Gaussian Parsimonious Clustering Models with Covariates and a Noise Component

Keefe Murphy[1,2] Thomas Brendan Murphy[1,2]

keefe.murphy@ucd.ie brendan.murphy@ucd.ie

[1] School of Mathematics and Statistics, University College Dublin
[2] Insight Centre for Data Analytics, University College Dublin

Abstract

We consider model-based clustering methods for continuous, correlated data that account for external information available in the presence of mixed-type fixed covariates by proposing the MoEClust suite of models. These models allow different subsets of covariates to influence the component weights and/or component densities by modelling the parameters of the mixture as functions of the covariates. A familiar range of constrained eigen-decomposition parameterisations of the component covariance matrices are also accommodated. This paper thus addresses the equivalent aims of including covariates in Gaussian parsimonious clustering models and incorporating parsimonious covariance structures into all special cases of the Gaussian mixture of experts framework. The MoEClust models demonstrate significant improvement from both perspectives in applications to both univariate and multivariate data sets. Novel extensions to include a uniform noise component for capturing outliers and to address initialisation of the EM algorithm, model selection, and the visualisation of results are also proposed.

Keywords: Model-based clustering, mixtures of experts, EM algorithm, parsimony, multivariate response, covariates, noise component.

1 Introduction

In many analyses using the standard mixture model framework, a clustering method is typically implemented on the outcome variables only. Reference is not made to the associated covariates until the structure of the produced clustering is investigated in light of the information present in the covariates. Therefore, interpretations of the values of the model parameters within each component are guided by covariates that are not actually used in the construction of the clusters. It is desirable to have covariates incorporated into the clustering process and not only into the interpretation of the clustering structure and model parameters, thereby making them endogenous rather than exogenous to the clustering model. This both informs the construction of the clusters and provides richer insight into the type of observation which characterises each cluster.

When each observation consists of a response variable $\mathbf{y}_i$ on which the clustering is based and covariates $\mathbf{x}_i$ there are, broadly speaking, two main approaches in the literature to having covariates guide construction of the clusters, neatly summarised by Lamont et al. (2016) and compared in Ingrassia et al. (2012). Letting $\mathbf{z}_i$ denote the latent cluster membership indicator vector, where $z_{ig} = 1$ if observation i belongs to cluster g

and $z_{ig} = 0$ otherwise, the first approach assumes that $\mathbf{z}_i$ affects the distribution of $\mathbf{x}_i$. In probabilistic terms, this means to replace the actual group-specific conditional distribution $f\big(\mathbf{y}_i \,|\, \mathbf{x}_i, z_{ig} = 1\big)\,\mathrm{Pr}\big(z_{ig} = 1\big)$ with $f\big(\mathbf{y}_i|\mathbf{x}_i, z_{ig} = 1\big)f\big(\mathbf{x}_i \,|\, z_{ig} = 1\big)\,\mathrm{Pr}\big(z_{ig} = 1\big)$. The name 'cluster-weighted model' (CWM) is frequently given to this approach, e.g. Dang et al. (2017) and Ingrassia et al. (2015); the latter provides a recent extension allowing for mixed-type covariates, with a further generalisation presented in Punzo & Ingrassia (2016). Noting the use of the alternative term 'mixtures of regressions with *random* covariates' to describe CWMs (e.g. Hennig 2000) provides opportunity to clarify that the remainder of this paper focuses on the second approach, with *fixed* potentially mixed-type covariates affecting cluster membership via $f\big(\mathbf{y}_i \,|\, \mathbf{x}_i, z_{ig} = 1\big)\,\mathrm{Pr}\big(z_{ig} = 1 \,|\, \mathbf{x}_i\big)$.

This is achieved using the mixture of experts (MoE) paradigm (Dayton & Macready, 1988; Jacobs et al., 1991) in which the parameters of the mixture are modelled as functions of fixed, potentially mixed-type covariates. We present, for finite mixtures of multivariate, continuous, correlated responses, a unifying framework combining all of the special cases of the Gaussian MoE model with the flexibility afforded by the covariance constraints in the Gaussian parsimonious clustering model (GPCM) family (Banfield & Raftery, 1993; Celeux & Govaert, 1995). This has, to date, been lacking for all but the mixture of regressions and the mixture of regressions with concomitant variables where the same covariates enter both parts of the model (Dang & McNicholas, 2015).

Parsimony is obtained in GPCMs by imposing constraints on the elements of an eigen-decomposition of the component covariance matrices. For MoE models, reducing the number of covariance parameters in this manner can help offset the number of regression parameters introduced by covariates, which is particularly advantageous when model selection is conducted using information criteria with penalty terms involving parameter counts. The main contribution of this paper is the development of a framework combining GPCM constraints with all of the special cases of the Gaussian MoE framework whereby different subsets of covariates can enter either, neither, or both the component densities and component weights. We also consider the special cases of the MoE framework for univariate response data with equal and unequal variance across components. Thus, this paper addresses the aim of incorporating potentially mixed-type covariates into the GPCM family and the equivalent aim of bringing GPCM covariance constraints into the Gaussian MoE framework, by proposing the MoEClust model family. The name MoEClust comes from the interest in employing MoE models chiefly for clustering purposes. From both perspectives, MoEClust models show significant improvement in applications to both univariate and multivariate response data.

Other novel contributions include the addition of a noise component for capturing outlying observations, and proposed solutions to initialising the EM algorithm sensibly, addressing the issue of model selection, and a means for visualising the results of MoEClust models. We also expand the number of special cases in the MoE framework from four to six, by considering more parsimonious counterparts to the standard mixture model and the mixture of regressions by constraining the mixing proportions. In addition, a software implementation of the full suite of MoEClust models is provided by the associated R package `MoEClust` (Murphy & Murphy, 2020), which is available from `www.r-project.org` (R Core Team, 2020), with which all results were obtained. The syntax of the popular `mclust` package (Scrucca et al., 2016) is closely mimicked, with formula interfaces for specifying covariates in the gating and/or expert networks.

The structure of the paper is as follows. For both Gaussian mixtures of experts and MoEClust models, the modelling frameworks and inferential procedures are described, respectively, in Section 2 and Section 3. Section 3.3 describes the addition of a noise component for capturing outliers. Section 4 discusses proposals for addressing some practical issues affecting performance, namely the initialisation of the EM algorithm used to fit the models (Section 4.1), and issues around model selection (Section 4.2). The performance of the proposed models is illustrated in Section 5 with applications to univariate response CO_2 emissions data (Section 5.1) and multivariate response data from the Australian Institute of Sports (Section 5.2). Finally, the paper concludes with a brief discussion in Section 6, with some additional results deferred to the Appendices.

2 Modelling

This section builds up the MoEClust models by first describing the mixture of experts (MoE) modelling framework in Section 2.1 — elaborating on the special cases of the MoE model in Section 2.1.1 — and then extending to the family of MoEClust models comprising Gaussian mixture of experts models with parsimonious covariance structures from the GPCM family in Sections 2.2 and 2.3. Finally, a brief review of existing models and software is given in Section 2.4.

2.1 Mixtures of Experts

The mixture of experts model (Dayton & Macready, 1988; Jacobs et al., 1991) extends the mixture model used to cluster response data $\mathbf{y}_i$ by allowing the parameters of the model for observation i to depend on covariates $\mathbf{x}_i$. An independent sample of response/outcome variables of dimension p, denoted by $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, is modelled by a G-component finite mixture model where the model parameters depend on the associated covariate inputs $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of dimension d. The MoE model is often referred to as a conditional mixture model (Bishop, 2006) because, given the set of covariates $\mathbf{x}_i$, the distribution of the response variable $\mathbf{y}_i$ is a finite mixture model:

$$f\big(\mathbf{y}_i \,|\, \mathbf{x}_i\big) = \sum_{g=1}^{G} \tau_g(\mathbf{x}_i)\, f\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\left(\mathbf{x}_i\right)\big).$$

Each component is modelled by a probability density function $f\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\left(\mathbf{x}_i\right)\big)$ with component-specific parameters $\boldsymbol{\theta}_g\left(\mathbf{x}_i\right)$ and mixing proportions $\tau_g(\mathbf{x}_i)$ which are only allowed to depend on covariates when $G \geq 2$. As usual, $\tau_g(\mathbf{x}_i) > 0$ and $\sum_{g=1}^{G} \tau_g(\mathbf{x}_i) = 1$.

The MoE framework facilitates flexible modelling. While the response variable $\mathbf{y}_i$ is modelled via a finite mixture, model parameters are modelled as functions of related covariates $\mathbf{x}_i$ from the context under study. Both the mixing proportions and the parameters of component densities can depend on $\mathbf{x}_i$. The terminology used to describe MoE models in the machine learning literature often refers to the component densities $f\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\left(\mathbf{x}_i\right)\big)$ as 'experts' or the 'expert network', and to the mixing proportions $\tau_g(\mathbf{x}_i)$ as 'gates' or the 'gating network', hence the nomenclature *mixture of experts*. Given that covariates can be continuous and/or categorical with multiple levels, we let $d + 1$ denote the number of columns in the corresponding design matrices, accounting also for the intercept term, in contrast to the number of covariates r, with $d \geq r$.

In the original formulation of the MoE model for continuous data (Jacobs et al., 1991), the mixing proportions (gating network) are modelled using multinomial logistic regression (MLR), though this need not strictly be the case; Geweke & Keane (2007) impose a multinomial probit structure here instead. The mixture components (expert networks) are generalised linear models (GLM; McCullagh & Nelder, 1983). Thus,

$$\widehat{\tau}_g(\mathbf{x}_i) = \frac{\exp\big(\widetilde{\mathbf{x}}_i\widehat{\boldsymbol{\beta}}_g\big)}{\sum_{h=1}^{G}\exp\big(\widetilde{\mathbf{x}}_i\widehat{\boldsymbol{\beta}}_h\big)}, \tag{1}$$

and

$$\widehat{\boldsymbol{\theta}}_g(\mathbf{x}_i) = \Big\{\psi\big(\widetilde{\mathbf{x}}_i\widehat{\boldsymbol{\gamma}}_g\big), \widehat{\boldsymbol{\Sigma}}_g\Big\}, \tag{2}$$

for some link function $\psi\left(\cdot\right)$, with a collection of parameters in the component densities (comprising a $(d+1)\times p$ matrix of expert network regression parameters $\widehat{\boldsymbol{\gamma}}_g$ and the $p \times p$ component covariance matrix $\widehat{\boldsymbol{\Sigma}}_g$), a $(d+1)$-dimensional vector of regression parameters $\widehat{\boldsymbol{\beta}}_g$ in the gates in (1), and $\widetilde{\mathbf{x}}_i = (1, \mathbf{x}_i)$. Note that expert network covariates influence only the component means, and not the component covariance matrices. Henceforth, we restrict our attention to continuous outcome variables as per the GPCM family. Therefore, component densities are assumed to be the p-variate Gaussian $\phi\left(\mathbf{y}_i\,|\,\cdot\right)$, and the link function $\psi\left(\cdot\right)$ in (2) is simply the identity, such that covariates are linearly related to the response variables, i.e.

$$f\big(\mathbf{y}_i\,|\,\mathbf{x}_i\big) = \sum_{g=1}^{G}\tau_g(\mathbf{x}_i)\,\phi\big(\mathbf{y}_i\,|\,\boldsymbol{\theta}_g\left(\mathbf{x}_i\right) = \big\{\widetilde{\mathbf{x}}_i\boldsymbol{\gamma}_g, \boldsymbol{\Sigma}_g\big\}\big). \tag{3}$$

2.1.1 The MoE Family of Models

It is possible that some, none, or all model parameters depend on the covariates. This leads to the four special cases of the Gaussian MoE framework shown in Figure 1, with the following interpretations, due to Gormley & Murphy (2011):

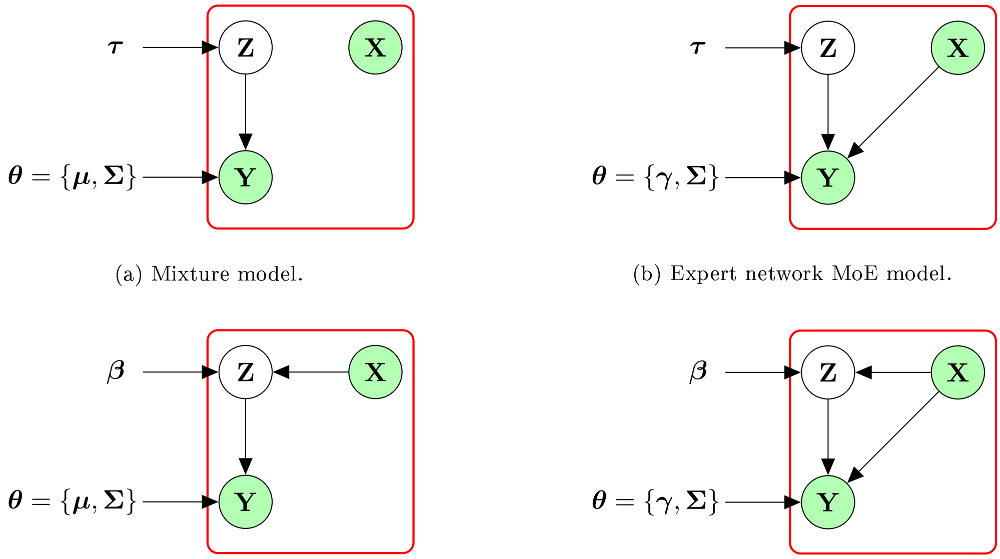
(a) in the *mixture model* the distribution of $\mathbf{y}_i$ depends on the latent cluster membership variable $\mathbf{z}_i$, the distribution of $\mathbf{z}_i$ is independent of the covariates $\mathbf{x}_i$, and $\mathbf{y}_i$ is independent of $\mathbf{x}_i$ conditional on $\mathbf{z}_i$: $f\big(\mathbf{y}_i\big) = \sum_{g=1}^{G}\tau_g\phi\big(\mathbf{y}_i\,|\,\boldsymbol{\theta}_g = \big\{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\big\}\big)$.

(b) in the *expert network MoE model* the distribution of $\mathbf{y}_i$ depends on the covariates $\mathbf{x}_i$ and the latent cluster membership variable $\mathbf{z}_i$, and the distribution of $\mathbf{z}_i$ is independent of $\mathbf{x}_i$: $f\big(\mathbf{y}_i\,|\,\mathbf{x}_i\big) = \sum_{g=1}^{G}\tau_g\phi\big(\mathbf{y}_i\,|\,\boldsymbol{\theta}_g\big(\mathbf{x}_i\big) = \big\{\widetilde{\mathbf{x}}_i\boldsymbol{\gamma}_g, \boldsymbol{\Sigma}_g\big\}\big)$.

(c) in the *gating network MoE model* the distribution of $\mathbf{y}_i$ depends on the latent cluster membership variable $\mathbf{z}_i$, $\mathbf{z}_i$ depends on the covariates $\mathbf{x}_i$, and $\mathbf{y}_i$ is independent of $\mathbf{x}_i$ conditional on $\mathbf{z}_i$: $f\big(\mathbf{y}_i\,|\,\mathbf{x}_i\big) = \sum_{g=1}^{G}\tau_g\big(\mathbf{x}_i\big)\phi\big(\mathbf{y}_i\,|\,\boldsymbol{\theta}_g = \big\{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\big\}\big)$.

(d) in the *full MoE model*, given by (3), the distribution of $\mathbf{y}_i$ depends on both the covariates $\mathbf{x}_i$ and on the latent cluster membership variable $\mathbf{z}_i$, and the distribution of the latent variable $\mathbf{z}_i$ depends in turn on the covariates $\mathbf{x}_i$.

For models (c) and (d), $\mathbf{z}_i$ has a multinomial distribution with a single trial and probabilities equal to $\tau_g(\mathbf{x}_i)$. The full MoE model thus has the following latent variable representation: $\big(\mathbf{y}_i\,|\,\mathbf{x}_i, z_{ig} = 1\big) \sim \phi\big(\mathbf{y}_i\,|\,\boldsymbol{\theta}_g\left(\mathbf{x}_i\right) = \big\{\widetilde{\mathbf{x}}_i\boldsymbol{\gamma}_g, \boldsymbol{\Sigma}_g\big\}\big), \mathrm{Pr}(z_{ig} = 1\,|\,\mathbf{x}_i) = \tau_g(\mathbf{x}_i)$.


$\boldsymbol{\tau}$
Z
X
$\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$
Y
$\boldsymbol{\tau}$
Z
X
$\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$
Y
$\boldsymbol{\beta}$
Z
X
$\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$
Y
$\boldsymbol{\beta}$
Z
X
$\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$
Y

(a) Mixture model.

(b) Expert network MoE model.

(c) Gating network MoE model.

(d) Full MoE model.

Figure 1: The graphical model representation of the mixture of experts models. The differences between the special cases are due to the presence or absence of edges between the covariates $\mathbf{X}$ and the latent variables $\mathbf{Z}$ and/or response variables $\mathbf{Y}$. Note that different subsets of the covariates in $\mathbf{X}$ can enter these two different parts of the full MoE model in (d).

The MoE family can be expanded further, from four to six special cases, by considering the models in (a) and (b), under which covariates do not enter the gating network, by constraining the mixing proportions to be equal across components, i.e. $\tau_g = {}^1\!/_G \,\forall\, g$. This leads, respectively, to the *equal mixing proportion mixture model* and *equal mixing proportion expert network MoE model*. Such models are more parsimonious than their counterparts with unconstrained $\boldsymbol{\tau}$, as they require estimation of $G-1$ fewer parameters. Note that the size of a cluster is proportional to τ_g, which is distinct from its volume (Celeux & Govaert, 1995). Thus, situations where $\tau_{ig} = \tau_g(\mathbf{x}_i)$, $\tau_{ig} = \tau_g$, or $\tau_{ig} = {}^1\!/_G$ can all be accommodated. The six special cases of this MoE framework can be applied to both univariate and multivariate response data.

It is worth noting that CWMs most fundamentally differ from MoE models in their handling of the mixing proportions τ_g and in how the joint density $f\left(\mathbf{x}_i, z_{ig} = 1\right)$ is treated, either as $\Pr(z_{ig} = 1 \,|\, \mathbf{x}_i) = \tau_g(\mathbf{x}_i)$ (MoE) or $f\left(\mathbf{x}_i \,|\, z_{ig} = 1\right) \Pr(z_{ig} = 1)$ (CWM). In other words, the direction of the edge between $\mathbf{X}$ and $\mathbf{Z}$ in the full MoE model in Figure 1d is reversed under CWMs (Ingrassia et al., 2012). By virtue of modelling the distribution of the covariates, CWMs are also inherently less parsimonious. The same covariate(s) can enter both parts of full MoE models, in principle. Such models can provide a useful estimation of the conditional density of the outcome given the covariates, but the interpretation of the clustering model and the effect of the covariates becomes more difficult in this case. Conversely, allowing different covariates enter different parts of the model further differentiates MoE models from CWMs. It is common to distinguish among the overall set of covariates between *concomitant* gating network variables and *explanatory* expert network variables. Thus, for clarity, $\mathbf{x}_i^{(G)}$ and $\mathbf{x}_i^{(E)}$ will henceforth refer, respectively, to the possibly overlapping subsets of gating and expert network covariates, such that $\mathbf{x}_i = \left\{\mathbf{x}_i^{(G)} \cup \mathbf{x}_i^{(E)}\right\}$, with the dimensions of the associated design matrices given by $d_G + 1$ and $d_E + 1$. Higher order terms, transformations, and interaction effects between covariates are also allowed in both networks.

2.2 Gaussian Parsimonious Clustering Models

Parsimony has been considered extensively in the model-based clustering literature. In particular, the volume of work on Gaussian and/or parsimonious mixtures has increased hugely since the work of Banfield & Raftery (1993) and Celeux & Govaert (1995). These works introduced the family of GPCMs, which are implemented in the popular R package `mclust` (Scrucca et al., 2016). The influence of GPCMs is clear on many other works which obtain parsimony in the component covariance matrices; e.g., using constrained factor-analytic structures (McNicholas & Murphy, 2008), the multivariate t-distribution and associated tEIGEN family (Andrews & McNicholas, 2012), and the multivariate contaminated normal distribution (Punzo & McNicholas, 2016).

Parsimonious covariance matrix parameterisations are obtained in GPCMs by means of imposing constraints on the components of an eigen-decomposition of the form $\mathbf{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^\top$, where λ_g is a scalar controlling the volume, $\mathbf{A}_g$ is a diagonal matrix, with entries proportional to the eigenvalues of $\mathbf{\Sigma}_g$ with $\det(\mathbf{A}_g) = 1$, specifying the shape of the density contours, and $\mathbf{D}_g$ is $p \times p$ orthogonal matrix, the columns of which are the eigenvectors of $\mathbf{\Sigma}_g$, governing the corresponding ellipsoid's orientation. Imposing constraints reduces the number of free covariance parameters from $Gp\,(p+1)\,/2$ in the unconstrained (VVV) model. This is desirable when p is even moderately large. Thus, GPCMs allow for intermediate component covariance matrices lying between homoscedasticity and heteroscedasticity. Table 1 summarises the geometric characteristics of the GPCM constraints, which are then shown in Figure 2.

Note for models with names ending with I that the number of parameters is linear in the data dimension p. Thus, the diagonal models are especially parsimonious and useful in $n \leq p$ settings. While there are 2 variance parameterisations for mixtures of univariate response data, and 14 covariance parameterisations for mixtures of multivariate response data, considering the equal mixing proportion constraint doubles the number of models available in each of these cases.

Table 1: Nomenclature, descriptions, and parameter counts of the parameterisations of the component covariance matrices $\mathbf{\Sigma}_g$ available under GPCMs, all of which are available when there is no dependency in any way on covariates. † indicates availability in the first four special cases of the Gaussian MoE framework shown in Figure 1 and the MoEClust family; • indicates other models available in the MoEClust family. While all models are possible when $G = 1$, they are all equivalent to one of the highlighted available models, otherwise missing entries correspond to models which are never available. The other central columns refer to $G > 1$ settings.

Name	Model	$G=1$	$n>p$	$n \leq p$	Distribution	Volume	Shape	Orientation	Covariance Parameters
E	σ	†	•		(univariate)	equal			1
V	σ_g		†		(univariate)	variable			G
EII	$\lambda \boldsymbol{\mathcal{I}}$	†	•	•	spherical	equal	equal	—	1
VII	$\lambda_g \boldsymbol{\mathcal{I}}$		•	•	spherical	variable	equal	—	G
EEI	$\lambda \mathbf{A}$	•	•	•	diagonal	equal	equal	axis-aligned	p
VEI	$\lambda_g \mathbf{A}$		•	•	diagonal	variable	equal	axis-aligned	$G+(p-1)$
EVI	$\lambda \mathbf{A}_g$		•	•	diagonal	equal	variable	axis-aligned	$1+G(p-1)$
VVI	$\lambda_g \mathbf{A}_g$		†	†	diagonal	variable	variable	axis-aligned	Gp
EEE	$\lambda \mathbf{DAD}^\top$	•	•		ellipsoidal	equal	equal	equal	$p(p+1)/2$
EVE	$\lambda \mathbf{D}\mathbf{A}_g\mathbf{D}^\top$		•		ellipsoidal	equal	variable	equal	$1+p(p-1)/2+G(p-1)$
VEE	$\lambda_g \mathbf{DAD}^\top$		•		ellipsoidal	variable	equal	equal	$G+p(p-1)/2+(p-1)$
EEV	$\lambda \mathbf{D}_g\mathbf{A}\mathbf{D}_g^\top$		•		ellipsoidal	equal	equal	variable	$1+Gp(p-1)/2+(p-1)$
VEV	$\lambda_g \mathbf{D}_g\mathbf{A}\mathbf{D}_g^\top$		•		ellipsoidal	variable	equal	variable	$G+Gp(p-1)/2+(p-1)$
EVV	$\lambda \mathbf{D}_g\mathbf{A}_g\mathbf{D}_g^\top$		•		ellipsoidal	equal	variable	variable	$1+Gp(p-1)/2+(p-1)$
VVE	$\lambda_g \mathbf{D}\mathbf{A}_g\mathbf{D}^\top$		•		ellipsoidal	variable	variable	equal	$G+p(p-1)/2+G(p-1)$
VVV	$\lambda_g \mathbf{D}_g\mathbf{A}_g\mathbf{D}_g^\top$		†		ellipsoidal	variable	variable	variable	$Gp(p+1)/2$

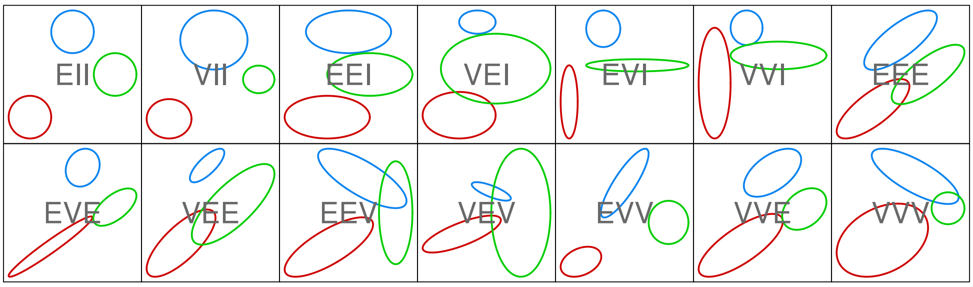


Figure 2: Ellipses of isodensity for each of the 14 parsimonious eigen-decomposition covariance parameterisations for multivariate data in GPCMs, with three components in two dimensions.

2.3 The MoEClust Family of Models

Interest lies in bringing parsimonious covariance structures to Gaussian MoE models with network-specific subsets of covariates:

$$f\left(\mathbf{y}_i \,|\, \mathbf{x}_i\right) = \sum_{g=1}^{G} \tau_g\left(\mathbf{x}_i^{(G)}\right)\phi\Big(\mathbf{y}_i \mid \boldsymbol{\theta}_g\left(\mathbf{x}_i^{(E)}\right) = \Big\{\widetilde{\mathbf{x}}_i^{(E)}\boldsymbol{\gamma}_g, \boldsymbol{\Sigma}_g\Big\}\Big),$$

where $\boldsymbol{\Sigma}_g$ can follow any of the GPCM constraints outlined in Table 1. It is equivalent to say that interest lies in incorporating covariate information into the GPCM model family. Using the covariance constraints, combined with the six special cases of the MoE model described in Section 2.1.1, yields the MoEClust family of models, which are capable of dealing with correlated responses and offering additional parsimony in the component densities compared to current implementations of Gaussian MoE models, by virtue of allowing the size, volume, shape, and/or orientation to be equal or unequal across components. For MoE models, every continuous covariate added to the gating and expert networks introduces $G-1$ and Gp additional regression parameters, respectively. Parsimonious MoEClust models allow the increase in the number of regression parameters to be offset by the reduction in the number of covariance parameters. This can be advantageous when model selection is conducted using information criteria which include penalty terms based on parameter counts (see Section 4.2).

2.4 Existing Models and Software

A number of tools for fitting MoE models are available in the R programming environment (R Core Team, 2020). These include `flexmix` (Grün & Leisch, 2007, 2008), `mixtools` (Benaglia et al., 2009), and others. Tools for fitting GPCMs without covariates include `mclust` (Scrucca et al., 2016) and `Rmixmod` (Lebret et al., 2015).

The `flexmix` package (Grün & Leisch, 2007, 2008) can accommodate the full range of MoE models outlined in Section 2.1.1, excluding those for which $\boldsymbol{\tau}$ is constrained to be equal, in the case of univariate $\mathbf{y}_i$, though only models with unequal variance can be fitted. The user can specify the form of the GLM and covariates (if any) to be used in the gating and expert networks, for which the package has a similar interface to the `glm` functions within R. In the case of a multivariate continuous response, there is functionality for multivariate Gaussian component distributions though only for models without expert network covariates. Furthermore, only the VVI and VVV constraints and models with unequal mixing proportions or gating concomitants are facilitated.

For univariate data, the mixtools package (Benaglia et al., 2009) can accommodate the expert network MoE model with equal or unequal variance; it can also accommodate the full MoE model, though only for $G = 2$, with unequal variance, and with the restriction that all covariates enter both part of the model. The package allows for nonaparametric estimation of the functional form for the mixing proportions (gating networks) and the component densities (expert networks), so it offers further flexibility beyond flexmix in these cases. However, the multivariate models in mixtools use the local independence assumption, so it does not directly offer the facility to model multivariate Gaussian component densities with non-diagonal covariance matrices. Furthermore, multivariate response models in mixtools do not yet incorporate covariates in any way, and the equal mixing proportions constraint is not facilitated in any way either.

The mclust package (Scrucca et al., 2016) and Rmixmod package (Lebret et al., 2015) can accommodate the full range of covariance constraints in Table 1, and are thus examples of existing software which can fit GPCMs, but only using the standard finite mixture model (model (a) in Figure 1) or the equal mixing proportions mixture model; i.e., they do not facilitate dependency on covariates in any way.

Another important contribution in this area is by Dang & McNicholas (2015). This work introduces eigen-decomposition parsimony to the MoE framework, though only for the expert network MoE model and the full MoE model. However, for the full MoE model, all covariates are assumed to enter into both parts of the model. Thus, the MoEclust model family completes the work of Dang & McNicholas (2015) by considering all six special cases of the MoE framework, whereby different subsets of covariates can enter either, neither, or both the component densities and/or component weights, as well as models with equal mixing proportions. In addition, our unifying MoEClust framework also incorporates such parsimonious models for univariate response data.

Finally, it should be noted that eigen-decomposition parsimony has been introduced to the alternative CWM framework, in which all covariates enter the same part of the model, by Dang et al. (2017), for the multivariate Gaussian distributions of both the response variables and the covariates, assuming only continuous covariates; see also Punzo & Ingrassia (2015) for eigen-decomposition parsimony applied to the covariates only. The flexCWM package (Mazza et al., 2018) allows GPCM covariance structures in the distribution of the continuous covariates only, though only univariate responses are accommodated. It also allows, simultaneously or otherwise, covariates of other types, as well as omitting the distribution for the covariates entirely, leading to non-parsimonious mixtures of regressions, with or without concomitant variables.

3 Model Fitting via EM

To estimate the parameters of MoEClust models, we focus on maximum likelihood estimation using the EM algorithm (Dempster et al., 1977). This is outlined first for MoE models in Section 3.1 and then extended to MoEClust models in Section 3.2. Model fitting details are described chiefly for the full MoE model only, for simplicity. A simple trick involving the residuals of the weighted linear regressions in the expert network assists fitting when using GPCM constraints. A uniform noise component to capture outlying non-Gaussian observations is added in Section 3.3. When gating concomitants are present, the noise component is treated in two different ways.

3.1 Fitting MoE Models

For the full mixture of experts model, the likelihood is of the form

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \tau_g\big(\mathbf{x}_i^{(G)}\big) \phi\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big)\big),$$

where $\tau_g\big(\mathbf{x}_i^{(G)}\big)$ and $\boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big)$ are as defined by (1). The data are augmented by imputing the latent cluster membership indicator $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})^\top$. Thus, the conditional distribution of $\big(\mathbf{y}_i, \mathbf{z}_i \,|\, \mathbf{x}_i\big)$ is of the form

$$f\big(\mathbf{y}_i, \mathbf{z}_i \,|\, \mathbf{x}_i\big) = \prod_{g=1}^{G} \Big[\tau_g\big(\mathbf{x}_i^{(G)}\big) \phi\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big)\big)\Big]^{z_{ig}}.$$

Hence, the complete data likelihood is of the form

$$\mathcal{L}_c(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \mathbf{Y}, \mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} \prod_{g=1}^{G} \Big[\tau_g\big(\mathbf{x}_i^{(G)}\big) \phi\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big)\big)\Big]^{z_{ig}},$$

and the complete data log-likelihood has the form

$$\begin{aligned} \ell_c(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \mathbf{Y}, \mathbf{X}, \mathbf{Z}) &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \Big[\log \tau_g\big(\mathbf{x}_i^{(G)}\big) + \log \phi\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big)\big)\Big] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \tau_g\big(\mathbf{x}_i^{(G)}\big) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \phi\big(\mathbf{y}_i \,|\, \boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big)\big). \end{aligned} \tag{4}$$

The iterative EM algorithm for MoE models follows in a similar manner to that for standard mixture models. It consists of an E-step (expectation) which replaces for each observation the missing data $\mathbf{z}_i$ with their expected values $\widehat{\mathbf{z}}_i$, followed by a M-step (maximisation) which maximises the expected complete data log-likelihood, computed with the estimates $\widehat{\mathbf{Z}} = (\widehat{\mathbf{z}}_1, \ldots, \widehat{\mathbf{z}}_n)$, to provide estimates of the component weight parameters $\widehat{\tau}_g\big(\mathbf{x}_i^{(G)}\big)$ and the component parameters $\widehat{\boldsymbol{\theta}}_g\big(\mathbf{x}_i^{(E)}\big)$. Aitken's acceleration criterion is used to assess convergence of the non-decreasing sequence of log-likelihood estimates (Böhning et al., 1994). Parameter estimates produced on convergence achieve at least a local maximum of the likelihood function. Upon convergence, cluster memberships are estimated via the maximum *a posteriori* (MAP) classification. The E-step involves computing

$$\widehat{z}_{ig}^{(t+1)} = \mathbb{E}\Big(z_{ig} \;\Big|\; \mathbf{y}_i, \mathbf{x}_i, \widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\gamma}}^{(t)}, \widehat{\boldsymbol{\Sigma}}^{(t)}\Big) = \frac{\widehat{\tau}_g^{(t)}\big(\mathbf{x}_i^{(G)}\big) \phi\big(\mathbf{y}_i \,|\, \widehat{\boldsymbol{\theta}}_g^{(t)}\big(\mathbf{x}_i^{(E)}\big)\big)}{\sum_{h=1}^{G} \widehat{\tau}_h^{(t)}\big(\mathbf{x}_i^{(G)}\big) \phi\big(\mathbf{y}_i \,|\, \widehat{\boldsymbol{\theta}}_h^{(t)}\big(\mathbf{x}_i^{(E)}\big)\big)},$$

where $\big\{\widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\gamma}}^{(t)}, \widehat{\boldsymbol{\Sigma}}^{(t)}\big\}$ are the estimates of the parameters in the gating and expert networks on the t-th iteration of the EM algorithm.

For the M-step, we notice that the complete data log-likelihood in (4) can be considered as a separation into the portion due to the gating network and the portion due to the expert network. Thus, the expected complete data log-likelihood (5) can be maximised separately under the EM framework:

$$\begin{aligned}\mathbb{E}\Big[\ell_c\big(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\Sigma}\,|\,\mathbf{Y},\mathbf{X},\mathbf{Z},\widehat{\boldsymbol{\beta}}^{(t)},\widehat{\boldsymbol{\gamma}}^{(t)},\widehat{\boldsymbol{\Sigma}}^{(t)}\big)\Big] &= \sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{ig}^{(t+1)}\log\tau_g\big(\mathbf{x}_i^{(G)}\big)\\ &+\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{ig}^{(t+1)}\log\phi\big(\mathbf{y}_i\,|\,\boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big)\big).\end{aligned} \tag{5}$$

The first term is of the same form as a MLR model, here written with component 1 as the baseline reference level, for identifiability reasons:

$$\log\frac{\tau_g\big(\mathbf{x}_i^{(G)}\big)}{\tau_1\big(\mathbf{x}_i^{(G)}\big)}=\log\frac{\Pr\big(\widehat{z}_{ig}^{(t+1)}=1\big)}{\Pr\big(\widehat{z}_{i1}^{(t+1)}=1\big)}=\widetilde{\mathbf{x}}_i^{(G)}\boldsymbol{\beta}_g\ \forall\ g\geq 2,\ \text{where}\ \boldsymbol{\beta}_1=(0,\ldots,0)^\top.$$

Thus, methods for fitting such models can be used to maximise this term and estimate the parameters in the gating network. The second term is of the same form as fitting G separate weighted multivariate linear regressions, and thus methods for fitting such models can be used to estimate the expert network parameters. Note that these are multivariate in the sense of a multivariate outcome $\mathbf{y}_i$; the associated design matrix having d_E+1 columns means these regressions are possibly also multivariate in terms of the explanatory variables. Thus, fitting MoE models is straightforward in principle.

3.2 Fitting MoEClust Models

Maximising the second term in (5), corresponding to the expert network, gives rise to the following expression

$$\begin{aligned}-\frac{1}{2}\Big(&p\log 2\pi+\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{ig}^{(t+1)}\log|\boldsymbol{\Sigma}_g|+\\ &\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{ig}^{(t+1)}\big(\mathbf{y}_i-\widetilde{\mathbf{x}}_i^{(E)}\boldsymbol{\gamma}_g\big)^\top\boldsymbol{\Sigma}_g^{-1}\big(\mathbf{y}_i-\widetilde{\mathbf{x}}_i^{(E)}\boldsymbol{\gamma}_g\big)\Big).\end{aligned} \tag{6}$$

When the same set of regressors are used for each dependent variable, as is always the case for MoEClust models, or when $\boldsymbol{\Sigma}_g$ is diagonal, it can be shown that $\boldsymbol{\gamma}_g$ does not depend on $\boldsymbol{\Sigma}_g$, much like a Seemingly Unrelated Regression model (SUR; Zellner, 1962). We first estimate $\widehat{\boldsymbol{\gamma}}_g$ and then $\widehat{\boldsymbol{\Sigma}}_g$. Fitting G separate multivariate regressions (weighted by $\widehat{z}_{ig}$), yields G sets of $n\times p$ SUR residuals $\widehat{\mathbf{r}}_{ig}=\mathbf{y}_i-\widetilde{\mathbf{x}}_i^{(E)}\widehat{\boldsymbol{\gamma}}_g$ which, crucially, satisfy $\sum_{i=1}^{n}\widehat{z}_{ig}\widehat{\mathbf{r}}_{ig}=0$. Thus, maximising (6) is equivalent to minimising

$$\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{ig}^{(t+1)}\log|\boldsymbol{\Sigma}_g|+\sum_{i=1}^{n}\sum_{g=1}^{G}\widehat{z}_{ig}^{(t+1)}\widehat{\mathbf{r}}_{ig}^\top\boldsymbol{\Sigma}_g^{-1}\widehat{\mathbf{r}}_{ig}\,, \tag{7}$$

which is of the same form as the criterion used in the M-step of a standard Gaussian finite mixture model with component covariance matrices $\widehat{\boldsymbol{\Sigma}}$, component means equal to zero, and new augmented data set $\widehat{\mathbf{R}}$. Thus, when estimating the component covariance matrices via (7), the same M-step function as used within `mclust` can be applied to augmented data, constructed so that each observation is represented as follows:

1. Stack the G sets of SUR residuals into the $(n \times G) \times p$ matrix $\widehat{\mathbf{R}}$:

$$\widehat{\mathbf{R}} = \left[\begin{array}{cccc}
\widehat{r}_{111} & \widehat{r}_{112} & \cdots & \widehat{r}_{11p}\\
\widehat{r}_{211} & \widehat{r}_{212} & \cdots & \widehat{r}_{21p}\\
\vdots & \vdots & \ddots & \vdots\\
\widehat{r}_{n11} & \widehat{r}_{n12} & \cdots & \widehat{r}_{n1p}\\ \hline
\widehat{r}_{121} & \widehat{r}_{122} & \cdots & \widehat{r}_{12p}\\
\widehat{r}_{221} & \widehat{r}_{222} & \cdots & \widehat{r}_{22p}\\
\vdots & \vdots & \ddots & \vdots\\
\widehat{r}_{n21} & \widehat{r}_{n22} & \cdots & \widehat{r}_{n2p}\\ \hline
\vdots & \vdots & \ddots & \vdots\\ \hline
\widehat{r}_{1G1} & \widehat{r}_{1G2} & \cdots & \widehat{r}_{1Gp}\\
\widehat{r}_{2G1} & \widehat{r}_{2G2} & \cdots & \widehat{r}_{2Gp}\\
\vdots & \vdots & \ddots & \vdots\\
\widehat{r}_{nG1} & \widehat{r}_{nG2} & \cdots & \widehat{r}_{nGp}
\end{array}\right]$$

2. Create the $(n \times G) \times G$ block-diagonal matrix $\widehat{\boldsymbol{\zeta}}$ from the columns of $\widehat{\mathbf{Z}}$:

$$\widehat{\boldsymbol{\zeta}} = \left[\begin{array}{cccc}
\widehat{z}_{11} & 0 & \cdots & 0\\
\widehat{z}_{21} & 0 & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots\\
\widehat{z}_{n1} & 0 & \cdots & 0\\ \hline
0 & \widehat{z}_{12} & \cdots & 0\\
0 & \widehat{z}_{22} & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots\\
0 & \widehat{z}_{n2} & \cdots & 0\\ \hline
\vdots & \vdots & \ddots & \vdots\\ \hline
0 & 0 & \cdots & \widehat{z}_{1G}\\
0 & 0 & \cdots & \widehat{z}_{2G}\\
\vdots & \vdots & \ddots & \vdots\\
0 & 0 & \cdots & \widehat{z}_{nG}
\end{array}\right]$$

Structuring the model in this manner allows GPCM covariance structures to be easily imposed on Gaussian MoE models with gating and/or expert network covariates. In the end, the M-step involves three sub-steps, each using the current estimate of $\widehat{\mathbf{Z}}$: i) estimating the gating network parameters $\widehat{\boldsymbol{\beta}}_g$ and hence the component weights $\widehat{\tau}_g\big(\mathbf{x}_i^{(G)}\big)$ via MLR, ii) estimating the expert network parameters $\widehat{\boldsymbol{\gamma}}_g$ and hence the component-specific means via weighted multivariate multiple linear regression, and iii) estimating the constrained component covariance matrices $\widehat{\boldsymbol{\Sigma}}_g$ using the augmented data set comprised of SUR residuals, as outlined above.

In the absence of covariates in the gating and/or expert networks, under the special cases outlined in Section 2.1.1, their respective contribution to (5) is maximised as per the corresponding term in a standard GPCM. In other words, the gating and expert networks without covariates can be seen as regressions with only an intercept term. Thus, the augmented data structure is not required when there are no expert covariates and the formula for estimating $\boldsymbol{\tau}$ in the absence of concomitant variables is $\widehat{\tau}_g = n^{-1}\sum_{i=1}^{n}\widehat{z}_{ig}$, rather than (1). As described in Section 2.1.1, it is sometimes useful to expand the model family further by considering more parsimonious alternatives to the special cases of models (a) and (b) in Figure 1, where gating covariates are omitted, by constraining the mixing proportions to be equal and fixed, i.e. $\tau_g = {}^1\!/_G \ \forall\ g$. Similarly, removing the corresponding regression intercept(s) from the part(s) of the model where covariates enter can yield further parsimony in appropriate settings, e.g. when there are strong *a priori* physical reasons for believing $\mathbb{E}\left(\mathbf{Y}\,|\,\mathbf{X}^{(E)} = \mathbf{0}\right) = \mathbf{0}$ (Eisenhauer, 2003).

3.3 Adding a Noise Component

For models with expert network covariates, and/or when the volume and/or shape differ across components, the mixture likelihood is unbounded. We restrict our interest only to solutions for which the log-likelihood at convergence is finite. As per the `eps` argument to the `mclust` R package's `emControl` function (Scrucca et al., 2016), we monitor the conditioning of the covariances and add a tolerance parameter (set to the relative machine precision, i.e. `2.220446e-16` on IEEE compliant machines) to the M-step estimation of the component covariances to control termination of the EM algorithm on the basis of small eigenvalues. For models with unconstrained $\boldsymbol{\Sigma}_g$, each cluster must contain at least $p+1$ units to avoid computational singularity. Thus, in practice, such spurious solutions with infinite likelihood occur especially for higher G values, whereby either solutions with empty components reduce to ones with fewer components, or uninteresting solutions with degenerate components containing too few units or even singletons are found. Sensible initial allocations (see Section 4.1) and/or the equal mixing proportion constraint, which help avoid empty or otherwise poorly populated clusters, can help to alleviate this problem. García-Escudero et al. (2018) offer an excellent discussion of the notions of spurious solutions and degenerate components.

Further extending MoEClust models via the inclusion of an additional uniform noise component can also help in addressing these issues, by capturing outlying observations which do not fit the prevailing pattern of Gaussian clusters and thus would otherwise be assigned to (possibly many) small clusters. In particular, the noise component for encompassing clusters with non-Gaussian distributions is here distributed as a homogeneous spatial Poisson process, as per Banfield & Raftery (1993). Such a noise component can be included regardless of where covariates (if any) enter, and regardless of the GPCM constraints employed. Model-fitting via the EM algorithm is not greatly complicated by the addition of a noise component, though it is required to estimate V, the hypervolume of the region from which the response data have been drawn, or to consider V as an independent tuning parameter as per Hennig & Coretto (2008), especially if $n \leq p$. For univariate responses V is given by the range of $y_1, \ldots, y_n$. For multivariate data, V can be estimated by the hypervolume of the convex hull, ellipsoid hull, or smallest hyperrectangle enclosing the data. We focus on the latter method.

For initialisation, a column in which each entry is τ_0 (the guess of the prior probability that observations are noise) is appended to the starting $\mathbf{Z}$ matrix, with other columns corresponding to non-noise components then multiplied by $1 - \tau_0$. The initial τ_0 should not be too high; it is set to 0.1 here. For models with a noise component and no gating concomitants, the mixing proportions can be, as before, either constrained or unconstrained. In the latter case, we estimate τ_0 and then constrain the remaining proportions. We add the extension that concomitants, when present, are allowed to affect (8) or not affect (9) the mixing proportion of the noise component. Henceforth, for clarity, we refer to these settings as the gated noise (NG) and non-gated noise (NGN) models, respectively. The NGN model assumes τ_0 is constant across observations and covariate patterns. It is thus the more parsimonious model; it requires only 1 extra gating network parameter, rather than $d_G + 1$ under the GN model, relative to models without a noise component, though it is only defined for $G \geq 2$.

$$\text{GN:} \quad f\big(\mathbf{y}_i \,|\, \mathbf{x}_i\big) = \sum_{g=1}^{G} \tau_g\big(\mathbf{x}_i^{(G)}\big)\phi\Big(\mathbf{y}_i \mid \boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big) = \Big\{\widetilde{\mathbf{x}}_i^{(E)}\boldsymbol{\gamma}_g, \boldsymbol{\Sigma}_g\Big\}\Big) + \frac{\tau_0\big(\mathbf{x}_i^{(G)}\big)}{V}. \tag{8}$$

$$\text{NGN:} \quad f\big(\mathbf{y}_i \,|\, \mathbf{x}_i\big) = \sum_{g=1}^{G} \tau_g\big(\mathbf{x}_i^{(G)}\big)\phi\Big(\mathbf{y}_i \mid \boldsymbol{\theta}_g\big(\mathbf{x}_i^{(E)}\big) = \Big\{\widetilde{\mathbf{x}}_i^{(E)}\boldsymbol{\gamma}_g, \boldsymbol{\Sigma}_g\Big\}\Big) + \frac{\tau_0}{V}. \tag{9}$$

4 Practical Issues

In this section, factors affecting the performance of MoEClust models are discussed; namely, the necessity of a good initial partition to prevent the EM algorithm from converging to a suboptimal local maximum (Section 4.1), and the necessity of model selection with regard to where and what covariates (if any) enter the model to yield further parsimony by reducing the number of gating and/or expert network regression parameters (Section 4.2). Novel strategies for dealing with both issues are proposed.

4.1 EM Initialisation

With regards to initialisation of the EM algorithm for $G > 1$ MoEClust models, model-based agglomerative hierarchical clustering and quantile-based clustering have been found to be suitable for multivariate and univariate data, respectively. Both `flexmix` and `mixtools` randomly initialise the allocations, despite the obvious computational drawback of the need to run the EM algorithm from multiple random starting points. However, when explanatory variables $\mathbf{x}_i^{(E)}$ enter the expert network, it is useful to use them to augment the initialisation strategy with extra steps. Algorithm 1 outlines the proposed initialisation strategy, similar to that of Ning et al. (2008). It takes the initial partition of the data (whether obtained by hierarchical clustering, random initialisation, or some other method) and iteratively reallocates observations in such a way that each subset can be well-modelled by a single expert.

When using a deterministic approach to obtain the starting partition for Algorithm 1, initialisation can be further improved by considering information in the expert network covariates to find a good clustering of the joint distribution of $\big(\mathbf{y}_i, \mathbf{x}_i^{(E)}\big)$. When $\mathbf{x}_i^{(E)}$ includes categorical or ordinal covariates, the model-based approach to clustering mixed-type data of McParland & Gormley (2016) can be employed at this stage, though this is not considered further here.

Algorithm 1: Iterative reallocation initialisation with expert network covariates

```
0 Concatenate the response data and expert network covariates into a matrix.
1 Obtain some non-overlapping hard starting partition $\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_G$.
2 Estimate the expert network regression $\eta_g(\boldsymbol{\gamma}_g, \cdot)$ on every subset $\{\boldsymbol{\Omega}_g\}_{g=1}^{G}$.
3 Compute the fitted values $\widehat{\mathbf{y}}_{ig} = \eta_g\big(\widehat{\boldsymbol{\gamma}}_g, \mathbf{x}_i^{(E)}\big)\ \forall\ (i, g)$ and hence the residuals $\widehat{\mathbf{r}}_{ig} = \mathbf{y}_i - \widehat{\mathbf{y}}_{ig}$.
4 Compute $\widehat{\boldsymbol{\Psi}}_g = \mathrm{Cov}\big(\widehat{\mathbf{R}}_g\big) = \frac{1}{n-d_E-1}\widehat{\mathbf{R}}_g^{\top}\widehat{\mathbf{R}}_g\ \forall\ g$.
5 Compute the squared Mahalanobis distance $\widehat{M}_{ig} = \mathrm{d}_{\mathrm{M}}^2\big(\mathbf{y}_i, \widehat{\mathbf{y}}_{ig}\big) = \widehat{\mathbf{R}}_g^{\top}\widehat{\boldsymbol{\Psi}}_g^{-1}\widehat{\mathbf{R}}_g$.
6 Let $k_i = \arg\min_{h\,\in\,\{1,\ldots,G\}}\big(\widehat{M}_{ih}\big)$ and reassign observation $i$ to subset $\boldsymbol{\Omega}_{k_i}$.
7 Repeat Steps 2–6 until convergence is achieved, i.e. until the partition ceases to change.
```

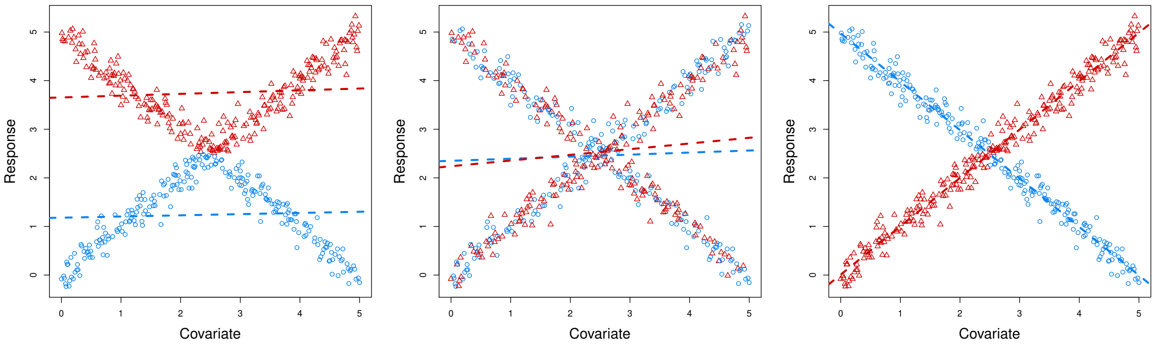
If at any stage a level is dropped from a categorical variable in subset $\boldsymbol{\Omega}_g$ the variable itself is dropped from the corresponding regressor for the observations with missing levels. Convergence of the algorithm is guaranteed and the additional computational burden incurred is negligible. By using the Mahalanobis distance metric (Mahalanobis, 1936), each observation is assigned to the cluster corresponding to the Gaussian ellipsoid to which it is closest. This has the added advantage of potentially speeding up the running of the EM algorithm. The estimates of $\widehat{\boldsymbol{\gamma}}_g$ at convergence are used as starting values for the expert network. The gating network is initialised by considering the partition itself at convergence as a discrete approximation of the gates.

While convergence is monitored via the partition itself, Algorithm 1 implicitly finds the hard partition which minimises the total intra-component regression error criterion

$$\sum_{g=1}^{G} \min_{\{\eta_g,\,\boldsymbol{\gamma}_g\}} \Bigg(\sum_{i\,\in\,\boldsymbol{\Omega}_g} \mathrm{d}_{\mathrm{M}}^2\Big(\mathbf{y}_i, \eta_g\big(\boldsymbol{\gamma}_g, \mathbf{x}_i^{(E)}\big)\Big)\Bigg). \tag{10}$$

However, there are a few small caveats. Firstly, it suffices to use the Euclidean distance in place of the Mahalanobis distance for applications to univariate response data. Secondly, the Moore-Penrose pseudo-inverse (Moore, 1920) or p-dimensional identity matrix $\boldsymbol{\mathcal{I}}_p$ is used in place of $\widehat{\boldsymbol{\Psi}}_g^{-1}$ when $n \leq p$. Finally, we note that Algorithm 1 applies only to the non-noise components; in the presence of a noise component, the $\widehat{\mathbf{Z}}$ matrix outputted by the algorithm at convergence is modified in the usual way.

Figure 3 illustrates the necessity of this procedure using a toy data set, with a single continuous covariate and a univariate response clearly arising from a mixture of two linear regressions, which otherwise would not be discerned without including the covariate in the initialisation routine via Algorithm 1. A further demonstration of the utility of this strategy is shown in Appendix B. Similar to the EM algorithm's susceptibility to local maxima, a limitation of our initialisation strategy is that the result at convergence may represent a suboptimal local minimum. However, the problem is transferred from the difficult task of initialising the EM algorithm to initialising Algorithm 1. Thus, it is feasible to repeat the algorithm with many different partitions and choose the best result, in the sense of minimising the criterion in (10), to initialise one run of the EM algorithm, since Algorithm 1 converges very quickly, requires much less computational effort than the EM algorithm itself, and generally reduces the number of required EM iterations. However, we caution against using the total intra-component regression error criterion to guide the inclusion of expert network covariates.



(a) Hierarchical clustering of the response variables only.

(b) Random allocation.

(c) Partition obtained via Algorithm 1.

Figure 3: Initial 2-component hard partitions on univariate data clearly arising from a mixture of two linear regressions, obtained using (a) agglomerative hierarchical clustering, (b) random allocation, and (c) Algorithm 1 applied to the initialisation in (b) upon convergence after 6 iterations, demonstrating the improvement achieved by incorporating expert network covariates into the initialisation strategy. Allocations are distinguished by blue circles and red triangles. Corresponding fitted lines are also shown.

4.2 Model Selection

Whether a variable should be considered as a covariate or part of the response is usually clear from the context of the data being clustered. However, within the suite of MoE models outlined in Section 2.1.1, it is natural to question which covariates, if any, are to be included, and if so in which part(s) of the MoE model. Unless the manner in which covariates enter is guided by the question of interest in the application under study, this is a challenging problem as the space of MoE models is potentially very large once variable selection for the covariates entering the gating and expert networks is considered. Thus, only models where covariates enter all mixture components or all component weights in a linear manner are typically considered in practice in order to restrict the size of the model search space. However, even within this reduced space, there are 2^r models to consider when $G = 1$ and 2^{2r} models to consider otherwise. Thus, the model space increases further if the number of components G is unknown.

Model comparison for the MoEClust family is even more challenging, especially for multivariate response data for which there are potentially 14 different GPCM covariance constraints to consider for models with $G \geq 2$ and 3 otherwise. When $p = 1$, there are 2 covariance constraints to consider when $G \geq 2$ and 1 otherwise. Considering constraints on the mixing proportions further increases the model search space. However, model selection can still be implemented in a similar manner to other model-based clustering methods: the Bayesian Information Criterion (BIC; Schwarz, 1978) and Integrated Completed Likelihood (ICL; Biernacki et al., 2000) have been shown to give suitable model selection criteria, both for the number of component densities (and thus clusters) required and for selecting covariates to include in the model. The number of free parameters in the penalty term for these criteria of course depends on the included gating and expert network covariates and the GPCM constraints employed.

For MoEClust models involving mixtures of GLMs, stepwise variable selection approaches can be used to find the optimal covariates for inclusion in either the multinomial logistic regression (gating network) or the weighted linear regression (expert network). Indeed, more parsimony can be achieved using variable selection, as there are a total of $G\left(d_G + 1\right) + Gp\left(d_E + 1\right)$ intercept and regression coefficients to estimate for a $G > 1$ full MoE model. However, the selected covariates may only be optimal for the given G and the given set of GPCM covariance matrix constraints. MoEClust models also allow for covariates entering only one part of the model. Thus, we propose a greedy stepwise search whereby each step could involve adding/removing a component or adding/removing a single covariate in either the gating or expert networks. We adopt a forward search, starting from a $G = 1$ model, as backward selection can be particularly cumbersome when r is large. In the considered applications, it sufficed to consider only additions (of components and covariates) rather than additions and removals in the sense that the same final model was obtained despite fewer models being evaluated over the course of the search. Hence, the recommended forward search algorithm proceeds as follows:

Algorithm 2: Greedy forward stepwise search for MoEClust models

```
1 Choose the best $G = 1$ model with no covariates among all allowable model types.
2 Either:
    • increase $G$ by 1,
    • add an explanatory variable to the expert network,
    • add a concomitant variable to the gating network (only when $G \geq 2$).
3 For every action in Step 2, consider the full range of allowable GPCM constraints.
4 Accept the change which yields the best improvement in terms of BIC or ICL.
5 Repeat Steps 2–4 until there is no further improvement in the selection criterion.
```

While one could consider changing the GPCM constraints as another potential action in Step 2 of Algorithm 2, our experience suggests that increasing G or adding covariates (especially in the expert network) can radically alter the covariance structure. Thus, we advise changing the GPCM constraints simultaneously and identifying the optimum action by first finding the optimum constraints for each action. While this is more computationally intensive than altering the GPCM constraints as a step in itself, this makes the search less likely to miss optimal models as it traverses the model space. See Appendix A for an example of how to conduct such a stepwise search using code from the `MoEClust` R package (Murphy & Murphy, 2020).

In certain special instances, some extra steps can be considered. When there are no gating network concomitants, a choice can be made, for each action, between fitted models with equal or unequal mixing proportions. We distinguish between G-component models without a noise component and models with $G-1$ Gaussian components plus an additional noise component. Thus, we recommend treating models with a noise component differently, by running a stepwise search for models excluding the possibility of a noise component, running a separate stepwise search starting from a $G = 0$ noise-only model, and ultimately choosing between the optimal models with and without a noise component identified by each search. In the presence of a noise component, one can also fit the GN and NGN models, given by (8) and (9) respectively, when evaluating every action involving models with gating network concomitants.

When r is not so prohibitively large as to render an exhaustive search infeasible, Gormley & Murphy (2010) demonstrate how model selection criteria such as the BIC can be employed to choose the appropriate number of components and guide the inclusion of covariates across the six special cases of the MoE model described in Section 2.1.1. Adapting this approach to MoEClust models where GPCM constraints must also be chosen requires fixing the covariates to be included in the component weights and densities and finding the G value and GPCM covariance structure which together optimise some criterion. Different fits with different combinations of covariates are then compared according to the same criterion. However, due to the highlighted computational difficulties when r is large, Algorithm 2 remains the recommended approach.

5 Results

The clustering performance of the MoEClust models is illustrated by application to two well-known data sets: univariate CO_2 data (Section 5.1) and multivariate data from the Australian Institute of Sports (Section 5.2). Additional results are provided for each data set in the Appendices. In particular, code examples (Appendix A) and details of the initialisation (Appendix B) for the CO_2 data and results of the stepwise search (Appendix C) for the AIS data are given.

Hereafter, any mention of methods for initialising the allocations, when covariates enter the expert network, refers to finding a single initial partition for Algorithm 1. The BIC and the stepwise search strategy outlined in Algorithm 2 were used to find the optimal number of components, choose the covariance type, and select the best subset of covariates, as well as where to put them. Results of exhaustive searches are also provided for demonstrative purposes. All results were obtained using the R package `MoEClust` (Murphy & Murphy, 2020).

5.1 CO_2 Data

As a univariate example of an application of MoEClust, data on the CO_2 emissions of $n = 28$ countries in the year 1996 (Hurn et al., 2003) are clustered, with Gaussian component densities. Studying the relationship between CO_2 and the covariate Gross National Product (GNP), both measured *per capita*, is of interest. As consideration is only being given to inclusion/exclusion of a single covariate in the gating and/or expert networks, an exhaustive search is feasible. A range of models ($G \in \{1, \ldots, 9\}$) are fitted, with either the equal (E) or unequal variance (V) models from Table 1. Quantile-based clustering of the CO_2 values is used to initialise Algorithm 1 when the expert network excludes GNP, otherwise hierarchical clustering of both CO_2 and GNP is used.

Table 2 gives BIC and ICL values for the top model under each of the six special cases of the MoE framework. The chosen model had $G = 3$, equal variances (i.e. the E constraint), equal mixing proportions, and GNP in the expert network; thus, this is an *equal mixing proportion expert network MoE model*. This model maximised both the BIC and ICL criteria, and was also identified by the forward stepwise search described in Algorithm 2, starting from a $G = 1$ model (BIC$=-163.90$), adding a component (BIC$=-163.16$), adding GNP to the expert network and changing to the V model type (BIC$=-157.20$), and finally adding a further component, constraining the mixing proportions, and changing back to the E model type (BIC$=-155.20$). Thereafter, neither adding a component nor adding GNP to the gating network improved the BIC. Code to reproduce both the exhaustive and stepwise searches using the `MoEClust` R package is given in Appendix A.

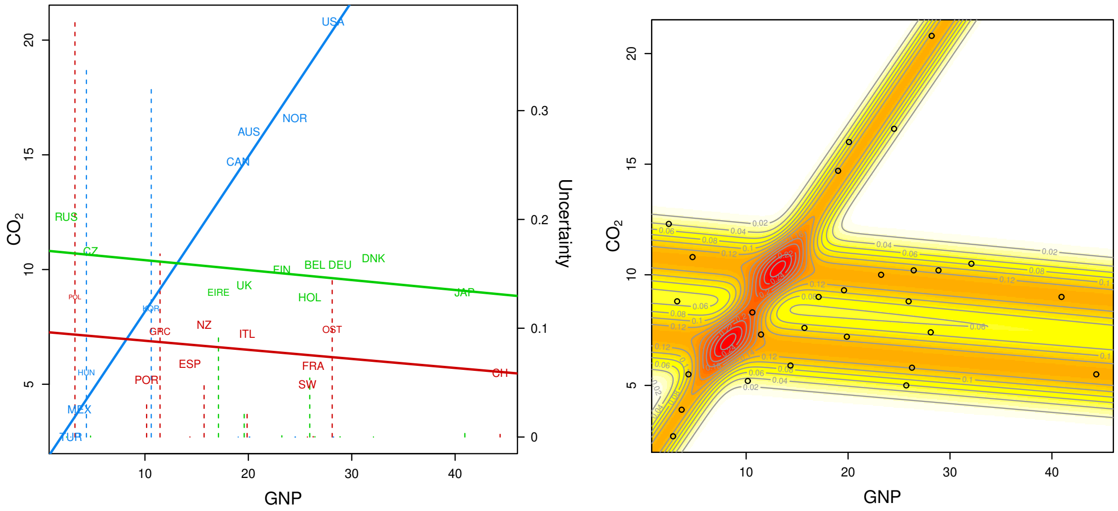
Table 2: The MoEClust BIC and ICL values of the top models under the six MoE special cases for the CO_2 data. Each row is optimal with respect to G and GPCM type, given the included covariates.

Special Case	Gating	Expert	G	GPCM	BIC	ICL
Mixture Model			2	E	−163.16	−163.91
Expert Network MoE Model		GNP	2	V	−157.20	−160.04
Gating Network MoE Model	GNP		2	E	−166.05	−166.68
Full MoE Model	GNP	GNP	2	V	−159.25	−161.47
Equal Mixing Proportion Mixture Model	Equal		2	V	−165.19	−184.71
Equal Mixing Proportion Expert Network MoE Model	Equal	GNP	3	E	**−155.20**	**−159.06**

Repeating both the exhaustive and stepwise searches with the addition of a noise component for all models also failed to yield any model with an improved BIC. The fourth row of Table 2 corresponds to a *full MoE*, with GNP included in both parts of the model; its sub-optimal BIC highlights the benefits of the model selection approach. The parameters of the optimal model are given in Table 3. Its fit is exhibited in Figure 4, which shows that the relationship between CO_2 and GNP is clustered around three different linear regression lines; one cluster of 8 countries with a large slope value and two equally-sized clusters, each with different intercepts but similar near-zero slope values. Clustering uncertainties, given by $\widehat{U}_i = \min_{g \in \{1,\ldots,\widehat{G}\}}(1 - \widehat{z}_{ig})$, are also shown.

Table 3: Estimated parameters of the optimal MoEClust model fit to the CO_2 data.

Parameter	Component 1	Component 2	Component 3
Proportion	1/3	1/3	1/3
(Intercept)	1.41	7.29	10.84
GNP	0.68	−0.04	−0.04
σ_g^2	0.98	0.98	0.98



(a) Fitted lines of the expert network GLMs. Text label size is proportional to a country's probability of belonging to its assigned cluster. Clustering uncertainty is also indicated by dotted vertical bars relating to the second y-axis. Colours correspond to the MAP classification.

(b) Heat map of the conditional density of the outcome variable CO$_2$, accounting for the gating and expert networks, the latter of which includes GNP as a covariate.

Figure 4: Scatter plots of GNP against CO$_2$ emissions for $n = 28$ countries with three linear regression components from the optimal MoEClust model with equal variances and mixing proportions.

The optimal model contains GNP in the expert network and has constraints on the component variances and mixing proportions. These are features of the MoEClust models which neither MoE nor GPCM models can fully accommodate. While `flexmix` and `mixtools` can fit the sub-optimal expert network MoE model in row four of Table 2, with unequal variances and mixing proportions (which achieves the second highest BIC value), our initialisation strategy ultimately leads to the same or higher BIC estimates. Across 50 random starts, BIC values of -157.29 and -157.20 are achieved using `flexmix` and `mixtools`, respectively. Among these random starts, BIC values as low as -163.67 are obtained. However, the `MoEClust` R package, with Algorithm 1 invoked, achieves a BIC of -157.20 using only a single initial partition. Using `MoEClust` without this initialisation strategy also yields the lower BIC value of -163.67. A further demonstration of the advantages of our initialisation strategy, using instead the optimal model for the the CO$_2$ data, is provided in Appendix B.

5.2 Australian Institute of Sport (AIS) Data

Various physical and hematological (blood) measurements were made on 102 male and 100 female athletes at the Australian Institute of Sport (AIS; Cook & Weisberg, 1994). The thirteen variables recorded in the study are detailed in Table 4.

Table 4: Australian Institute of Sports data variables. The $p = 5$ in the first column are hematological response variables and the others, the $r = 8$ covariates, are physical measurements for the athlete.

Response	Description	Covariate	Description (Units)
RCC	red cell count	BMI	body mass index (kg/m^2)
WCC	white cell count	SSF	sum of skin folds (mm)
Hc	Hematocrit	Bfat	body fat percentage (%)
Hg	Hemoglobin	LBM	lean body mass (kg)
Fe	plasma ferritin	Ht	height (cm)
	concentration	Wt	weight (kg)
		sex	a factor with levels: female, male
		sport	a factor with levels: Basketball, Field, Gymnastics, Netball, Rowing, Swimming, Tennis, Track 400m, Track Sprint, Water Polo

MoEClust models are used to investigate the clustering structure in the athletes' hematological measurements and investigate how covariates may influence these measurements and the clusters. Cluster allocations are initialised using model-based agglomerative hierarchical clustering. Results of the forward stepwise model search described in Algorithm 2, with all covariates considered for inclusion, are given in Appendix C. The optimal model (BIC$=-4010.14$) is a 2-component EVE *equal mixing proportion expert network MoE model*, which thus has clusters of equal size, volume, and orientation, and unequal shape. Notably, the only covariate (sex), only enters in one part of the model, the expert network.

The sub-optimal BIC values for the best model with all covariates in both parts of the model ($G = 2$, VVE, BIC$=-4563.12$), the best model with all covariates in the expert network only ($G = 1$, EEE, BIC$=-4234.79$), regardless of $\boldsymbol{\tau}$ being constrained or not, and the best model with all covariates in the gating network only ($G = 2$, VEE, BIC$=-4092.79$), highlight the need for the model selection strategy employed. As the optimal model uses the EVE constraints, it has 19 covariance parameters; an otherwise exactly equivalent VVV model, having 30 such parameters, yields a lower BIC of -4056.19, thus showcasing the benefits of the parsimonious covariance constraints. The difference of 11 covariance parameters between these models is exactly one more than the number of regression parameters introduced by the expert network covariate.

Subsequently, and purely for the purposes of comparing certain special cases of interest, an exhaustive search over a range of MoEClust models is conducted, with $G \in \{1, \ldots, 9\}$. This is rendered feasible by only considering the covariates BMI and sex; allowing either, neither, or both to enter either, neither, or both of the gating and expert networks. Note that BMI is itself computed using the covariates measuring weight (Wt) and height (Ht). With 3 permissible covariance parameterisations for the single component models (i.e. those without gating network covariates) and 14 otherwise, 16 possible combinations of gating and/or expert network covariate settings, and consideration also being given to $G > 1$ models with equal mixing proportions, this still requires fitting $2,252$ MoEClust models. However, some 26 spurious solutions were found (and ultimately discarded), particularly for higher values of G, in the sense that models with empty components or degenerate components with few observations reduced to equivalent models with fewer non-empty components (see Section 3.3). Table 5 gives the BIC and ICL values of a selection of these fitted models, representing the optimal models for certain special cases of interest.

Table 5: The BIC and ICL values for a selection of MoEClust models fitted to the Australian Institute of Sports data. Rows 1 and 2 give the optimal models under settings available in `flexmix`; models without expert network covariates, using either the VVV or VVI covariance constraints. Among the more general MoEClust family, the last row gives the top model according to the ICL criterion and the remaining rows give the top models according to the BIC criterion for each of the six special cases of the MoE framework. Thus, row 3 corresponds to the optimal model according to `mclust`.

Rank (BIC)	Gating	Expert	G	GPCM	BIC	ICL	No. Parameters
198	sex		2	VVV	−4113.31	−4121.32	42
880	sex		5	VVI	−4319.85	−4345.55	58
293			2	EVE	−4146.16	−4201.61	30
3		sex	2	EVE	−4015.35	−4059.54	40
24	sex		3	EVE	−4037.32	−4066.66	42
2	BMI	sex	2	EVE	−4013.40	−4074.11	41
269	Equal		2	EVE	−4140.98	−4192.21	29
1	Equal	sex	2	EVE	**−4010.14**	−4057.87	39
26	BMI, sex		3	EEE	−4038.75	**−4043.01**	36

Clearly, the inclusion of covariates improves the fit compared to GPCM models. Similarly, using GPCM covariance constraints improves the fit compared to standard Gaussian MoE models. In particular, it is notable that the optimal models using the VVV and VVI constraints only have covariates enter the gating network. This suggests that the parsimony afforded by the remaining GPCM settings somewhat offsets the number of regression parameters introduced to the expert network.

The top three models according to BIC all have 2 components, the EVE covariance constraints, and the covariate sex in the expert network; they differ only in their treatment of the gating network. Models with equal and unequal mixing proportions, and with BMI as a gating concomitant, have zero, one, and two associated gating network parameters, respectively. The optimal model has equal mixing proportions and was also identified above via Algorithm 2. The full MoE model with BMI in the gating network and sex in the expert network is an interesting case as it does not fit the framework of Dang & McNicholas (2015), which assumes that when covariates enter the model, they enter in both parts. The best such model has 'sex' in both networks ($G = 2$, EVE) and achieves a BIC of -4020.22 with a corresponding rank of 8.

Up to now, models with a noise component have not yet been considered for the AIS data. Thus, another stepwise search is conducted, including a noise component for all candidate models and starting from a $G = 0$ noise-only model (see Appendix C). Consideration was also given to both the GN and NGN model types, in (8) and (9) respectively, where models included gating concomitants, and to models with equal/unequal mixing proportions for the non-noise components for models without gating concomitants. The optimal full MoE model thus found has two EEE Gaussian clusters and an additional noise component. The covariate 'sex' enters the expert network (see Table 6). Both 'SSF' and 'Ht' enter the gating network, though not for the noise component, which has a constant mixing proportion ($\widehat{\tau}_0 \approx 0.08$), as per the NGN model in (9). Thus, the Gaussian clusters have equal volume, shape, and orientation, but unequal size. This model achieves a BIC value of -3989.83, which compares favourably to the previously optimal model from Table 5, adding a noise component to a model otherwise identical to the optimal model from Table 5 (BIC$=-3992.81$), and to models with a noise component but no stepwise selection of covariates (or no covariates at all).

Table 6: Coefficients of the expert network linear regressions for the $G = 2$ Gaussian clusters in the optimal 'full' MoEClust model (with an extra noise component and gating concomitants entering the non-noise clusters only) fit to the AIS data, with female as the reference level for the explanatory variable 'sex'.

	RCC	WCC	Hc	Hg	Fe
Cluster 1					
(Intercept)	4.56	6.89	42.33	14.08	49.73
sexmale	0.42	0.12	2.95	1.30	28.19
Cluster 2					
(Intercept)	4.26	6.93	38.91	13.11	59.70
sexmale	0.86	0.59	7.36	2.80	132.66

The gating network has an intercept of 10.58 and slope coefficients of 0.04 (SSF) and -0.08 (Ht) with corresponding odds ratios of 1.04 and 0.93. Thus, higher SSF values increase the probability of belonging to the second Gaussian cluster, to which taller athletes are less likely to belong, and the probability of belonging to the noise component is constant. Though every observation has its own mean parameter in the presence of expert covariates, given by the fitted values of the expert network (shown in Table 6), the means are summarised in Table 7 by the posterior mean of the fitted values of the model according to (11). The noise component is accounted for by $\overline{\boldsymbol{\mathcal{V}}}$, the p-dimensional centroid of the region used to estimate V:

$$\widehat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^n \widehat{z}_{ig}\widehat{\mathbf{y}}_i}{\sum_{i=1}^n \widehat{z}_{ig}} = \frac{\sum_{i=1}^n \widehat{z}_{ig}\left(\sum_{g=1}^G \widehat{z}_{ig}\left(\widetilde{\mathbf{x}}_i^{(E)}\widehat{\boldsymbol{\gamma}}_g\right) + \widehat{z}_{i0}\overline{\boldsymbol{\mathcal{V}}}\right)}{\sum_{i=1}^n \widehat{z}_{ig}}. \tag{11}$$

Given that there exists a binary variable, 'sex', in the expert network for the optimal MoEClust model, there are effectively four Gaussian components plus an additional noise component. By virtue of the EEE constraint on the Gaussian components, all four components and thus both clusters in fact share the same covariance matrix. Components 1 and 2, corresponding to females and males in Cluster 1, share the same covariance matrix but differ according to their means. The same is true for females and males (Components 3 and 4) in Cluster 2. Table 7 gives the means and average gates in terms of both components and clusters, as well as the common $\widehat{\boldsymbol{\Sigma}}$ matrix.

Table 7: Estimated parameters of the $G = 2$ Gaussian clusters in the optimal 'full' MoEClust model fit to the AIS data (with an extra noise component and gating concomitants entering the non-noise clusters), with further splitting due to the binary covariate sex in the expert network, giving average gates and component means (for females and males) and the common EEE covariance matrix. While every observation has its own mean parameter, given by the fitted values of the expert network in Table 6, the means are summarised by the posterior mean of the model's fitted values, given by (11).

	Cluster 1			Cluster 2			$\widehat{\boldsymbol{\Sigma}}$ (EEE)				
	All	Female	Male	All	Female	Male					
$\overline{\widehat{\tau}_g(\mathbf{x}_i)}$	0.60	0.21	0.39	0.33	0.25	0.08	RCC	WCC	Hc	Hg	Fe
RCC	4.81	4.51	4.98	4.51	4.33	5.12	0.08	0.08	0.46	0.15	-0.83
WCC	7.02	6.95	7.06	7.10	6.96	7.57		2.50	0.60	0.21	5.12
Hc	44.06	41.79	45.35	41.14	39.61	46.29			3.84	1.33	-7.55
Hg	14.88	13.94	15.51	13.91	13.32	15.90				0.57	-1.05
Fe	70.18	53.05	79.87	87.84	58.96	184.67					821.68

Though the plots in Figure 4 are suitable for univariate data with a single continuous expert network covariate, visualising MoEClust results for multivariate data with $r > 1$ mixed-type covariates constitutes a much greater challenge. For the optimal full MoE model fit to the AIS data, the data and clustering results are shown using a generalised pairs plot in Figure 5. This plot depicts the pairwise relationships between the hematological response variables, the included gating and expert network covariates, and the MAP classification, coloured according to the MAP classification. The marginal distributions of each variable are given along the diagonal. For the hematological response variables, ellipses with axes related to the within-cluster covariances are drawn. For the purposes of visualising Figure 5, owing to the presence of an expert network covariate in the fitted model, the multivariate Gaussian ellipses in panels depicting two response variables are centred on the posterior mean of the fitted values, as described in (11). Their shape and size are also modified for the same reason: they are derived by adding the extra variability in the component means to $\widehat{\Sigma}_g$. Thus, the depicted ellipses do not conform to the EEE covariance constraints of the optimal model.

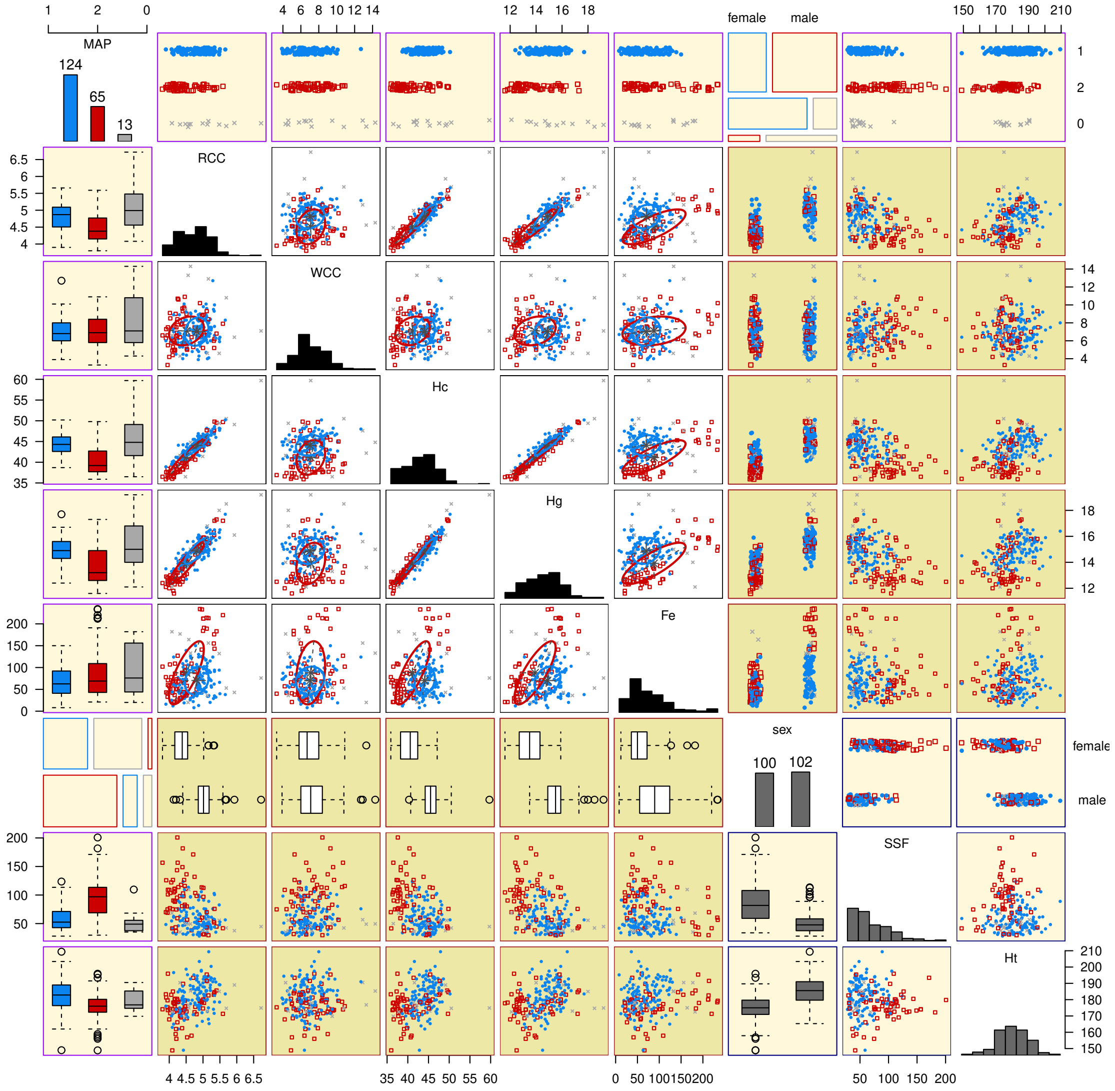


Figure 5: Generalised pairs plot for the optimal 'full' MoEClust model fit to the AIS data, depicting pairwise relationships between the hematological response variables, the expert network covariate sex, the gating concomitants SSF and Ht, and the MAP classification. Colours and plotting symbols correspond to the MAP classification: blue circles and red squares for the two Gaussian clusters; grey crosses for the 4 female and 9 male outlying observations assigned to the uniform noise component. Mosaic plots are used to depict two categorical variables, scatter plots are used for panels involving two continuous variables, and a mix of box-plots and jittered strip-plots are used for mixed pairs.

It is clear from Figure 5 that the variables 'Hematocrit' (Hc), 'Hemoglobin' (Hg), and 'plasma ferritin concentration' (Fe), and the gating network concomitants 'SSF' and 'Ht', are driving much of the separation between the clusters. On the other hand, the expert network covariate 'sex' is driving separation within the Gaussian clusters. The correspondence between the MAP classification and the sex label is notably poor, with an adjusted Rand index (Hubert & Arabie, 1985) of just 0.11. This index is higher for models where sex does not enter the expert network, especially when it instead enters the gating network, though such fitted models all have sub-optimal BIC values (see Table 5). This is because, under the optimal model, the athletes' size in terms of their SSF and height measurements, rather than their sex, influences the probability of cluster membership, and athletes are divided by sex within each cluster rather than the clusters necessarily capturing their sex.

Indeed, Table 6 implies that males, on average, have elevated levels of all five blood measurements in both Gaussian clusters. However, the magnitude of this effect is more pronounced in Cluster 2, related to athletes with higher average SSF measurements (a proxy for body fat) and lower average height. Interestingly, Figure 5 also shows that females have higher average SSF measurements and lower average height; this may explain why there are more males than females in Cluster 1, and the reverse in Cluster 2, given the signs of the gating network coefficients for SSF (0.04) and Ht (-0.08).

6 Discussion

The development of a suite of MoEClust models has been outlined, clearly demonstrating the utility of mixture of experts models for parsimonious model-based clustering where covariates are available. A novel means of visualising such models has also been proposed. The ability of MoEClust models to jointly model the response variable(s) and related covariates provides deeper and more principled insight into the relations between such data in a mixture-model based analysis, and provides a principled method for both creating and explaining the clustering, with reference to information contained in covariates. Their demonstrated use to cluster observations and appropriately capture heterogeneity in cross-sectional data provides only a glimpse of their potential flexibility and utility in a wide range of settings. Indeed, given that general MoE models have been used, under different names, in several fields, including but not limited to statistics (Grün & Leisch, 2007, 2008), biology (Wang et al., 1996), econometrics (Wang et al., 1998), marketing (Wedel & Kamakura, 2012), and medicine (Thompson et al., 1998), MoEClust models could prove useful in many domains.

Improvement over GPCM models has been introduced by accounting for external information available in the presence of potentially mixed-type covariates. Similarly, improvement over Gaussian mixture of experts models which incorporate fixed covariates has been introduced by allowing GPCM family covariance structures in the component densities. MoEClust models are thus Gaussian parsimonious MoE models where the size, volume, shape, and/or orientation can be equal or unequal across components. MoEClust models have been further extended to accommodate the presence of an additional uniform noise component to capture departures from Gaussianity, in such a way that observations are smoothly classified as belonging to Gaussian clusters or as outliers. In particular, two means of doing so have been proposed for models which include gating concomitants. Due to sensitivity of the final solution obtained by the EM algorithm to the initial values, an iterative reallocation procedure based on the Mahalanobis distance has been proposed to mitigate against convergence to suboptimal local maxima for models with expert network covariates. This initialisation algorithm converges quickly and also speeds up convergence of the EM algorithm itself.

Previous parsimonious Gaussian mixtures of experts (Dang & McNicholas, 2015) accommodated only the cases where all covariates enter the expert network MoE model, or the full MoE model with the restriction that all covariates enter both parts of the model. MoEClust constitutes a unifying framework whereby different subsets of covariates can enter either, neither, or both the gating and/or expert networks of Gaussian parsimonious MoE models. Considering the standard mixture model (with no dependence on covariates), or the expert network MoE model, with the equal mixing proportion constraint expands the model family further.

On a cautionary note, care must be exercised in choosing how and where covariates enter when a MoEClust model is used as a clustering tool, as the interpretation of the analysis fundamentally depends on where covariates enter, which of the six special cases of the MoE framework is invoked, and on which GPCM constraints are employed. To this end, a novel greedy forward stepwise search algorithm has been employed for model/variable selection purposes. This strategy has the added advantage of introducing additional parsimony, by potentially reducing the number of regression parameters in the gating and/or expert networks.

Gating network MoEClust models may be of particular interest to users of GPCMs; while concomitants influence the probability of cluster membership, the correspondence thereafter between component densities and clusters has the same interpretation as in standard GPCMs. When covariates enter the component densities, we warn that observations with very different response values can be clustered together, because they are being modelled using the same GLM; similarly, regression distributions with distinct parameters do not necessarily lead to well-separated clusters.

MoEClust models allow the number of parameters introduced by gating and expert network covariates to be offset by a reduction in the number of covariance parameters. This is particularly advantageous when model selection is conducted using the BIC or ICL, which include a penalty term based on the parameter count. Thus, MoEClust models may tend to favour including covariates more than standard Gaussian MoE models would. This is particularly true for explanatory variables in the expert network, which tend to necessitate more regression parameters (Gp) than concomitant variables in the gating network ($G-1$) per additional continuous covariate or level of categorical covariates included. Thus, in cases where a MoE model might elect to include a concomitant variable in the gating network, a MoEClust model with fewer covariance parameters may elect to include it as an explanatory expert network variable instead. While this does lead to a better fit, it can complicate interpretation.

Possible directions for future work in this area include investigating the utility of nonparametric estimation of the gating network (Young & Hunter, 2010), as well as exploring the use of regularisation penalties in the gating and expert networks to help with variable selection when the number of covariates r is large. Regularisation in another, Bayesian sense, by specifying a prior on the component variances/covariances in the spirit of Fraley & Raftery (2007), and/or component regression parameters, could also prove useful for avoiding spurious solutions due to computational singularity described in Section 3.3. MoEClust models could also be developed in the context of hierarchical mixtures of experts (Jordan & Jacobs, 1994), and/or extended to the supervised or semi-supervised model-based classification settings, where some or all observations are labelled.

Beyond the family of GPCM constraints, MoEClust models could be extended to avail of parsimonious factor-analytic covariance structures for high-dimensional data (McNicholas & Murphy, 2008). These could be incorporated into Gaussian mixture of experts models using residuals in an equivalent fashion to Section 3.2 above. Similarly, MoEClust models could benefit from the heavier tails of the multivariate t-distribution, and the robustness to outliers it affords, by considering the associated tEIGEN family of covariance constraints (Andrews & McNicholas, 2012). However, the inclusion of a uniform noise component has the advantage of drawing a clearer distinction between observations belonging to clusters or designated as outliers.

Acknowledgements

This work was supported by the Science Foundation Ireland funded Insight Centre for Data Analytics in University College Dublin under grant number SFI/12/RC/2289_P2.

7 References

J. L. Andrews & P. D. McNicholas. Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: the tEIGEN family. *Statistics and Computing*, 22(5): 1021–1029, 2012. 6, 25

J. Banfield & A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3): 803–821, 1993. 2, 6, 12

T. Benaglia, D. Chauveau, D. R. Hunter, & D. Young. mixtools: an R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6): 1–29, 2009. 7, 8

C. Biernacki, G. Celeux, & G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7): 719–725, 2000. 15

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. 3

D. Böhning, E. Dietz, R. Schaub, P. Schlattmann, & B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2): 373–388, 1994. 9

G. Celeux & G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5): 781–793, 1995. 2, 5, 6

R. D. Cook & S. Weisberg. *An Introduction to Regression Graphics*. John Wiley & Sons, New York, 1994. 18

U. J. Dang & P. D. McNicholas. Families of parsimonious finite mixtures of regression models. In I. Morlini, T. Minerva, & M. Vichi, editors, *Advances in Statistical Models for Data Analysis: Studies in Classification, Data Analysis, and Knowledge Organization*, pages 73–84. Springer, Switzerland, 2015. 2, 8, 20, 24

U. J Dang, A. Punzo, P. D. McNicholas, S. Ingrassia, & R. P. Browne. Multivariate response and parsimony for Gaussian cluster-weighted models. *Journal of Classification*, 34(1): 4–34, 2017. 2, 8

C. M. Dayton & G. B. Macready. Concomitant-variable latent-class models. *Journal of the American Statistical Association*, 83(401): 173–178, 1988. 2, 3

A. P. Dempster, N. M. Laird, & D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1): 1–38, 1977. 8

J. G. Eisenhauer. Regression through the origin. *Teaching Statistics*, 25(3): 76–80, 2003. 11

C. Fraley & A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2): 155–181, 2007. 24

L. A. García-Escudero, A. Gordaliza, F. Greselin, S. Ingrassia, & A. Mayo-Iscar. Eigenvalues and constraints in mixture modeling: geometric and computational issues. *Advances in Data Analysis and Classification*, 12(2): 203–233, 2018. 12

J. Geweke & M. Keane. Smoothly mixing regressions. *Journal of Econometrics*, 138 (1): 252–290, 2007. 4

I. C. Gormley & T. B. Murphy. Clustering ranked preference data using sociodemographic covariates. In S. Hess & A. Daly, editors, *Choice Modelling: The State-of-the-art and The State-of-practice – Proceedings from the Inaugural International Choice Modelling Conference*, chapter 25, pages 543–569. Emerald, United Kingdom, 2010. 16

I. C. Gormley & T. B. Murphy. Mixture of experts modelling with social science applications. In K. Mengersen, C. Robert, & D. M. Titterington, editors, *Mixtures: Estimation and Applications*, chapter 9, pages 101–121. John Wiley & Sons, New York, 2011. 4

B. Grün & F. Leisch. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51(11): 5247–5252, 2007. 7, 23

B. Grün & F. Leisch. FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4): 1–35, 2008. 7, 23

C. Hennig. Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17(2): 273–296, 2000. 2

C. Hennig & P. Coretto. The noise component in model-based cluster analysis. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker, editors, *Data Analysis, Machine Learning and Applications: Studies in Classification, Data Analysis, and Knowledge Organization*, pages 127–138. Springer, Berlin, Heidelberg, 2008. 12

L. Hubert & P. Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985. 23

M. Hurn, A. Justel, & C. P. Robert. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1): 55–79, 2003. 16

S. Ingrassia, S. C. Minotti, & G. Vittadini. Local statistical modeling via the cluster-weighted approach with elliptical distributions. *Journal of Classification*, 29(3): 363–401, 2012. 1, 5

S. Ingrassia, A. Punzo, G. Vittadini, & S. C. Minotti. The generalized linear mixed cluster-weighted model. *Journal of Classification*, 32(1): 85–113, 2015. 2

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, & G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1): 79–87, 1991. 2, 3, 4

M. I. Jordan & R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2): 181–214, 1994. 25

A. E. Lamont, J. K. Vermunt, & M. L. Van Horn. Regression mixture models: does modeling the covariance between independent variables and latent classes improve the results? *Multivariate Behavioural Research*, 51(1): 35–52, 2016. 1

R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, & G. Govaert. Rmixmod: the R package of the model-based unsupervised, supervised, and semi-supervised classification mixmod library. *Journal of Statistical Software*, 67(6): 1–29, 2015. 7, 8

P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings National Institute of Science, India*, 2(1): 49–55, 1936. 13

A. Mazza, A. Punzo, & S. Ingrassia. flexCWM: a flexible framework for cluster-weighted models. *Journal of Statistical Software*, 86: 1–27, 2018. 8

P. McCullagh & J. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983. 4

P. D. McNicholas & T. B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3): 285–296, 2008. 6, 25

D. McParland & I. C. Gormley. Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2): 155–169, 2016. 13

E. H. Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26(9): 394–395, 1920. 13

K. Murphy & T. B. Murphy. `MoEClust`*: Gaussian parsimonious clustering models with covariates and a noise component*, 2020. URL https://cran.r-project.org/package=MoEClust. R package version 1.3.1. 2, 16, 29

H. Ning, Y. Hu, & T. S. Huang. Efficient initialization of mixtures of experts for human pose estimation. In *Proceedings of the International Conference on Image Processing, ICIP 2008, October 12-15, 2008, San Diego, California, USA*, pages 2164–2167, 2008. 13

A. Punzo & S. Ingrassia. Parsimonious generalized linear Gaussian cluster-weighted models. In I. Morlini, T. Minerva, & M. Vichi, editors, *Advances in Statistical Models for Data Analysis: Studies in Classification, Data Analysis, and Knowledge Organization*, pages 201–209. Springer, Switzerland, 2015. 8

A. Punzo & S. Ingrassia. Clustering bivariate mixed-type data via the cluster-weighted model. *Computational Statistics*, 31(3): 989–103, 2016. 2

A. Punzo & P. D. McNicholas. Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, 58(6): 1506–1537, 2016. 6

R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. 2, 7

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464, 1978. 15

L. Scrucca, M. Fop, T. B. Murphy, & A. E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1): 289–317, 2016. 2, 6, 7, 8, 11

T. J. Thompson, P. J. Smith, & J. P. Boyle. Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3): 393–404, 1998. 23

P. Wang, M. L. Puterman, I. Cockburn, & N. Le. Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 52(2): 381–400, 1996. 23

P. Wang, M. L. Puterman, & I. Cockburn. Analysis of patent data – a mixed-Poisson regression-model approach. *Journal of Business & Economic Statistics*, 16(1): 27–41, 1998. 23

M. Wedel & W. A. Kamakura. *Market Segmentation: Conceptual and Methodological Foundations*. International Series in Quantitative Marketing. Springer, US, 2012. 23

D. S. Young & D. R. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10): 2253–2266, 2010. 24

A. Zellner. An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association*, 57 (298): 348–368, 1962. 10

Appendices

Appendix A CO$_2$ Data: Code Examples

Code to reproduce both the exhaustive (Listing 1) and greedy forward stepwise (Listing 2) searches for the CO$_2$ data described in Section 5.1, using the `MoEClust` R package (Murphy & Murphy, 2020), is provided below. The code in Listing 1 can be used to reproduce the results in Table 2.

Listing 1: Exhaustive search R code for the CO$_2$ data.

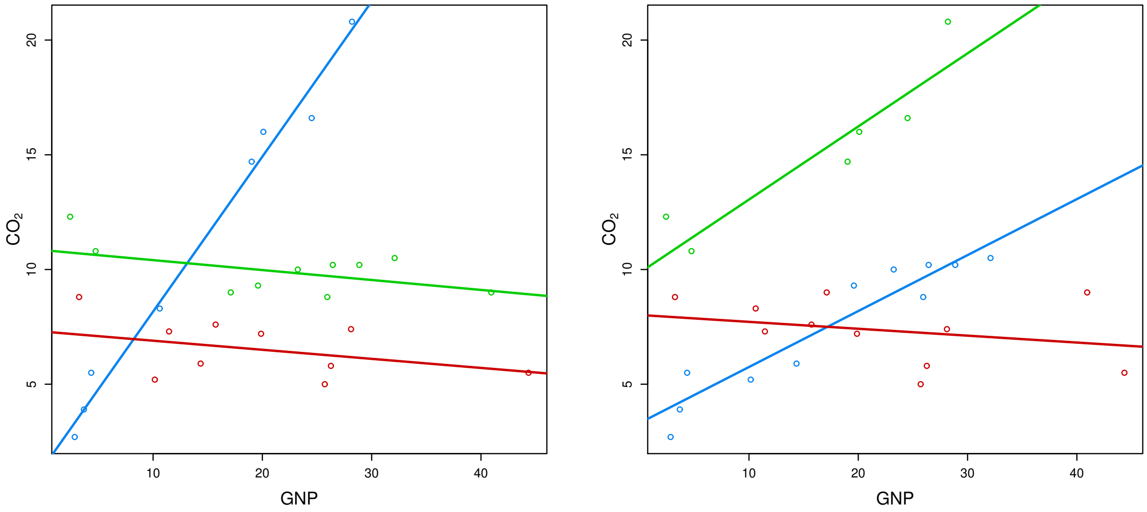
```r
 library(MoEClust)
 data(CO2data)
 CO2   <- CO2data$CO2
 GNP   <- CO2data$GNP

# Fit models under the 6 special cases of the MoE framework
 m1    <- MoE_clust(CO2, G=1:9)
 m2    <- MoE_clust(CO2, G=2:9, gating=~GNP)
 m3    <- MoE_clust(CO2, G=1:9, expert=~GNP)
 m4    <- MoE_clust(CO2, G=2:9, gating=~GNP, expert=~GNP)
 m5    <- MoE_clust(CO2, G=2:9, equalPro=TRUE)
 m6    <- MoE_clust(CO2, G=2:9, expert=~GNP, equalPro=TRUE)

# Collate results and rank (by BIC) only the 6 optimal models
 res   <- list(m1=m1, m2=m2, m3=m3, m4=m4, m5=m5, m6=m6)
 (comp <- MoE_compare(res, optimal.only=TRUE))
```

Listing 2: Stepwise search R code for the CO$_2$ data.

```r
 library(MoEClust)
 data(CO2data)
 CO2   <- CO2data$CO2
 GNP   <- CO2data$GNP

# Conduct a stepwise search
 (mod1 <- MoE_stepwise(CO2, GNP))

# Conduct a stepwise search for models with a noise component
 (mod2 <- MoE_stepwise(CO2, GNP, noise=TRUE))

# Compare both sets of results to choose the optimal model
 (best <- MoE_compare(mod1, mod2, optimal.only=TRUE)$optimal)
```

Appendix B CO$_2$ Data: EM Initialisation

The regression lines for the optimal $G = 3$ equal mixing proportion expert network MoEClust model with equal component variances and the explanatory variable GNP fitted to the CO$_2$ data, with and without the initial partition obtained by model-based agglomerative hierarchical clustering being passed through Algorithm 1, are shown in Figure B.1. A BIC value of -155.20 is achieved after 21 EM iterations (starting, after 6 iterations of our proposed initialisation strategy, from a log-likelihood of -66.01) compared to a value of -161.06 after 28 EM iterations without Algorithm 1 (starting from a log-likelihood of -76.39). While the models differ only in terms of the initialisation strategy employed, Table 2 shows that the model would not have been identified as optimal according to the BIC criterion had Algorithm 1 not been used. The superior solution in Figure B.1a has one cluster with a steep slope and two clusters with near-zero slopes but different intercepts. Notably, supplying 100 random starts to Algorithm 1 did not yield an improved BIC in any instance. Similarly, the optimal BIC value obtained by supplying 100 random starts to the EM algorithm directly was greater than -161.06 but still less than -155.20.



(a) With Algorithm 1 invoked for initialisation, achieving a BIC value of -155.20.

(b) Without Algorithm 1 invoked for initialisation, achieving a BIC value of -161.06.

Figure B.1: Scatter plots of GNP against CO$_2$ emissions for $n = 28$ countries showing $G = 3$ coloured linear regression components from MoEClust models with equal variances and mixing proportions, with (a) and without (b) the initialisation strategy described in Algorithm 1 invoked.

Appendix C AIS Data: Stepwise Model Search

For the AIS data, Table C.1 gives the results of the greedy forward stepwise model selection strategy described in Algorithm 2, showing the action yielding the best improvement in terms of BIC for each step. This forward search is less computationally onerous than its equivalent in the backwards direction. A 2-component EVE *equal mixing proportion expert network MoE* model is chosen, in which the mixing proportions are constrained to be equal and sex enters the expert network. This same model was identified after an exhaustive search over a range of G values, the full range of GPCM covariance constraints, and every possible combination of the BMI and sex covariates in the gating and expert networks (see Table 5). Note, however, that the remaining covariates in Table 4 are also considered for inclusion here.

To give consideration to outlying observations departing from the prevailing pattern of Gaussianity, a separate stepwise search is conducted, starting from a $G = 0$ noise-only model, with all candidate models having an additional noise component. Thus, a distinction is made between the model found by following the steps shown in Table C.1 with $G = 2$ EVE Gaussian components, and the model found by the second stepwise search described in Table C.2 with three, of which two are EEE Gaussian and one is uniform. Ultimately, the model with the noise component identified in Table C.2 is chosen, based on its superior BIC. Aside from the noise component, it similarly includes 'sex' in the expert network, but differs in its treatment of the gating network and the GPCM constraints employed for the Gaussian clusters. It is a *full MoE* model where the Gaussian clusters have equal volume, shape, and orientation, the expert network includes the covariate 'sex', and the both 'SSF' and 'Ht' influence the probability of belonging to the Gaussian clusters but not the additional noise component, as per (9).

Table C.1: Results of the forward stepwise model selection algorithm applied to the AIS data where candidate models do not include a noise component. All covariates in Table 4 are considered for inclusion in both parts of the model. The optimal action and associated BIC value is detailed for each step. The resulting models are described in terms of the number of Gaussian components G, the GPCM constraints used, and the treatment of the gating and expert networks.

Step	Optimal Action	G	GPCM	Gating	Expert	BIC
1	—	1	EEE	—		−4202.79
2	Add explanatory variable (Expert)	1	EEE	—	sex	−4050.64
3	Add component and constrain mixing proportions	2	EVE	Equal	sex	−4010.14
4	Stop	2	EVE	Equal	sex	−4010.14

Table C.2: Results of the forward stepwise model selection algorithm applied to the AIS data where all candidate models explicitly include a noise component. All covariates in Table 4 are considered for inclusion in both parts of the model. The optimal action and associated BIC value is detailed for each step. The resulting models are described in terms of the number of Gaussian (i.e. non-noise) components G, the GPCM constraints used, and the treatment of the gating and expert networks. When gating concomitants are included, the chosen models here correspond to the NGN model in (9). Thus, the noise component's mixing weight is constant and independent of the included concomitants.

Step	Optimal Action	G	GPCM	Gating	Expert	BIC
1	—	0	—	—	—	−4869.82
2	Add component	1	EEE			−4149.46
3	Add explanatory variable (Expert)	1	EEE		sex	−4013.55
4	Add component	2	EVE		sex	−3992.81
5	Add concomitant (Gating)	2	EVE	NGN: SSF	sex	−3990.09
6	Add concomitant (Gating)	2	EEE	NGN: SSF, Ht	sex	−3989.83
7	Stop	2	EEE	NGN: SSF, Ht	sex	−3989.83