Bioinformatics, 36(2), 2020, 621–628
doi: 10.1093/bioinformatics/btz599
Advance Access Publication Date: 1 August 2019
Original Paper

OXFORD

Data and text mining

Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping

Siva Rajesh Kasa[1], Sakyajit Bhattacharya[2] and Vaibhav Rajan [1,*]

[1]Department of Information Systems and Analytics, School of Computing, National University of Singapore, 117418 Singapore and [2]TCS Innovation Labs, Kolkata 700156, India

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on December 11, 2018; revised on May 27, 2019; editorial decision on July 23, 2019; accepted on July 26, 2019

Abstract

Motivation: The identification of sub-populations of patients with similar characteristics, called patient subtyping, is important for realizing the goals of precision medicine. Accurate subtyping is crucial for tailoring therapeutic strategies that can potentially lead to reduced mortality and morbidity. Model-based clustering, such as Gaussian mixture models, provides a principled and interpretable methodology that is widely used to identify subtypes. However, they impose identical marginal distributions on each variable; such assumptions restrict their modeling flexibility and deteriorates clustering performance.

Results: In this paper, we use the statistical framework of copulas to decouple the modeling of marginals from the dependencies between them. Current copula-based methods cannot scale to high dimensions due to challenges in parameter inference. We develop HD-GMCM, that addresses these challenges and, to our knowledge, is the first copula-based clustering method that can fit high-dimensional data. Our experiments on real high-dimensional gene-expression and clinical datasets show that HD-GMCM outperforms state-of-the-art model-based clustering methods, by virtue of modeling non-Gaussian data and being robust to outliers through the use of Gaussian mixture copulas. We present a case study on lung cancer data from TCGA. Clusters obtained from HD-GMCM can be interpreted based on the dependencies they model, that offers a new way of characterizing subtypes. Empirically, such modeling not only uncovers latent structure that leads to better clustering but also meaningful clinical subtypes in terms of survival rates of patients.

Availability and implementation: An implementation of HD-GMCM in R is available at: https://bitbucket.org/cdal/hdgmcm/.

Contact: vaibhav.rajan@nus.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In several diseases, including cancer, patients exhibit a remarkable degree of variation, in genetic landscape, lifestyle and environmental factors, and this heterogeneity presents a formidable challenge to healthcare (Saria and Goldenberg, 2015). Precision medicine attempts to address this challenge by tailoring diagnostic, prognostic and therapeutic strategies to patient sub-populations with similar characteristics; the identification of such groups is called patient subtyping or stratification (Mirnezami *et al.*, 2012). Development of computational approaches that can leverage high-throughput molecular data for subtyping is an active research area, with a broader goal of gaining deeper understanding of associations between biological entities such as genes, proteins and drugs, to expand our scientific knowledge (Lin *et al.*, 2019).

Clustering has been successfully used for subtyping, e.g. in lung cancer subtyping using gene-expression data (Chen *et al.*, 2013), and pan-cancer integrative analysis using multiple high-throughput measurements (Hoadley *et al.*, 2014), that have led to improved

outcome prediction and new taxonomies with valuable prognostic information. High-dimensionality, where the number of samples (n) is lower than the number of dimensions (p) is a common characteristic of omics datasets, and is known to deteriorate the performance of classical clustering algorithms (McWilliams and Montana, 2014). A common approach is to employ dimensionality reduction (e.g. through PCA) before clustering: such methods often lack interpretability, are not sufficiently robust or yield suboptimal results (Bouveyron and Brunet-Saumard, 2014).

Mixture models are a principled statistical approach to clustering, where inferred clusters can be interpreted through the lens of the underlying distributional assumptions. On gene-expression datasets, mixture models were found to outperform widely used classical methods like *K*-means and hierarchical clustering (Thalamuthu *et al.*, 2006). Although mixture models are over-parameterized in high dimensions, which make parameter inference difficult, variable selection techniques and parsimonious covariance structures alleviate the problems to a large extent and enable their use in subtyping (Baek and McLachlan, 2011; Städler *et al.*, 2017; Xie *et al.*, 2010). However, through the choice of the multivariate distribution, model-based clustering imposes distributional assumptions on the marginals, along each dimension, and these marginal distributions are assumed or forced to be identical (e.g. a multivariate normal imposes univariate normal distribution on each marginal); such assumptions restrict their modeling flexibility.

The statistical framework of *copulas* provides a modular parameterization of multivariate distributions that decouples the modeling of marginals from the dependencies between them. Various parametric families model both the strength and shape of dependencies (see Fig. 1). This allows each marginal to be chosen independently from any distribution and the dependency model offers a richer characterization than single-number metrics like Pearson's or Spearman's correlation coefficients. Thus, when interest lies mainly in discovering feature dependencies, copulas provide an elegant model of dependencies with no restrictive assumptions on the marginals. Such models have been used extensively in finance (Patton, 2009) and more recently in *dependency clustering* that discovers clusters based on their dependency patterns (Rajan and Bhattacharya, 2016; Rey and Roth, 2012; Tekumalla *et al.*, 2017). For many applications, including clustering, a semi-parametric model works well, where copulas are used to model dependency patterns, assuming no fixed parametric model for the marginals. However, copulas are rarely used in high-dimensional settings—either parameter estimation is intractable or they lose their modeling flexibility (Joe, 2014).

In this paper, we present a copula-based method, called HD-GMCM, for dependency clustering of high-dimensional data. We use a specific copula model, the Gaussian mixture copula model (GMCM) that can model a wide variety of dependencies including asymmetric, tail and multimodal dependencies (Bhattacharya and Rajan, 2014; Bilgrau *et al.*, 2016). HD-GMCM uses Alternating Expectation Conditional Maximization (AECM) for parameter estimation. To overcome the limitations of previous parameter estimation methods for GMCM and enable HD-GMCM to scale to high dimensions, we use constrained covariance structures (McNicholas and Murphy, 2008), to reduce the local dimensionality of each cluster. This reduces the number of parameters to be estimated at high dimensions but induces a model selection problem that we address through the use of a penalized likelihood approach with the LASSO penalty (Khalili and Chen, 2007). To our knowledge, HD-GMCM is the first copula-based clustering model that can fit high-dimensional data, where $p > n$.

Our experiments on several real high-dimensional gene-expression and clinical datasets show that HD-GMCM outperforms state-of-the-art model-based clustering methods. With a marginal-free copula-based approach, HD-GMCM is better at modeling non-Gaussian data and is found to be robust to outliers in our experiments. We present a case study on lung cancer, where we illustrate the benefits of HD-GMCM for both dependency analysis and clustering. Clusters obtained from HD-GMCM can be interpreted based on the dependencies they model. Such modeling not only uncovers latent structure that leads to better clustering but also meaningful subtypes in terms of survival rates of patients. We believe that further analysis of such dependency patterns may also lead to more fine-grained characterization of associations between biological entities to gain deeper insights on their interactions.

2 Related work

2.1 Model-based clustering of high-dimensional data

For a review on model-based clustering of high-dimensional data and a discussion on information loss due to dimensionality reduction before clustering, see Bouveyron and Brunet-Saumard (2014). Two categories of approaches have been developed for model-based clustering of high-dimensional data:

(1) *Subspace clustering* methods cluster data and simultaneously attempt to reduce locally each cluster's dimensionality. Mixture of

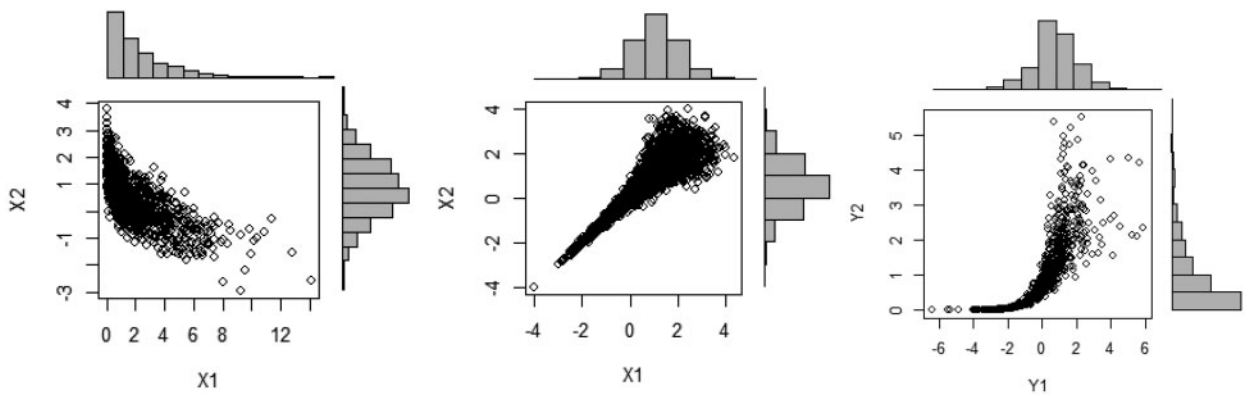

Fig. 1. Copulas enable independent parameterization of marginals and dependencies: (left) Gaussian copula models symmetric dependencies, with Gamma and Gaussian marginals; (center) Clayton copula models lower tail dependencies (where lower values have higher dependence than higher values in the two variables), with Gaussian marginals; (right) Clayton copula, with Student's t and exponential marginals. Various copula families define the shape of the dependencies while their parameters determine the strength of association

factor analyzers (MFA) (Ghahramani and Hinton, 1997; McLachlan *et al.*, 2003) is one of the earliest such approaches. Since the number of covariance parameters grows quadratically with the data dimensionality, constrained covariance structures were introduced in MFA through a family of parsimonious Gaussian mixture models (PGMMs; McNicholas and Murphy, 2008; McNicholas *et al.*, 2010). Bouveyron *et al.* (2007) developed HDDC that uses a combination of subspace clustering and parsimonious modeling for GMMs.

(2) *Variable selection* methods select relevant variables for clustering that are assumed to primarily determine the cluster structure in the data. A recent review on variable selection methods in model-based clustering can be found in Fop *et al.* (2018). A broad class of techniques uses penalized clustering criteria (Pan and Shen, 2007). Marbac and Sedki (2017) proposed a clustering method, VarSelLCM, with an efficient inference algorithm through the use of a new information criterion. Using this criterion simplifies model selection and works particularly well for $p > n$ cases, for moderately large n. In a recent work on subtyping, Gaussian graphical models were used for high-dimensional clustering in MixGlasso (Städler *et al.*, 2017). They also use a penalized likelihood that adapts to the number of clusters, sample size and scale of the clusters.

2.2 Copulas and mixture models

Due to their flexible characterization of multivariate distributions, mixtures of copulas have been used in various contexts (Fujimaki *et al.*, 2011; Kosmidis and Karlis, 2016; Rey and Roth, 2012). None of these address the problems of clustering high-dimensional data. Vine copulas, that are hierarchical collections of bivariate copulas, can scale to moderately high dimensions but at the cost of exponentially increasing complexity for model selection and estimation (Müller and Czado, 2018). Elidan (2013) provides a comparison of copulas with machine learning models including a discussion on fitting copulas to high-dimensional data.

The GMCM was proposed by Tewari *et al.* (2011). Unlike mixtures of copulas, GMCM is a copula family where the (latent) copula density follows a Gaussian mixture model (GMM; the following section has details). This has considerable advantages for copula-based clustering since clusters can be inferred directly from the dependencies obviating the need for marginal parameter estimation. This was leveraged for clustering by Bhattacharya and Rajan (2014) who designed an expectation maximization (EM) algorithm for GMCM parameter estimation. An improved mixed EM and Gibbs sampling-based approach, for clustering was designed by Rajan and Bhattacharya (2016) to fit both real and ordinal data. Bilgrau *et al.* (2016) discuss computational and statistical hurdles in GMCM parameter estimation and offer some resolutions, but none of these methods work well for clustering high-dimensional data. In a related work, Li *et al.* (2011) studied a specific case of GMCM to design a meta-analysis method called reproducibility analysis, that can be used to verify the reliability and consistency of multiple high-throughput experiments.

3 Background

Consider n i.i.d. instances of p-dimensional data, $\mathbf{X} = [x_{ij}]_{n\times p} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$, where i denotes the observation and j denotes the dimension. Single subscript denotes the index along dimension, unless specified otherwise. For example $\mathbf{X}_j$ denotes the jth dimension of data.

3.1 Gaussian mixture model (GMM)

Let ϕ denote the multivariate normal distribution. The probability density function (PDF) of a G-component GMM is given by $\mathcal{G}(\mathbf{x}; \vartheta) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, with mixing proportions $\pi_g > 0$ such that $\sum_{g=1}^{G} \pi_g = 1$, component-specific mean vectors, $\boldsymbol{\mu}_g$, and covariance matrices, $\boldsymbol{\Sigma}_g$. We use $\vartheta = (\pi_1, \ldots \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G)$ to denote all the parameters.

3.2 Copulas

Let F_j denote the marginal cumulative distribution function (CDF) of $\mathbf{X}_j$ along the jth dimension. A CDF transformation, $\mathbf{U}_j = F_j(\mathbf{X}_j)$, maps a random variable to a scalar that is uniformly distributed in $[0, 1]$. However, the joint distribution of all p marginal CDFs is not uniform and is modeled by a copula, which is a multivariate distribution function $C : [0, 1]^p \to [0, 1]$, defined on random variables $\mathbf{U}_j$. Copulas uniquely characterize continuous joint distributions (Sklar, 1959): for every joint distribution with continuous marginals, $F(\mathbf{X}_1, \ldots, \mathbf{X}_p)$, there exists a unique copula function such that $F(\mathbf{X}_1, \ldots, \mathbf{X}_p) = C(F_1(\mathbf{X}_1), \ldots, F_p(\mathbf{X}_p))$; and the converse is also true. It can be shown that the corresponding joint density is given by the product of the individual marginal densities f_j and the *copula density* c:

$$f(\mathbf{x}) = c(F_1(\mathbf{x}_1), \ldots, F_p(\mathbf{x}_p))\Pi_{j=1}^{p} f_j(\mathbf{x}_j). \tag{1}$$

Equation (1) shows how copulas enable flexible constructions of multivariate densities by decoupling the specification of marginals (f_j) and dependence structure (c), thus allowing us to choose each parametric family independently from each other (as shown in Fig. 1). Equation (2) [derived from (1)], illustrates how copula families can be defined by the choice of the joint density f that determines the dependence structure:

$$c(\mathbf{U}_1, \ldots, \mathbf{U}_p) = \frac{f(\mathbf{x})}{\Pi_{j=1}^{p} f_j(\mathbf{x}_j)}. \tag{2}$$

3.3 GMCM

In GMCM, the dependence is obtained from a GMM. Let $\Psi_j(\vartheta)$ and $\psi_j(\vartheta)$ denote the jth marginal CDF and PDF, respectively, of $\mathcal{G}(\vartheta)$. Let Ψ_j^{-1} denote the inverse CDF and $\mathbf{Y}_j = \Psi_j^{-1}(\mathbf{U}_j)$. Using Eq. (2), we obtain the GMCM copula density:

$$c_{\mathcal{G}}(\mathbf{U}; \vartheta) = \frac{\mathcal{G}(\Psi_j^{-1}(\mathbf{U}))}{\prod_{j=1}^{p} \psi_j(\Psi_j^{-1}(\mathbf{U}_j))} \tag{3}$$

GMCM can be used to obtain cluster labels $l \in \{1, \ldots, G\}$ through a semi-parametric MAP estimate $\arg\max_l P(l = g|\vartheta, \mathbf{X})$, using rank-transformed marginals in the data as estimates of $\mathbf{U}_j$ (Bhattacharya and Rajan, 2014). Supplementary Appendix SA provides a more detailed description of copulas and GMCM.

3.4 GMCM parameter inference

Semi-parametric inference of GMCM, for applications that use only the dependence structure, obtains estimates of copula parameters ϑ, without estimating the marginal parameters. Maximizing the copula likelihood $c_{\mathcal{G}}$ is difficult and, in practice, the pseudo-likelihood:

$$\mathcal{L} = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(\mathbf{y}_i|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \tag{4}$$

is used. Genest *et al.* (1995) study the properties of estimates based on the pseudo-likelihood and show that for continuous-valued marginals, the estimator is consistent and asymptotically normal. Such estimates have been used for the Gaussian copula (Hoff *et al.*, 2007). However, even obtaining a maximum likelihood estimate through the pseudo-likelihood, $\arg\max_\vartheta \mathcal{L}(\mathbf{U})$, poses challenges for GMCM that are discussed in detail by Bilgrau *et al.* (2016).

The main challenge is due to the inverse CDF Ψ_j^{-1} that has no closed-form expression. A Pseudo Expectation Maximization (PEM) approach (Bilgrau *et al.*, 2016; Tewari *et al.*, 2011) iteratively alternates between estimating $y_{ij} = \Psi_j^{-1}(\hat{\mathbf{U}}_j; \vartheta)$ and updating ϑ by E and M steps. They use a grid search-based heuristic to compute the inverse CDF. However, this is prohibitively expensive as it scales exponentially with dimensionality. Moreover, since the estimates $\hat{\mathbf{U}}_j$ are not constant across iterations of PEM, there is no guarantee of convergence, resulting in biased estimates. Bhattacharya and Rajan (2014) design an algorithm using an approximation to y_{ij} (Eq. (5)) and prove its convergence. But empirically, its clustering performance is found to deteriorate with increasing data dimensionality, particularly when $p > n$.

Another reason why previous inference methods cannot fit high-dimensional data is due to a matrix inversion step during covariance estimation. When $n \leq p$ the matrix is singular and the inversion fails. To address this problem we impose constraints on the covariance matrix through PGMM (McNicholas and Murphy, 2008) (see Supplementary Appendix SA for a summary of PGMM). The use of PGMM enables a parsimonious GMCM model in a q-dimensional space where $q < p$ and avoids singular matrices. But this poses a new problem of model selection, since now we have to select the covariance structure family as well as the number of latent factors. So, we use a model selection criterion—the LASSO-penalized BIC (LPBIC), which is designed for high-dimensional settings for the PGMM family (Bhattacharya and McNicholas, 2014).

4 HD-GMCM: our inference algorithm

Our new inference algorithm, called HD-GMCM, uses the penalized likelihood approach (Khalili and Chen, 2007), with an LASSO penalty for the mean parameters of $\mathcal{G}$. We maximize:

$$\mathcal{L}_{\text{pen}} = \log \mathcal{L}(\vartheta|\mathbf{x}) - \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} \varphi(\mu_{gj})$$

where $\varphi(\mu_{gj}) = n\lambda_n|\mu_{gj} - c_j|$, where μ_{gj} is the jth element in $\boldsymbol{\mu}_g$, $\mathbf{c}$ is the mean of all the data points and λ_n is a tuning parameter that depends on n.

HD-GMCM uses the approximate expression for y_{ij} in terms of the CDF u_{ij} derived in Bhattacharya and Rajan (2014). For each i and j, y_{ij} can be approximated as follows (Proof in Supplementary Appendix SB):

$$y \approx \left(\sum_{g=1}^{G} \frac{\pi_g}{\sigma_g\sqrt{2\pi}}\right)^{-1} \left[u - 0.5 + \left(\sum_{g=1}^{G} \frac{\pi_g\mu_g}{\sigma_g\sqrt{2\pi}}\right)\right] \tag{5}$$

In addition, we apply another crucial condition on the inverse distribution values: in any iteration, if the computed y_{ij} value is outside the range defined by condition 6, we set the value of y_{ij} to the nearest boundary point of the range, to satisfy the condition. This step does not decrease the likelihood at each iteration of the algorithm, as shown in the proof of Theorem 1.

$$|(y_{ij}^{(t+1)} - y_{ij}^{(t)})| \leq |2\kappa - 2y_{ij}^{(t)}|, \text{ where } \kappa = \min_{j,g}(\hat{\mu}_{jg}) \tag{6}$$

To estimate the parameters, we use the AECM algorithm (McNicholas *et al.*, 2010; Meng and Van Dyk, 1997). There are two stages of the algorithm. At the first stage of the algorithm, when estimating $\hat{\pi}_g$ and $\hat{\boldsymbol{\mu}}_g$, we define $\hat{\mathbf{z}}_i = (\hat{z}_{i1}, \ldots, \hat{z}_{iG})$ showing the component membership of the ith observation: $\hat{z}_{ig} = 1$ if $\mathbf{y}_i$ belongs to the gth component and $\hat{z}_{ig} = 0$ otherwise. The component memberships

Algorithm 1. HD-GMCM

Input: Observed n datapoints [as $n \times p$-dimensional matrix $X = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$] and number of clusters G.

Initialize: Set $\vartheta^{(0)}$ using random start or K-means clustering under the constraints that $\hat{\pi}_g^{(0)} > 0$, $\sum_{g=1}^{G} \hat{\pi}_g^{(0)} = 1$ and $\hat{\boldsymbol{\Sigma}}_g^{(0)}$ is positive definite. Set $u_{ij} = \tilde{\mathbf{F}}_j(x_{ij})$ (percentile ranks).

repeat

Reset y:

$$y_{ij}^{(t+1)} = \left(\sum_{g=1}^{G} \frac{\hat{\pi}_g^{(t)}}{\sqrt{2\pi}\hat{\sigma}_{g,jj}^{(t)}}\right)^{-1} \left[u_{ij} + \frac{1}{\sqrt{2\pi}} \sum_{g=1}^{G} \frac{\hat{\pi}_g^{(t)}\hat{\mu}_{gj}^{(t)}}{\hat{\sigma}_{g,jj}^{(t)}} - \frac{1}{2}\right]$$

ensure that
$|(y_{ij}^{(t+1)} - y_{ij}^{(t)})| \leq |2\kappa - 2y_{ij}^{(t)}|$ where $\kappa = \min_{j,g}(\hat{\mu}_{jg})$.

Stage I

$$\hat{z}_{ig}^{(t+1)} = \frac{\hat{\pi}_g^{(t)}\phi(\mathbf{y}_i^{(t)}|\hat{\boldsymbol{\mu}}_g^{(t)}, \hat{\boldsymbol{\Sigma}}_g^{(t)})}{\sum_{g=1}^{G} \hat{\pi}_g^{(t)}\phi(\mathbf{y}_i^{(t)}|\hat{\boldsymbol{\mu}}_g^{(t)}, \hat{\boldsymbol{\Sigma}}_g^{(t)})}$$

$$\hat{\pi}_g^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(t)}}{n}$$

$$\hat{\boldsymbol{\mu}}_g^{t+1} = \frac{\sum_{i=1}^{n} \hat{z}_{ig}^{(t+1)}\mathbf{y}_i}{\sum_{i=1}^{n} \hat{z}_{ig}^{(t+1)}} - n\lambda_n\hat{\pi}_g\hat{\boldsymbol{\Sigma}}_g^{(t)}\boldsymbol{\beta}_g^{\text{t}}$$

Stage II

$\hat{\boldsymbol{\Sigma}}_g^{(t+1)}$ is estimated based on the PGMM family used.

until convergence criterion is met:

$\hat{\mathbf{z}}_i$ are treated as the missing data at the first stage. So, the expected complete data log-likelihood is

$$Q(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig} \log \hat{\pi}_g + \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{z}_{ig} \log\{\phi(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)\} - \varphi(\hat{\boldsymbol{\mu}}),$$

where $\hat{z}_{ig} = \hat{\pi}_g\phi(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g) / \sum_{j=1}^{G} \hat{\pi}_j\phi(\mathbf{y}_i|\hat{\boldsymbol{\mu}}_g, \hat{\boldsymbol{\Sigma}}_g)$. The M-step maximizes Q to update the parameter estimates π_g and $\boldsymbol{\mu}_g$. The estimation of π_g is complicated and as seen in Khalili and Chen (2007), we also empirically observe good results with $\hat{\pi}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{n}$. The mean parameter estimate is as follows (derivation in Supplementary Appendix SC), where $\boldsymbol{\beta}_g$ is a vector with p elements, its j^{th} element being $\text{sign}(\hat{\mu}_{gj}^m - c_j)$:

$$\hat{\boldsymbol{\mu}}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}\mathbf{y}_i}{\sum_{i=1}^{n} \hat{z}_{ig}} - n\lambda_n\hat{\pi}_g\hat{\boldsymbol{\Sigma}}_g\boldsymbol{\beta}_g \tag{7}$$

At the second stage of the AECM algorithm, we take the missing data as the group labels $\mathbf{z}_i$ and the unobserved latent factors q to estimate the covariance matrix using a PGMM structure. The component covariance matrices $\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_G$ are updated, depending on the family of PGMM model used, using analytic expressions found in McNicholas and Murphy (2008). These two stages are iterated until convergence, i.e. until the change in pseudo-likelihood is less than a pre-defined threshold γ ($|\mathcal{L}^{(t+1)} - \mathcal{L}^{(t)}| < \gamma$) or until a monotonic increase in copula likelihood ($c_{\mathcal{G}}$) is observed. The tuning parameter λ_n can be chosen by cross-validation. Some choices suggested

by previous authors include $\lambda_n = (\log n)^{1/2}/np$ and $\lambda_n = 1/p$ (Bhattacharya and McNicholas, 2014). The complete HD-GMCM method is shown in Algorithm 1. Due to the inexact update used for $\hat{\pi}_g$, stages I and II of AECM do not guarantee monotonic increase of likelihood. However, empirically we observe monotonic increase in all our experiments as discussed in Supplementary Appendix SH. In Supplementary Appendix SD, we prove that the introduction of our approximation in the 'Reset y' step preserves the property of monotonic increase in likelihood in each iteration.

Theorem 1. *In Algorithm HD-GMCM, the estimates of* $\mathbf{y}$ *from (5) and (6) in the 'Reset* $\mathbf{y}$*' step preserves the property of monotonic increase in the likelihood* $\mathcal{L}$ *over each iteration.*

4.1 Computational complexity

The time complexity is dominated by the computation of ϕ for calculating $z_{ig}^{(t+1)}$ in stage I of each iteration. The worst-case time complexity for each iteration is given by $O(nqGp^2)$. Empirically, we obtained good results with just a few (<10) iterations.

4.2 Model selection

The Bayesian Information Criterion (BIC) is commonly used for selecting the number of components and was empirically found to be effective for the GMCM model (Bhattacharya and Rajan, 2014). However, in high dimensions, BIC is prone to under-estimating the number of components (Giraud, 2014). To address this problem for mixture models an LPBIC was proposed by Bhattacharya and McNicholas (2014), that can be used to select the number of latent factors (q), PGMM covariance structure family as well as the number of components (G). LPBIC can be applied in our case since we use a penalized likelihood ($\mathcal{L}_{pen}$)-based model.

5 Experiments

5.1 Clustering performance on real datasets

Baseline methods: We compare the performance of HD-GMCM on clustering tasks (here, the number of clusters is assumed to be known) with state-of-the-art model-based clustering algorithms HDDC (Bouveyron and Brunet-Saumard, 2014), VarSelLCM (Marbac and Sedki, 2017) and MixGlasso (Städler *et al.*, 2017). We also compare with the previous best GMCM-based clustering algorithms: GMCM (Bilgrau *et al.*, 2016) and EGMCM (Rajan and Bhattacharya, 2016). We use GMMs, PGMM (McNicholas *et al.*, 2018) and K-means as additional baselines. R implementations (R Core Team, 2018) were used in all cases. For all the EM-based algorithms including HD-GMCM, five different K-means initializations were used, where K-means itself uses a random initialization. The result with the best BIC is reported.

Evaluation metrics. We use Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and Adjusted Mutual Information (AMI) (Vinh *et al.*, 2010). In both metrics higher values indicate better clustering. Both metrics are 0 for independent clusterings, have maximum value 1 for identical clusterings and can be negative.

Datasets. We use eight publicly available gene-expression datasets: Leukamia (Wouters, 2011), Colon Cancer (Boulesteix *et al.*, 2011), Prostrate (Chung *et al.*, 2018), Breast Cancer (Hothorn, 2018), Khan500 and NC160 (James *et al.*, 2017) and Lusc-Methyl and Lusc-rnaseq (Weinstein *et al.*, 2013). We also use a clinical dataset, SCADI that has attributes of children with physical and motor disability (Zarchi *et al.*, 2018). We follow the preprocessing steps described in McWilliams and Montana (2014). Statistics of the datasets are shown in Table 1.

Results. Table 1 shows the ARI and AMI obtained by HD-GMCM and baseline methods on each dataset. In six out of the nine cases, HD-GMCM outperforms GMCM, GMM, PGMM, K-means, VarSelLCM, HDDC and MixGlasso. Out of the remaining three cases, performance of HD-GMCM is comparable to that of the best performing algorithm. Algorithms GMCM and GMM (not shown) fail to cluster any of the datasets. GMCM fails in the covariance estimation step (details in Section 3) while GMM and HDDC, that assume Gaussian data, are affected by outliers (also discussed in Section 5.2).

We evaluate the effect of varying cluster signal in the data. Our experiments (in Supplementary Appendix SF) show that HD-GMCM outperforms other baselines when the clusters are not well separated, while the simpler K-means outperforms HD-GMCM as well as other baselines designed for high-dimensional data (HDDC

Table 1. Performance of HD-GMCM and baselines on real datasets

Dataset	n	p	G	Metric	HD-GMCM	EGMCM	GMCM	HDDC	VarSelLCM	MixGlasso	K-means	PGMM
Leukamia	38	999	3	ARI	**0.59**	0.185	—	0.133	0.291	—	0.226	0.226
				AMI	**0.49**	0.256	—	0.245	0.4	—	0.158	0.158
SCADI	70	206	7	ARI	0.297	0.034	—	0.247	0	—	0.210	**0.334**
				AMI	0.377	0.058	—	0.341	0	—	0.397	**0.434**
Colon cancer	62	2000	2	ARI	**0.156**	−0.005	—	−0.025	0.018	—	−0.026	–
				AMI	**0.083**	−0.006	—	0.031	0	—	0.031	–
Khan500	63	500	4	ARI	0.116	**0.181**	—	0.057	0.085	0.16	0.178	0.159
				AMI	0.201	**0.337**	—	0.167	0.198	0.32	0.163	0.319
Breast cancer	49	500	2	ARI	**0.132**	0.010	—	0.004	−0.011	—	0.000	0
				AMI	**0.098**	0.015	—	0	−0.016	—	0.004	0
NC160	64	500	14	ARI	0.37	0.36	—	0	0.309	—	**0.439**	–
				AMI	**0.434**	0.428	—	0	0.385	—	0.426	–
Prostrate	102	300	2	ARI	**0.13**	0	—	0	0	—	0.058	0.026
				AMI	**0.097**	0	—	0	0	—	0.052	0.012
Lusc-rnaseq	130	206	2	ARI	**0.0157**	−0.010	—	—	−0.008	−0.01	−0.0073	0
				AMI	**0.021**	−0.044	—	—	−0.004	−0.0027	−0.0055	0
Lusc-Methyl	130	234	2	ARI	**0.004**	−0.004	—	—	−0.005	−0.007	−0.007	0
				AMI	**0.004**	−0.001	—	—	0.008	−0.005	−0.005	0

Note: Row-wise best results in bold. ARI, Adjusted Rand Index; AMI, Adjusted Mutual Information; n, number of samples; p, dimensions; G, number of clusters, — indicates that algorithm fails to run.

and VarSelLCM) when the clusters are well separated. Our preliminary experiments (in Supplementary Appendix SG) suggest that LPBIC is effective for model selection particularly when the number of components is low.

Illustration. Figure 2 illustrates the advantage of using a copula-based model, through a scatterplot of two features from the Leukamia dataset and the corresponding latent variables $[\mathbf{Y}_j = \Psi_j^{-1}(\mathbf{U}_j)]$. The feature selection procedure is outlined in Supplementary Appendix SJ. No cluster structure is seen in the data but in the latent copula space, GMCM captures the cluster structure well.

5.2 Simulation study

We compare the performance of HD-GMCM and baselines on synthetic high-dimensional data with varying proportions of Gaussian and non-Gaussian features. Our experimental results (Supplementary Appendix SE) show that HD-GMCM outperforms state-of-the-art algorithms for non-Gaussian data, while being comparable to baselines for Gaussian data. We also observe that outliers are assigned a separate cluster that results in singularity during parameter inference, for HDDC and GMM, and is one of the reasons for their failure (also in real data). Rank-based models, including copulas, are known to be robust to outliers (Huber, 1981), and is one of the strengths of GMCM.

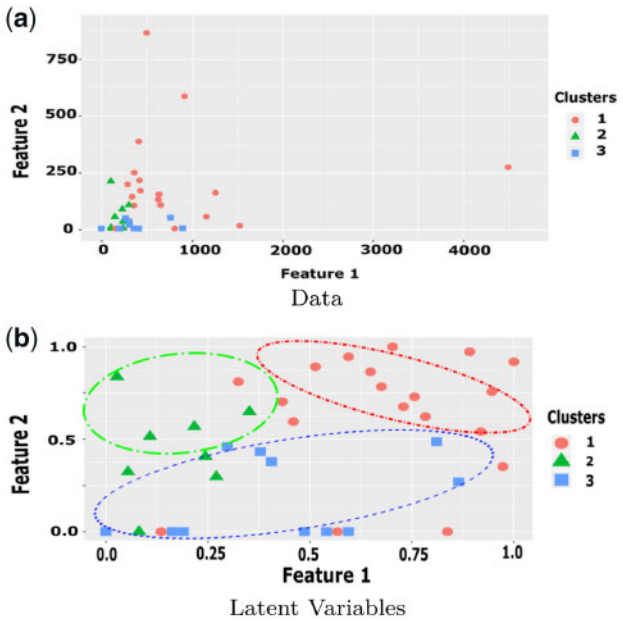

Fig. 2. Scatterplot of features from the Leukamia dataset (**a**) and the corresponding latent variables (Y) in copula space, where clusters are apparent (**b**). Ellipses denote components inferred by GMCM

5.3 Case study: lung cancer

Survival separability. We evaluate the clusters obtained on lung cancer RNA-Seq data (lusc-rnaseq) from TCGA, by HD-GMCM, EGMCM and MixGlasso, with respect to their clinical phenotypes, in terms of survival probability. The survival probabilities are estimated using Kaplan–Meier method and the *P*-values are obtained by the log-rank test on the survival analysis based on the cluster memberships. Figure 3 shows the survival curves obtained on clusters from HD-GMCM, EGMCM and MixGlasso. We observe that only HD-GMCM yields significant clusters ($P-$value < 0.05) that are the least overlapping indicating a good separation of patient subtypes.

Dependency analysis. A copula-based model discovers latent structure based on dependencies. We illustrate its advantages through a visualization of bivariate dependency patterns in the lung cancer data. Note that GMCM models the dependencies of all the dimensions; here, we only show three. Figure 4 shows the differences in bivariate dependency patterns between the two clusters obtained from HD-GMCM. Each scatterplot shows pairwise feature associations and the univariate distribution (in diagonal cells). Note that survival does not have a normal distribution and the distribution varies in the two clusters. The bivariate pattern between smoking and survival is distinctly different between the two clusters: in the cluster with higher survival probability patients tend to smoke less when they are older. GMCM, that can model non-linear and asymmetric dependence, distinguishes the clusters based on such patterns. The strength of these associations can be measured using correlations of fitted bivariate copulas (Table 2). Note that other copula families could also be used. In contrast, a linear regression line that is shown in each of the subplots only measures linear dependence. Such copula-based models are succinct and statistically principled way of characterizing dependencies within subtypes.

6 Conclusion

In this paper, we present a new copula-based algorithm for clustering high-dimensional data. We use the GMCM model, that can model non-linear and asymmetric dependencies, particularly in non-Gaussian data. We overcome the limitations of previous GMCM-based clustering methods through the use of constrained covariance matrices and LASSO-penalized likelihood and design an AECM algorithm that can scale to high dimensions. Our experiments on real gene-expression and clinical datasets show that HD-GMCM outperforms state-of-the-art model-based clustering methods and obtains

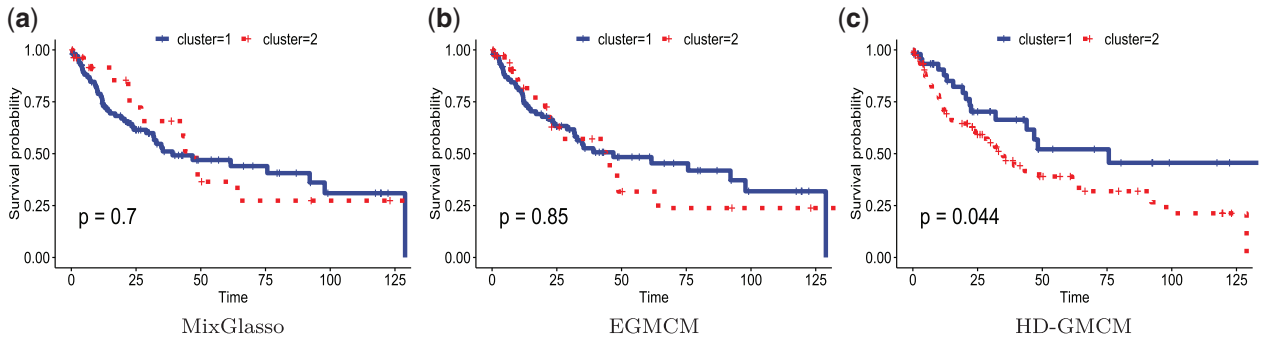

Fig. 3. Kaplan-Meier survival curves with *P*-values using log-rank test: clusters obtained from (**a**) MixGlasso, (**b**) EGMCM and (**c**) HD-GMCM

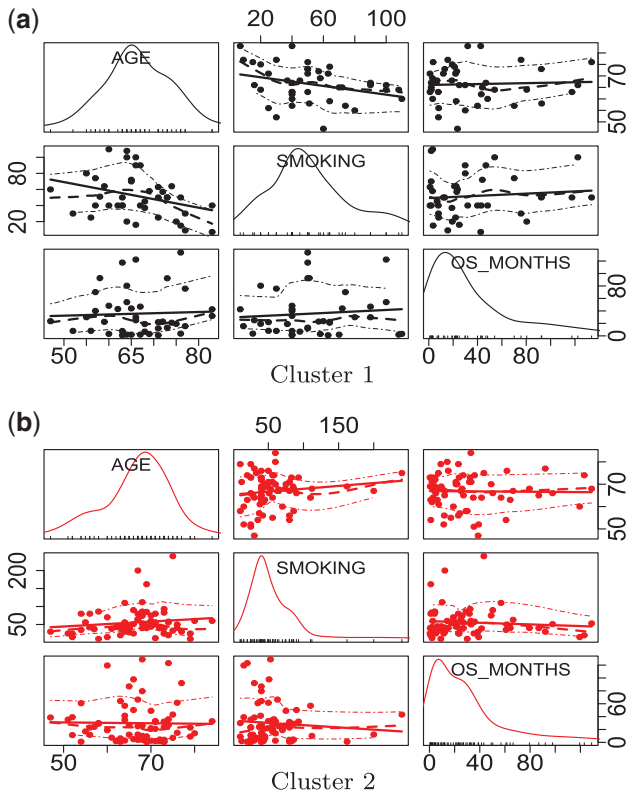

Fig. 4. Bivariate scatterplots of three features—age (in years), survival (in months) and smoking (number of packs per year)—for two clusters obtained from HD-GMCM. (**a**) Cluster 1, (**b**) Cluster 2

Table 2. Correlation coefficients of bivariate Gaussian copulas, fitted pairwise on three features of TCGA lung cancer data, in each cluster

Clusters	Age versus survival	Smok versus survival	Age versus smoking
Cluster 1	0.0387	0.2454	0.1071
Cluster 2	−0.0291	0.1456	0.1518

Note: See text for more details.

meaningful patient subtypes from high-dimensional data. Copulas have been effectively used in modeling dependencies, and our clustering method, for the first time, enables their use for high-dimensional omics data (where $n < p$). The clusters obtained from HD-GMCM are interpretable through the modeled dependency patterns. Such dependency patterns offer a novel and statistically principled way of characterizing subtypes that can potentially lead to deeper insights on interactions between clinical and genetic entities.

Funding

This work was supported by the Singapore Ministry of Education Academic Research Fund [R-253-000-139-114] to V.R.

Conflict of Interest: none declared.

References

Baek,J. and McLachlan,G.J. (2011) Mixtures of common t-factor analyzers for clustering high-dimensional microarray data. *Bioinformatics*, **27**, 1269–1276.

Bhattacharya,S. and McNicholas,P.D. (2014) A LASSO-penalized BIC for mixture model selection. *Adv. Data Anal. Class.*, **8**, 45–61.

Bhattacharya,S. and Rajan,V. (2014) Unsupervised learning using Gaussian mixture copula model. In: 21st International Conference on Computational Statistics. Geneva, Switzerland.

Bilgrau,A.E. *et al.* (2016) GMCM: unsupervised clustering and meta-analysis using Gaussian mixture copula models. *J. Stat. Software*, **70**, 1–23.

Boulesteix,A.-L. *et al.* (2011) *plsgenomics: PLS Analyses for Genomics. R Package Version 1.5-1.*

Bouveyron,C. and Brunet-Saumard,C. (2014) Model-based clustering of high-dimensional data: a review. *Comput. Stat. Data Anal.*, **71**, 52–78.

Bouveyron,C. *et al.* (2007) High-dimensional data clustering. *Comput. Stat. Data Anal.*, **52**, 502–519.

Chen,G. *et al.* (2013) Biclustering with heterogeneous variance. *Proc. Natl. Acad. Sci. USA*, **110**, 12253–12258.

Chung,D. *et al.* (2018) *spls: Sparse Partial Least Squares (SPLS) Regression and Classification. R Package Version 2.2-2.*

Elidan,G. (2013) Copulas in machine learning. In *Copulae in Mathematical and Quantitative Finance*, Springer, pp. 39–60.

Fop,M. *et al.* (2018) Variable selection methods for model-based clustering. *Stat. Surv.*, **12**, 18–65.

Fujimaki,R. *et al.* (2011) Online heterogeneous mixture modeling with marginal and copula selection. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 645–653.

Genest,C. *et al.* (1995) A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, **82**, 543–552.

Ghahramani,Z. and Hinton,G.E. (1997) *The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1.* University of Toronto.

Giraud,C. (2014) *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, Boca Raton, Florida.

Hoadley,K.A. *et al.* (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, **158**, 929–944.

Hoff,P.D. *et al.* (2007) Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.*, **1**, 265–283.

Hothorn,T. (2018) TH.*data: TH Data Archive*. R package. Available at: https://CRAN.R-project.org/package=TH.data.

Huber,P.J. (1981) *Robust Statistics*. Wiley, New York.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Class.*, **2**, 193–218.

James,G. *et al.* (2017) *ISLR: Data for an Introduction to Statistical Learning with Applications in R. R Package Version 1.2.*

Joe,H. (2014) *Dependence Modeling with Copulas*. CRC Press, New York.

Khalili,A. and Chen,J. (2007) Variable selection in finite mixture of regression models. *J. Am. Stat. Assoc.*, **102**, 1025–1038.

Kosmidis,I. and Karlis,D. (2016) Model-based clustering using copulas with applications. *Stat. Comput.*, **26**, 1079–1099.

Li,Q. *et al.* (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

Lin,C.-H. *et al.* (2019) Multimodal network diffusion predicts future disease–gene–chemical associations. *Bioinformatics*, **35**, 1536–1543.

Marbac,M. and Sedki,M. (2017) Variable selection for model-based clustering using the integrated complete-data likelihood. *Stat. Comput.*, **27**, 1049–1063.

McLachlan,G.J. *et al.* (2003) Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Stat. Data Anal.*, **41**, 379–388.

McNicholas,P.D. and Murphy,T.B. (2008) Parsimonious Gaussian mixture models. *Stat. Comput.*, **18**, 285–296.

McNicholas,P.D. *et al.* (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Comput. Stat. Data Anal.*, **54**, 711–723.

McNicholas,P.D. *et al.* (2018) *pgmm: Parsimonious Gaussian Mixture Models. R Package Version 1.2.3.*

McWilliams,B. and Montana,G. (2014) Subspace clustering of high-dimensional data: a predictive approach. *Data Min. Knowl. Disc.*, **28**, 736–772.

Meng,X.-L. and Van Dyk,D. (1997) The EM algorithm—an old folk-song sung to a fast new tune. *J. R. Stat. Soc. B*, **59**, 511–567.

Mirnezami,R. *et al*. (2012) Preparing for precision medicine. *N. Engl. J. Med*., **366**, 489–491.

Müller,D. and Czado,C. (2018) Representing sparse Gaussian DAGs as sparse R-vines allowing for non-Gaussian dependence. *J. Comput. Graph. Stat*., **27**, 334.

Pan,W. and Shen,X. (2007) Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res*., **8**, 1145–1164.

Patton,A.J. (2009) Copula-based models for financial time series. In *Handbook of Financial Time Series*. Springer, pp. 767–785.

Rajan,V. and Bhattacharya,S. (2016) Dependency clustering of mixed data with Gaussian mixture copulas. In The 25th International Joint Conference on Artificial Intelligence.

R Core Team (2018) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rey,M. and Roth,V. (2012) Copula mixture model for dependency-seeking clustering. In Proceedings of the 29th International Conference on Machine Learning.

Saria,S. and Goldenberg,A. (2015) Subtyping: what it is and its role in precision medicine. *IEEE Intell. Syst*., **30**, 70–75.

Sklar,A. (1959) Fonctions de rpartition n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231.

Städler,N. *et al*. (2017) Molecular heterogeneity at the network level: high-dimensional testing, clustering and a TCGA case study. *Bioinformatics*, **33**, 2890–2896.

Tekumalla,L.S. *et al.* (2017) Vine copulas for mixed data: multi-view clustering for mixed data beyond meta-Gaussian dependencies. *Mach. Learn*., **106**, 1331–1357.

Tewari,A. *et al*. (2011) Parametric characterization of multimodal distributions with non-Gaussian modes. In: *The 11th IEEE International Conference on Data Mining Workshops*. IEEE, pp. 286–292.

Thalamuthu,A. *et al*. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

Vinh,N.X. *et al*. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach. Learn. Res*., **11**, 2837–2854.

Weinstein,J.N. *et al*. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet*., **45**, 1113.

Wouters,L. (2011) *MPM: multivariate Projection Methods*. *R Package Version*, 1.0–22.

Xie,B. *et al*. (2010) Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. *Bioinformatics*, **26**, 501–508.

Zarchi,M. *et al*. (2018) SCADI: a standard dataset for self-care problems classification of children with physical and motor disability. *Int. J. Med. Inform*., **114**, 81.