G = 5$ components. The model is a mixture of multivariate skew distributions with heavy tails to handle outliers. The BIC is not reported in the online version of the software. The optimal candidate among the **mixsmsn** models, as determined by BIC, is a mixture of $G = 6$ skew normal distributions with unequal covariance structure across groups and BIC $= -37783.1$. Employing the **clustCombi** procedure in **mclust** suggests that a model with $G = 5$ components is appropriate. This produces a clustering solution broadly similar to that of the FLAME software.

Using a mixture of MNIG distributions an optimum BIC of -34531.0 is obtained for a $G = 6$ model with scale matrix $\tilde{\boldsymbol{\Sigma}}_g = \lambda \mathbf{A}_g$ and constraints $\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}$ (equal level of skew across groups) and $\tilde{\gamma}_g = \tilde{\gamma}$. The greater flexibility in component shape afforded by the mixture of MNIG distributions means the model separates two clusters that, visually, are clearly distinct from one another in the pairs plots for the data. These observations are not separated by the FLAME software. The fact that the groups are not symmetric results in additional components being fitted by the mixture of Gaussians model. In the case of the *FLAME* data there is similar skew across all components. Hence the result that the constraint that $\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}$ is chosen by the mixture of MNIGs model is to be expected.

As shown in Table 3 there is a moderate level of disparity among the competing methodologies. The Van Dongen criterion again indicates that this is largely attributable to the differing number of groups between solutions.

Figure 6 presents the pairs plots for the *FLAME* data. Observations are distinguished by classification under the optimal (a) **mclust**, (b) FLAME and (c) mixture of MNIGs models. Separation of components is masked for several of the pairs plots, perhaps explaining why the FLAME software identifies only 5 groups. However Figure 6(c) highlights a clear separation between two of the clusters (magenta and gray)

that are grouped together by the FLAME software (red).

Figure 7(a) shows that the optimal mixture of MNIGs fits the data well, but has a slight tendency to overestimate the density in the middle of the distribution. The somewhat unusual shape exhibited at the top of this plot is attributable to the fact that there is a small set of observations with a near-linear trend at the extremes of the marginal distributions. Hence they have high values of log density and the ECDF approaches 1 slowly as it accounts for their appearance in the data. Figure 7(b) presents the equivalent plot for the optimal **mixsmsn** solution, which is clearly a poorer fit than the optimal mixture of MNIGs model.

Table 3: Information metric results for the *FLAME* data.

Metric	*Optimum*	**mclust** *versus MNIG*	**FLAME** *versus MNIG*
Mirkin	0	0.1010	0.1634
Van Dongen	0	0.0763	0.0373
VI_{soft}	0	1.1185	NA
Adjusted Rand	1	0.6798	0.6148

5.3. Iris data

The optimal mixture of Gaussians has $G = 2$ and $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A} \mathbf{D}_g$, with BIC $= -561.7$. The optimal candidate under **mixsmsn**, as determined by BIC, is a mixture of $G = 2$ t distributions with unequal covariance structure across groups and BIC $= -573.8$. Using a mixture of $G = 2$ MNIG distributions, the optimum BIC of -515.3 is greater than its **mclust** and **mixsmsn** counterparts. The results are attained for a model with $\tilde{\boldsymbol{\Sigma}}_g = \lambda \mathbf{A}_g$ under the constraints that $\tilde{\boldsymbol{\beta}}_g = 0$ and $\tilde{\gamma}_g = \tilde{\gamma}$. All three models produce identical clusterings of the data.

In the case of the *Iris* data clustered using a mixture of two Gaussian or two t distributions, neither of the components exhibit skew. Hence the result that $\tilde{\boldsymbol{\beta}}_g = 0$ is chosen verifies that in the case where a simpler model suffices, the mixture of MNIG distributions is sufficiently flexible and responsive. If a mixture of $G = 3$ components is fitted, in keeping with the number of distinct flower types, all three clustering methodologies produce identical results.

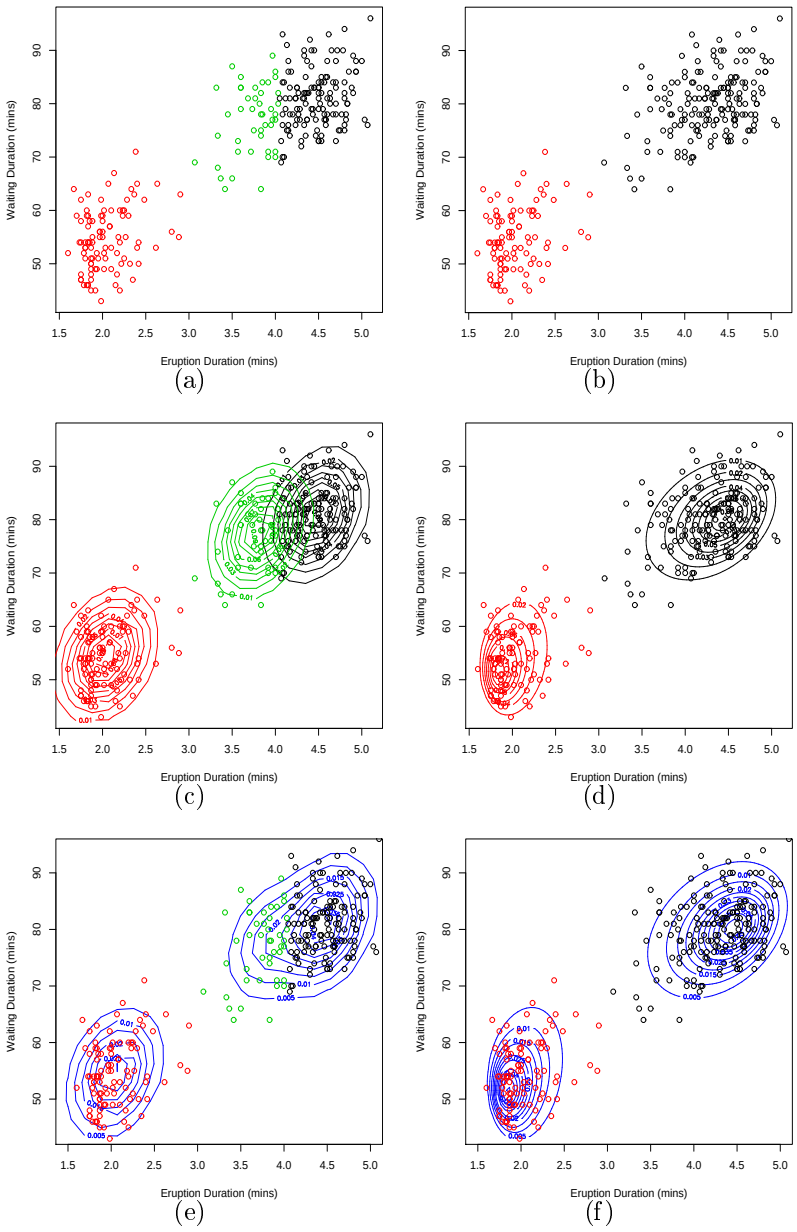
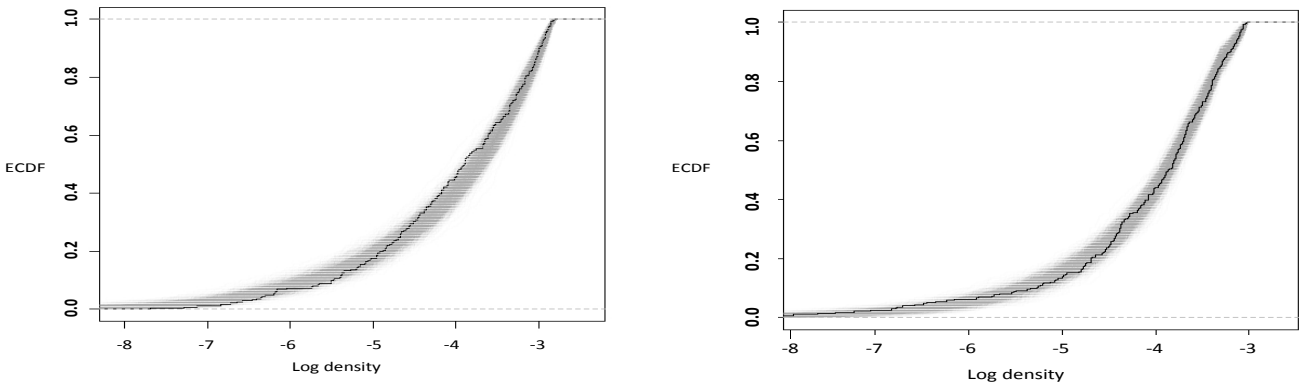
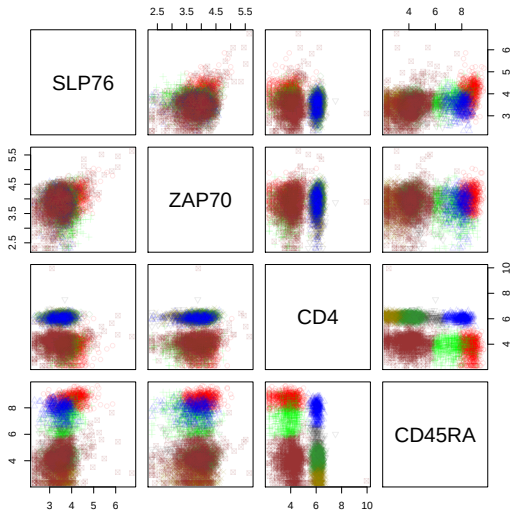


Figure 4: Clustering solutions for the *Old Faithful* data under (a) optimal **mclust** model: $G = 3$ and $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma} = \lambda \mathbf{DAD}'$ (b) optimal mixture of MNIGs model: $G = 2$ and $\tilde{\boldsymbol{\Sigma}}_g = \lambda \mathbf{A}_g$, $\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}_g$ and $\tilde{\gamma}_g = \tilde{\gamma}$ (c) optimal **mclust** model with component densities superimposed (d) optimal mixture of MNIGs model with component densities superimposed (e) optimal **mclust** model with mixture density superimposed (f) optimal mixture of MNIGs model with mixture density superimposed.

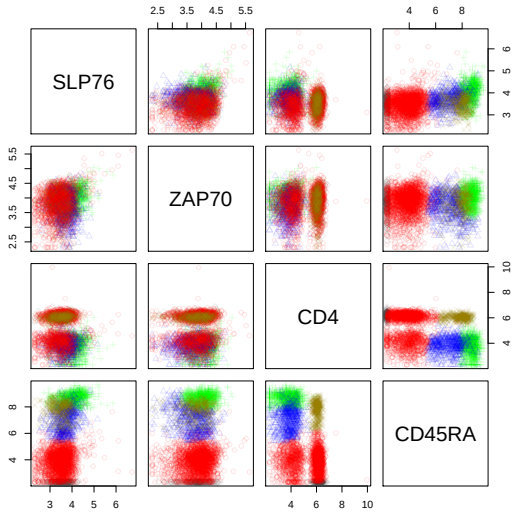


(a) (b)

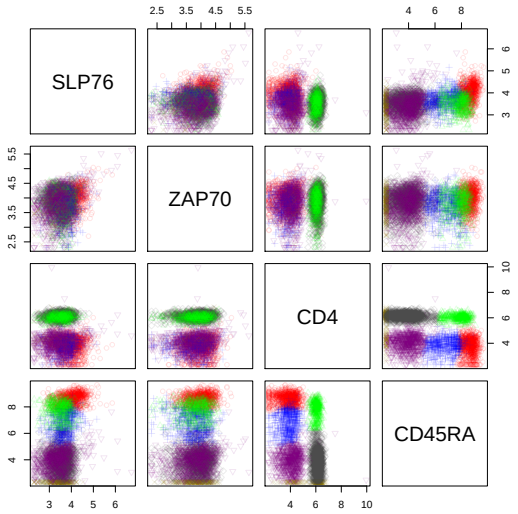
Figure 5: Adapted Kolmogorov-Smirnov goodness of fit plot for the *Old Faithful* data under (a) the optimal mixture of MNIGs model and (b) the optimal **mclust** model. The black line represents the ECDF of the original data, the gray lines represent the ECDFs of each of the 1000 data sets simulated from the optimal model.



(a)

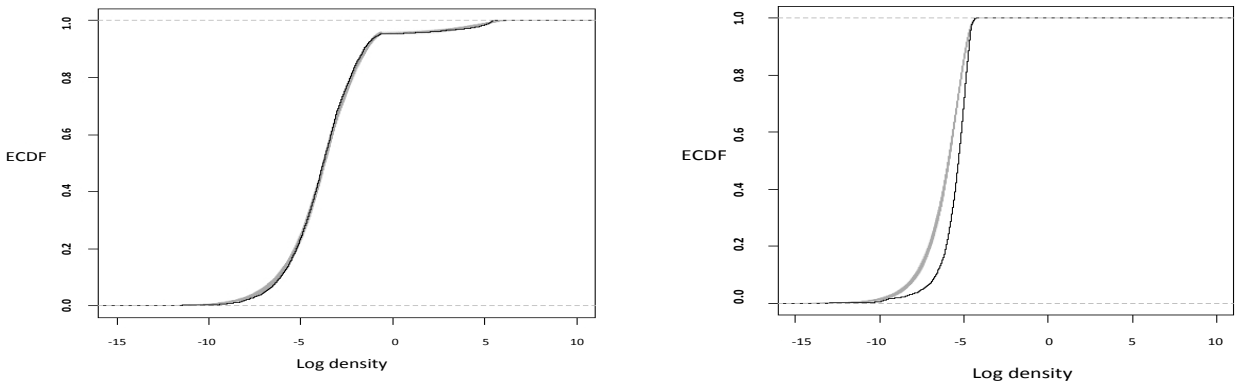


(b)



(c)

Figure 6: Pairs plot of the *FLAME* flow cytometry data under (a) optimal **mclust** model: $G = 7$, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma} = \lambda \mathbf{DAD}'$ (b) optimal FLAME model: $G = 5$ (c) optimal mixture of MNIGs model: $G = 6$, $\tilde{\boldsymbol{\Sigma}}_g = \lambda \mathbf{A}_g$, $\tilde{\boldsymbol{\beta}}_g = \tilde{\boldsymbol{\beta}}$ and $\tilde{\gamma}_g = \tilde{\gamma}$. The distinct group colorings present in the plot are: red, green, blue, magenta, gray, mustard, maroon.



(a) (b)

Figure 7: Adapted Kolmogorov-Smirnov goodness of fit plot for the *FLAME* data under (a) the optimal mixture of MNIGs model and (b) the optimal **mixsmsn** mixture of skew normals model. The black line represents the ECDF of the original data, the gray lines represent the ECDFs of each of the 1000 data sets simulated from the optimal model.

6. Conclusions and further work

A mixture of MNIG distributions provides a valuable alternative to a mixture of Gaussian distributions as a means of performing model-based clustering. Through a combination of improved log-likelihood at convergence and simplified scale matrix structure, it can achieve a superior BIC value to **mclust**, despite the additional number of parameters. The implementation of the subset of constrained parameter models helps significantly in this regard.

There appears to be a notable synergy between the $\lambda \mathbf{A}_g$ scale matrix structure and the use of the multivariate normal inverse Gaussian distribution. The $\tilde{\boldsymbol{\beta}}$ parameter that governs skew in the MNIG case provides the off-diagonal elements in the covariance matrix that would otherwise be missing in the multivariate Gaussian case. Simultaneously it captures the skew and heavy tails in the original data. This "two for the price of one" effect is a powerful one in the context of model parsimony.

The set of information metrics presented have widespread applicability in model-fitting beyond their use in this paper. While goodness of fit in a regression setting is commonly monitored using statistics such as the coefficient of determination, many authors have been critical of their use (King, 1986). The goodness of fit test developed, based on the Kolmogorov-Smirnov approach, provides a useful diagnostic tool for investigating whether a model focuses density on regions in a manner consistent with the underlying data.

In terms of further work, it would be extremely useful to automate the fitting of mixtures of MNIG distributions via a package similar to **mclust**. It is envisaged that such a facility (tentatively titled **MNIGclust**) would allow the number of components, covariance structure and any desired parameter constraints to be specified by the user or, alternatively, in the absence of user-input, to automatically assess all candidate models using BIC in a manner similar to **mclust**. However this presents a significant increase in scale.By default **mclust** assesses 83 models (3 covariance structures when $G = 1$ and 10 covariance structures when $G =$ 2-9). However, **MNIGclust** would have to handle the extra constraints on $\tilde{\boldsymbol{\beta}}_g$ and $\tilde{\gamma}_g$, giving rise to a far greater number of models. While any individual combination can be run in **R** in reasonable time for a data set of moderate size and dimensionality, the computational burden from evaluating all models necessitates that the package be coded in a compiled language. This is certainly feasible, given that all E-step and M-step updates are available in closed form.

Clearly there are additional applications of the MNIG clustering methodology. For example, financial returns data are known to generally exhibit both skew and heavy tails and failure to adequately model these features has precipitated several financial crises (Chen et al., 2001). Longitudinal data is another application where

model-based clustering could be extended beyond the standard mixture of Gaussians approach (McNicholas and Murphy, 2010). Both examples present compelling areas for further exploration within the MNIG clustering framework. The approach could further be extended to accommodate "asymmetric" missing values, as has already been carried out for mixtures of other skew distributions, such as mixtures of skew-normals, mixtures of skew-t; and for mixtures of common factor analyzers; with incomplete and missing data respectively (Lin et al., 2009; Lin and Lin, 2011; Wang, 2013).

Acknowledgements

The authors wish to thank the associate editor and reviewers for their comments, which greatly improved this work. This work is supported by Science Foundation Ireland under the Research Frontiers Programme (2007/RFP/MATH281) and the Insight Research Centre (SFI/12/RC/2289).

References

Andrews, J.L., McNicholas, P.D., 2011. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. the teigen family. Statistics and Computing 22, 1021–1029.

Azzalini, A., Bowman, A.W., 1990. A look at some data on the Old Faithful geyser. Applied Statistics 39, 357–365.

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

Baudry, J.P., Raftery, A., Celeux, G., Lo, K., Gottardo, R., 2010. Combining mixture components for clustering. Journal of Computational and Graphical Statistics 9, 332–353.

Bensmail, H., Celeux, G., 1996. Regularized Gaussian discriminant analysis through eigenvalue decomposition. Journal of the American Statistical Association 91, 1743–1748.

Cabral, C.R., Lachos, V.H., Prates, M., 2012. Robust multivariate mixture modelling using scale mixtures of skew-normal distributions. Computational Statistics and Data Analysis 56, 226–246.

Cabral, C.R., Lachos, V.H., Zeller, C.B., 2014. Multivariate measurement error models using finite mixtures of skew-student t distributions. Journal of Multivariate Analysis 124, 179–198.

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

Chen, J., Hong, H., Stein, J., 2001. Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices. Journal of Financial Economics 61, 345–381.

Dasgupta, A., Raftery, A.E., 1998. Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American Statistical Association 93, 294–302.

Fisher, R., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.

Fraley, C., Raftery, A.E., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. Computer Journal 41, 578–588.

Fraley, C., Raftery, A.E., 1999. Mclust: Software for model-based clustering. Journal of Classification 16, 297–306.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97, 611–612.

Fruhwirth-Schnatter, S., Pyne, S., 2009. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. Biostatistics 11, 317–336.

Gottardo, R., Lo, K., 2011. Robust Model-based Clustering of Flow Cytometry Data: The flowClust Package. Technical Report. UBC. Vancouver, Canada.

Hennig, C., 2010. Methods for merging gaussian mixture components. Advances in Data Analysis and Classification 4, 3–34.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of Classification 2, 193–218.

Hunter, D., Wang, S., Hettmansperger, T., 2007. Inference for mixtures of symmetric distributions. The Annals of Statistics 35, 224–251.

Karlis, D., Santourian, A., 2008. Model-based clustering with non-elliptically contoured distributions. Statistics and Computing 19, 73–83.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90, 773–795.

King, G., 1986. How not to lie with statistics: avoiding common mistakes in quantitative political science. American Journal of Political Science 30, 666–687.

Lin, T.I., Ho, H.J., Chen, C.L., 2009. Analysis of multivariate skew normal models with incomplete data. Journal of Multivariate Analtsis 100, 2337–2351.

Lin, T.I., Lee, J.C., Hsieh, W.J., 2007a. Robust mixture modeling using the skew t distribution. Statistics and Computing 17, 81–92.

Lin, T.I., Lee, J.C., Yen, S.Y., 2007b. Finite mixture modelling using the skew normal distribution. Statistica Sinica 17, 909–927.

Lin, T.I., Lin, T., 2011. Robust statistical modelling using the multivariate skew t distribution with complete and incomplete data. Statistics and Computing 11, 253–277.

MacLean, C., Morton, N., Elston, R., Yee, S., 1976. Skewness in commingling distributions. Biometrics 32, 695–699.

Massey, F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association 46, 68–78.

McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley Interscience, New York.

McNicholas, P.D., Murphy, T.B., 2010. Model-based clustering of longitudinal data. The Canadian Journal of Statistics 38, 153–168.

McNicholas, S.M., McNicholas, P.D., Browne, R.P., 2013. Mixtures of Variance-Gamma Distributions. ArXiv e-prints `arXiv:1309.2695`.

Mechel, F., 1966. Calculation of the modified bessel functions of the second kind with complex argument. Mathematics of Computation 20, 407–412.

Meila, M., 2007. Comparing clusterings - an information based distance. Journal of Multivariate Analysis 98, 873–895.

Mirkin, B.G., Cherny, L.B., 1970. Measurement of the partition between distinct partitions of a finite set of objects. Automation and Remote Control 31, 786–792.

Prates, M.O., Cabral, C.R., Lachos, V.H., 2013a. Fitting finite mixture of scale mixture of skew-normal distributions. Journal of Statistical Software 54.

Prates, M.O., Lachos, V.H., Cabral, C.R., 2013b. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. R package version 0.2-9.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T.I., Maier, L.M., Allan, C.B., McLachlan, G., Tamayo, P. Hafler, D., De Jager, P.L., Mesirov, J.P., 2009. Automated high-dimensional flow cytometric data analysis. PNAS 106, 8519–8524.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: `http://www.R-project.org`. ISBN 3-900051-07-0.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Van Dongen, S., 2000. Performance criteria for graph clustering and Markov cluster experiments. Technical Report. National Research Institute for Mathematics and Computer Science. Amsterdam, Holland.

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

Wang, W.L., 2013. Mixtures of common factor analyzers for high-dimensional data with missing information. Journal of Multivariate Analysis 117, 120–133.

Wasserman, R.E.K.L., 1995. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. Journal of the American Statistical Association 90, 928–934.

Yu, Y., 2011. On Normal Variance-Mean Mixtures. Technical Report. University of California. Irvine, California, USA.