Dependency Clustering of Mixed Data with Gaussian Mixture Copulas

Vaibhav Rajan, Sakyajit Bhattacharya
Xerox Research Centre India
{vaibhav.rajan, sakyajit.bhattacharya}@xerox.com

Abstract

Heterogeneous data with complex feature dependencies is common in real-world applications. Clustering algorithms for mixed – continuous and discrete valued – features often do not adequately model dependencies and are limited to modeling meta–Gaussian distributions. Copulas, that provide a modular parameterization of joint distributions, can model a variety of dependencies but their use with discrete data remains limited due to challenges in parameter inference. In this paper we use Gaussian mixture copulas, to model complex dependencies beyond those captured by meta–Gaussian distributions, for clustering. We design a new, efficient, semiparametric algorithm to approximately estimate the parameters of the copula that can fit continuous, ordinal and binary data. We analyze the conditions for obtaining consistent estimates and empirically demonstrate performance improvements over state-of-the-art methods of correlation clustering on synthetic and benchmark datasets.

1 Introduction

The fundamental task of clustering has been studied extensively in machine learning [Jain, 2010]. Real world data often contains *mixed* – continuous and discrete – features as well as complex dependencies between features. *Dependency clustering*, that detects clusters corresponding to feature dependency patterns, for mixed data remains a challenging open problem [Plant, 2012; Wang *et al.*, 2015].

Clustering methods for mixed data, that are much fewer than those for continuous data, often do not address dependencies satisfactorily. For example, *K-prototypes* [Huang, 1998] quantifies similarity between mixed observations but does not address dependency; *K-Means-Mixed* [Ahmad and Dey, 2007] models dependencies of discrete features only; *mixture* [Hunt and Jorgensen, 2011], assuming normally distributed data, captures dependence of continuous features only. A coupled representation (which can subsequently be used for clustering) for nominal data [Wang *et al.*, 2011] and for continuous data [Wang *et al.*, 2013] has recently been combined for mixed data [Wang *et al.*, 2015]. The method, called *CoupledMC*, transforms mixed to continuous valued data through a series of steps that capture dependencies both within and between continuous and discrete features using non-parametric metrics like correlations and co-occurrence. The high computational complexity (upto cubic in dimension and quadratic in size of data, for CoupledMC) is a disadvantage of many of these distance-based methods.

Correlation clustering methods discover clusters in subspaces based on correlations revealed by low dimensional representations, e.g. through Principal Component Analysis. This approach has been extended to mixed data in INCONCO [Plant and Böhm, 2011], SCENIC [Plant, 2012] and *SpectralCAT* [David and Averbuch, 2012] that find low dimensional embeddings of the data, in different ways, to detect clusters. INCONCO models dependencies by distinct Gaussian distributions for each category of each discrete feature. While SCENIC is not as restrictive in the dependencies, it also assumes a Gaussian distribution to find the embedding space. *SpectralCAT* discretizes continuous features before spectral clustering using an adaptive Gaussian kernel. Normality assumptions limit the modeling capability of these techniques.

Model–based approaches for continuous data can leverage the flexible framework of copulas that provides a modular parameterization of multivariate distributions – arbitrary marginals independent of *dependency models* from copula families which can model a wide variety of linear and non–linear dependencies. For example, with the Gaussian copula itself, using different marginals, many different joint distributions can be constructed, called *meta–Gaussian* distributions, that have been used in several applications (see [Elidan, 2013]) including multi-view clustering [Rey and Roth, 2012]. However, meta–Gaussian dependencies from the Gaussian copula cannot model many kinds of dependencies, notably asymmetric and tail dependencies, and the Gaussian mixture copula (GMCM) was proposed by [Tewari *et al.*, 2011] to allow flexible dependency modeling going beyond meta–Gaussian distributions. These copula–based models can be used with continuous features only.

With discrete data, copula dependencies are not marginal-free but they can still be used effectively [Genest and Neslehova, 2007]. Clustering with a mixture of copulas was proposed by [Kosmidis and Karlis, 2015] but for discrete marginals it is computationally inefficient, the runtime for parameter estimation being exponential in data dimensions.

An efficient alternative not yet explored for dependency

clustering, is the semiparametric *extended rank likelihood (ERL)* approach of [Hoff, 2007] that allows us to use the modeling flexibility of copulas with mixed data. We exploit this technique to enable the efficient use of GMCM on mixed data, to model complex dependencies, that in turn, improves clustering performance. We analyze, both theoretically and empirically, and find that the ERL approximation yields accurate parameter estimates, and good clustering performance, when all features are continuous or ordinal with not many levels.

To summarize, our contributions are:

1. We present a Gaussian mixture copula based clustering algorithm that:
   - can model complex dependencies beyond those captured by meta-Gaussian distributions,
   - can fit mixed – continuous, ordinal and binary valued – data, which previous GMCM-based methods cannot and
   - scales linearly with size and quadratically with dimensions of input, which is significantly faster than state-of-the-art correlation clustering methods for mixed data.
2. We theoretically analyze the conditions necessary to obtain consistent and asymptotically unbiased estimates of GMCM parameters, which, to our knowledge, is the first such analysis of an ERL approach for mixed data clustering.
3. Our experimental results demonstrate the efficacy of our method, that outperforms state-of-the-art methods for correlation clustering on synthetic and real benchmark data sets with mixed features, thus illustrating the advantage of our copula-based approach for dependency clustering.

2 Gaussian Mixture Copulas

A p–dimensional copula is a multivariate distribution function $C : [0,1]^p \mapsto [0,1]$. A theorem by [Sklar, 1959] proves that copulas can uniquely characterize continuous joint distributions: for every joint distribution with continuous marginals, $F(Y_1, \ldots, Y_p)$, there exists a unique copula function such that $F(Y_1, \ldots, Y_p) = C(F_1(Y_1), \ldots, F_p(Y_p))$ as well as the converse. In the discrete case, the copula is uniquely determined on $Ran(F_1) \times \ldots \times Ran(F_p)$, where $Ran(F_j)$ is the range of marginal F_j. See [Joe, 2014] for a comprehensive treatment of copulas. Parametric copula families are typically defined on uniform random variables obtained through CDF transformations from the marginals. In a *Gaussian Mixture Copula Model (GMCM)* [Tewari *et al.*, 2011], the dependence is obtained from a Gaussian Mixture (GMM). Note the difference from a mixture of copulas [Kosmidis and Karlis, 2015]. GMCM copula density is given by:

$$\mathcal{C}(\boldsymbol{\vartheta}, \mathbf{z}_i) = \frac{\sum_{g=1}^{G} \pi_g \phi(\mathbf{z}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)}{\prod_{j=1}^{p} \psi_j(z_{ij})} \tag{1}$$

The generative process is:

$$z_{ij} \sim p\text{-dimensional}, G\text{-component GMM, parameters: } \boldsymbol{\vartheta}$$
$$u_{ij} = \Psi_j(z_{ij})$$
$$y_{ij} = F_j^{-1}(u_{ij}) \text{ for } j = 1, \ldots, p \text{ and } i = 1, \ldots, n.$$

Notation: The observed data matrix is $\mathbf{Y} = \{y_{ij}\}$; indices i, j denote the observation ($i = 1, \ldots, n$) and dimension ($j = 1, \ldots, p$); F_j is the unknown marginal distribution for the j^{th} dimension, and uniform random variables $u_{ij} = F_j(y_{ij})$. Latent variables $\mathbf{Z} \in \mathbb{R}^{n \times p} = \{z_{ij}\}$, $\mathbf{z}_i$ denotes p–dimensional row vectors and $\mathbf{z}_j$ denotes n–dimensional column vector along the j^{th} dimension; Ψ_j and ψ_j are the marginal CDF and density of the GMM along the j^{th} dimension, ϕ is the multivariate Gaussian density and $\boldsymbol{\vartheta} = \{\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\}$ is the (unknown) parameter set representing mixing proportions ($\pi_g > 0$, with $\sum_{g=1}^{G} \pi_g = 1$), mean vectors ($\boldsymbol{\mu}_g$) and covariance matrices ($\boldsymbol{\Sigma}_g$; $\sigma_{gij} : ij^{th}$ entry of $\boldsymbol{\Sigma}_g$) for GMM component $g = 1, \ldots, G$. Vectors and matrices are in bold font.

Copula Parameter Estimation. Standard maximum likelihood inference requires specifying a parametric family for each marginal and inferring the parameters simultaneously with the copula parameters which is computationally prohibitive for even moderate dimensions. The two–step IFM procedure [Joe, 2014] requires fitting a marginal to each feature and then using the CDF transformation to obtain *pseudo data* for estimating copula parameters. Often marginal families are unknown and a semiparametric approach is to use rank–transformed scaled empirical marginals to obtain the pseudo data. The copula function for a joint distribution is invariant to monotone increasing transformations of marginals and the resulting pseudo–likelihood estimator is consistent for continuous marginals [Genest *et al.*, 1995].

With discrete marginals, rank transformation leads to ties with non–zero probability. In general, technical hurdles have to be surmounted to use copulas with discrete marginals: the copula may be unidentifiable and may not be margin–free, inference algorithms based on Kendall's tau or Spearman's rho may be biased or inconsistent [Genest and Neslehova, 2007]. The Extended Rank Likelihood (ERL) approach is an approximation to the full likelihood that can accommodate both continuous and discrete (ordinal and binary) data [Hoff, 2007].

The key idea of ERL is that, since F_j^{-1} is monotonic, although we don't observe variables z_i directly, we still know the order of z_i's induced by the data. Estimation of the copula parameters $\boldsymbol{\vartheta}$ using rank likelihood can be implemented by conditioning on the partial ordering induced by $\mathbf{Y}$. The ERL function is given by $P(\mathbf{Z} \in D|\boldsymbol{\vartheta})$; D is the rank–induced set: $D = \{\mathbf{Z} \in \mathbb{R}^{n \times p} : z_{lj} < z_{ij} < z_{uj}\}$ where $z_{lj} = max\{z_{kj} : y_{kj} < y_{ij}\}$ and $z_{uj} = min\{z_{kj} : y_{ij} < y_{kj}\}$, along the j^{th} dimension. Since the ERL approach *bypasses the estimation of marginals*, it can be applied to both discrete and continuous data. Note that the ERL function is equivalent to the distribution of the ranks for continuous data. For discrete data, the distribution of the ranks depends on the marginals, which is ignored resulting in some loss of information. Bayesian inference can be achieved by constructing a Markov chain having a stationary distribution equal to $P(\boldsymbol{\vartheta}|\mathbf{Z} \in D) \propto P(\boldsymbol{\vartheta}) \times P(\mathbf{Z} \in D|\boldsymbol{\vartheta})$.

3 Clustering with Gaussian Mixture Copulas

Our algorithm, EGMCM (*Extended* GMCM), fits a Gaussian mixture copula to data in a semiparametric manner and infers clusters from the obtained parameters. Algorithm 1 outlines our iterative scheme for obtaining approximate samples to estimate the posterior $P(\boldsymbol{\vartheta}|\mathbf{Z} \in D)$. Each iteration has 2 steps:

estimating GMM parameters $\boldsymbol{\vartheta}$ and resampling $\mathbf{Z}$.

The algorithm is initialized with two steps. First, rank transformed data is obtained by transforming each data point y_{ij} of the feature vector Y_j to $R(y_{ij}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_{ij} \leq y_{ij})$, the scaled rank of y_{ij} in all the observations of Y_j, the j^{th} feature (where $\mathbf{1}(c)$ denotes the indicator function which is 1 if the condition c is true and 0 otherwise). To initialize the values of $\mathbf{Z}$, we take the inverse CDF of a standard normal on the rank transformed data. We empirically find that this simple choice works well; note that the values of $\mathbf{Z}$ are resampled through the iterations.

We use Expectation Maximization (EM) to estimate $\boldsymbol{\vartheta}$. Gibb's sampling could also be used. If a single Gibb's sweep is used to sample $\boldsymbol{\vartheta}$ instead of a full estimation, the clustering performance deteriorates – also explained by our theoretical analysis which shows that initial unbiased estimates of $\boldsymbol{\vartheta}$ are important to obtain final unbiased mean estimates.

In the *Resample* $\mathbf{Z}$ step, we sample from each Gaussian in the GMM, truncated by the bounds z_{uj} and z_{lj} which constrains the sampled z to be in the set D. The Truncated Normal (**TN**) random variates are then combined using the mixing proportions of the estimated GMM. The function *unique*, that returns the set of unique values in the input, is used to avoid repeated computation for ties.

Algorithm 1 EGMCM

Input: $R(\mathbf{Y})$, scaled rank transformed data $\mathbf{Y}$; G, number of clusters
Initialization
  $\mathbf{Z} = \Phi^{-1}(R(\mathbf{Y}))$
loop
  Estimate, via EM, GMM parameters $\boldsymbol{\vartheta} = [\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g]$
  Resample $\mathbf{Z}$:
  for j: 1 **to** p **do**
    for all $y \in$ *unique* $\{y_{1j}, \ldots, y_{nj}\}$ **do**
      Compute $z_{lj} = \max\{z_{ij} : y_{ij} < y\}$ and $z_{uj} = \min\{z_{ij} : y < y_{ij}\}$
      For each i such that $y_{ij} = y$:
        Sample r_{gij} from $TN(\mu_{gj}, \sigma_{gij}, z_{lj}, z_{uj})$
        Set $z_{ij} = \sum_{g=1}^{G}\pi_g r_{gij}$
    end for
  end for
end loop
Output: Cluster labels (latent variables of GMM ($\mathbf{Z}|\boldsymbol{\vartheta}$))

The algorithm can easily handle data missing–at–random. If y_{ij} is missing, z_{ij} is obtained (in each *Resample* $\mathbf{Z}$ step) from the unconstrained GMM with current estimates of $\boldsymbol{\vartheta}$.

Computational Complexity Comparison

The time complexity of EGMCM is $\mathcal{O}(tGnp^2)$ where t is the number of iterations of the outer loop. The runtime is dominated by EM ($\mathcal{O}(Gnp^2)$). Empirically we obtain good results with $t < 50$ and so, the runtime is comparable to that of GMM estimation algorithms.

In comparison, note that the runtime for SCENIC is $\mathcal{O}(t'n^2p_v^2)$ where t' includes terms proportional to number of iterations in their algorithm and p_v is the dimension of the low–dimensional embedding; the runtime for coupledMC is $\mathcal{O}(n^2R^3)$ where R is the maximal number of attribute values for all features, including continuous features that are discretized in their method, hence $p \leq R \leq np$.

4 Theoretical Analysis

Our iterative algorithm has two steps: estimation of $\boldsymbol{\vartheta}$ through EM and the second *Resample* $\mathbf{Z}$ step. EM estimates of GMM parameters are consistent and asymptotically unbiased when the component means are not too close to each other; and for large sample sizes (> 200) the bias in mean and variance can be ignored for practical purposes [Nityasuddhi and Bohning, 2003]. Thus, assuming that the initial estimates obtained via EM in the first iteration of the loop are consistent and asymptotically unbiased, for large number of iterations m, we analyze the deviation in the mean from the initial estimate in subsequent *Resample* $\mathbf{Z}$ steps for discrete (theorem 2) and continuous (theorem 3) marginals. We show that the deviation tends to zero under simple conditions that can be maintained in the algorithm.

Similar properties cannot be inferred for the variance parameters because we are assuming a full variance–covariance matrix; thus the behaviour of the variance components after truncating the GMM will be dependent on the data, the nature of truncation and position of the component means with respect to the mixture means, among other factors. To our knowledge, similar analysis of previous ERL based methods has not been done before.

Notation: Superscript (m) denotes the iteration; z_l and z_u depend on g and j, i.e. the corresponding cluster and the dimension but we do not use them for notational simplicity. Let σ^2_{gj} denote the j^{th} diagonal element of $\boldsymbol{\Sigma}_g$, and δ_{gj} denote the contribution of the component mean of cluster g to the mixture mean along the j^{th} dimension, i.e. $\delta_{gj} = \pi_g\mu_{gj}/\sum_{g=1}^{G}\pi_g\mu_{gj}$. Also denote:

$$\alpha_{gj}^{(m)} = \frac{{z_l}^{(m)} - \mu_{gj}^{(m)}}{\sigma_{gj}^{(m)}}, \beta_{gj}^{(m)} = \frac{{z_u}^{(m)} - \mu_{gj}^{(m)}}{\sigma_{gj}^{(m)}},$$

$$\lambda_{gj}^{(m)} = \frac{\phi(\beta_{gj}^{(m)}) - \phi(\alpha_{gj}^{(m)})}{\Phi(\beta_{gj}^{(m)}) - \Phi(\alpha_{gj}^{(m)})}\frac{\sigma_{gj}^{(m)}}{\mu_{gj}^{(m)}}.$$

$\lambda_{gj}^{(m)}$ is the deviance in mean due to truncation of the g^{th} component of the normal distribution along the j^{th} dimension in the m^{th} iteration (which depends on the coefficient of variation, $\frac{\sigma}{\mu}$). Note that the mixture mean of the GMM is $\boldsymbol{\mu} = \sum_{g=1}^{G}\pi_g\boldsymbol{\mu}_g$ and the mixture variance is $\boldsymbol{\Sigma} = \sum_{g=1}^{G}\pi_g\left[\left(\boldsymbol{\mu}_g - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_g - \boldsymbol{\mu}\right)^T + \boldsymbol{\Sigma}_g\right]$. In addition to set D defined in section 2, we define the set $D_1 = \{\mathbf{Z} : z_{l1} \leq z_{ij} \leq z_{u1}\}$ where $z_{l1} = max\{z_{kj} : y_{kj} = y_{ij}\}$, $z_{u1} = min\{z_{kj} : y_{kj} = y_{ij}\}$. The initial estimate of mean, $\mu^{(0)}$ is obtained in the first iteration of the outer loop, through EM, and is considered asymptotically unbiased.

Lemma 1 *After m iterations of the loop in algorithm EGMCM, the deviation of the g-th component*

of mixture mean $\boldsymbol{\mu}^{(m)}$ from the g-th component of the initial estimate $\boldsymbol{\mu}^{(0)}$ along the j^{th} dimension is $\mu_{gj}^{(0)}\left[\prod_{k=1}^{m}(1-\lambda_{gj}^{(k-1)})-1\right]$.

Proof: According to the algorithm, for each $j \in \{1,2,...,p\}$, for each $i \in \{1,2,...,n\}$ and for each $y \in$ unique $\{y_{1j}, y_{2j}, ..., y_{nj}\}$, at the m^{th} iteration we sample r_{gij} from the truncated normal distribution $N(\mu_{gj}^{(m-1)}, {\sigma_{gj}^{(m-1)}}^2)$ with the truncation points (z_l, z_u), as described in the algorithm, for $g = 1, 2, ..., G$. Then z_{ij} is taken as $\sum_{g=1}^{G}\pi_g r_{gij}$ and $\boldsymbol{\mu}$ is estimated based on z_{ij}. But since r_{gij} is taken from a truncated normal distribution,

$$\begin{aligned}\mu_{gj}^{(m)} &= \mu_{gj}^{(m-1)} - \frac{\phi(\beta_{gj}^{(m-1)}) - \phi(\alpha_{gj}^{(m-1)})}{\Phi(\beta_{gj}^{(m-1)}) - \Phi(\alpha_{gj}^{(m-1)})}\sigma_{gj}^{(m-1)} \\ &= \mu_{gj}^{(m-1)}\left(1-\lambda_{gj}^{(m-1)}\right).\end{aligned}$$

Repeating the iteration and applying the above equation, after m iterations we have $\mu_{gj}^{(m)} = \mu_{gj}^{(0)}\prod_{k=1}^{m}(1-\lambda_{gj}^{(k-1)})$.

Theorem 1 *If the marginal distribution, F_j, is discrete and there exists some m_0 such that (z_l, z_u) does not contain 0 at the m_0^{th} iteration, and if $(z_u^{(m)} - z_l^{(m)})/\sigma_{gj}^{(m)} \to 0$ as $m \to \infty$ then $\lambda_{gj}^{(m)} \to 0$ as $m \to \infty$.*

Proof: For discrete F_j, if $z_l^{(m)}$ and $z_u^{(m)}$ are the two extremes at the m^{th} iteration, then $z_u^{(m)} - z_l^{(m)} \leq z_u^{(m-1)} - z_l^{(m-1)}$ for all m.

What if equality holds for all subsequent iterations? λ will not tend to 0 then. This is because of the discreteness of the marginals. Recall, for the extended rank likelihood, we consider the set D and not the set D_1. For discrete F_j, the measure of D_1 is non–zero. Thus $(z_l^D, z_u^D) \subset (z_l^{D\cup D_1}, z_u^{D\cup D_1})$ where z_l^D is the lowest order statistic taken based on the set D only, and $z_l^{D\cup D_1}$ is the lowest order statistic taken based on the union of D and D_1 (similarly for z_u). Thus at each iteration some positive measure will be assigned to the set D_1 and hence $z_u^{(m)} - z_l^{(m)} < z_u^{(m-1)} - z_l^{(m-1)}$, i.e. the relation will actually be a strict inequality.

Let us consider the sequence $s_m = z_u^{(m)} - z_l^{(m)}$. From the discussion above, we can infer that s_m is a positive sequence and $s_m < s_{m-1}$. Thus $s_m \to 0$ as $m \to \infty$. In other words, for each $\bar{m}$ there exists an ϵ such that $s_m < \epsilon$ for all $m \geq \bar{m}$. Thus $z_u^{(m)} - z_l^{(m)} \to 0$.

Now, since $\mu_{gj}^{(m)} \in \left(z_l^{(m)}, z_u^{(m)}\right)$, $(z_u^{(m)} - z_l^{(m)})/\sigma_{gj}^{(m)} \to 0$ implies that $(z_u^{(m)} - \mu_{gj}^{(m)})/\sigma_{gj}^{(m)} \to 0$ and $(z_l^{(m)} - \mu_{gj}^{(m)})/\sigma_{gj}^{(m)} \to 0$ as $m \to \infty$. Thus, using the expression of $\lambda_{gj}^{(m)}$ we get

$$\lim_{m\to\infty}\lambda_{gj}^{(m)} = \lim_{\delta\to 0}\frac{\left[\phi(\alpha_{gj}^{(m)}+\delta) - \phi(\alpha_{gj}^{(m)})\right]}{\left[\Phi(\alpha_{gj}^{(m)}+\delta) - \Phi(\alpha_{gj}^{(m)})\right]}\frac{\sigma_{gj}^{(m)}}{\mu_{gj}^{(m)}}, \quad (2)$$

where $\delta = (z_u^{(m)} - \mu_{gj}^{(m)})/\sigma_{gj}^{(m)}$. Using L'Hospital's rule,

$$\lim_{\delta\to 0}\frac{\phi(\alpha_{gj}^{(m)}+\delta) - \phi(\alpha_{gj}^{(m)})}{\Phi(\alpha_{gj}^{(m)}+\delta) - \Phi(\alpha_{gj}^{(m)})} = -2\alpha_{gj}^{(m)} \to 0.$$

Thus we show that z_l and z_u become closer to each other, and the change in difference of their respective CDFs also becomes smaller (since the Gaussian CDF is a uniformly continuous function of its arguments). Also, as z_l and z_u becomes closer and the subsequent iterations are based on samples taken from the shrinking set (z_l, z_u), $\sigma_{gj}^{(m)}$ tends to 0. Also, since there exists m_0 such that (z_l, z_u) does not contain 0 at the m_0^{th} iteration, it will not contain 0 for all the subsequent iterations since (z_l, z_u) is a shrinking set. Thus $\mu_{gj}^{(m)} \not\to 0$ for all $m \geq m_0$. Using these conditions in equation 2, we have $\lambda_{gj}^{(m)} \to 0$ as $m \to \infty$.

Theorem 2 *If the marginal distribution, F_j, is continuous and the initial values are sampled unrestrictedly (i.e. from the entire sample space), the deviation in the estimate of $\boldsymbol{\mu}$ along the j^{th} dimension after the m^{th} iteration, $\lambda_{gj}^{(m)} \to 0$ as $m \to \infty$.*

Proof: For continuous F_j, we have $\{\mathbf{Z}\} = D \cup D_1 = D$ (since the measure of D_1 is zero). Thus z_l and z_u will not change over the iterations, if the initial values are taken over the entire domain space of $\mathbf{z}$ along the j^{th} dimension. Thus, after the m^{th} iteration, as $m \to \infty$,

$$\phi(\alpha_{gj}^{(m)}) \to 0, \quad \phi(\beta_{gj}^{(m)}) \to 0$$

and

$$\Phi(\beta_{gj}^{(m)}) - \Phi(\alpha_{gj}^{(m)}) \to 1.$$

Similarly $\sigma_{gj}^{(m)} - \sigma_{gj}^{(m-1)} \to 0$ and $\mu_{gj}^{(m)} - \mu_{gj}^{(m-1)} \to 0$ as $m \to \infty$. Thus after iteration m, $\lambda_{gj}^{(m)} \to 0$ as $m \to \infty$.

Summary of Results. The deviation (due to repeated *Resample* $\mathbf{Z}$ steps) from an initial asymptotically unbiased estimate and subsequent loss of information depends on the initial values of $\mathbf{Z}$. For a continuous feature, the deviation along the dimension is less if the initial values range over the entire domain of $\mathbf{Z}$. More interestingly, it turns out that for the discrete case the deviation is less if the highest and lowest order statistics of $\mathbf{Z}$ along the particular dimension are close to each other, that is, when the number of levels in the discrete feature is less. The analysis implies that EGMCM works well when all features are continuous or when there is mixed data with ordinal features that do not have many levels. In the presence of discrete valued features, the number of iterations should be increased to obtain better estimates.

5 Experiments

We compare the performance of our algorithm, EGMCM, with that of SCENIC and CoupledMC. Previous experiments (in [Plant, 2012] and [Wang *et al.*, 2015] respectively) have shown that SCENIC outperforms *K-Means-Mixed* and INCONCO; and CoupledMC outperforms *K-Means-Mixed*, *K-Prototype*, *mixture* and *SpectralCAT*. In addition, we use GMM, with 'binarized' categorical features as a baseline.

Performance Metrics. To evaluate the performance of clustering algorithms we use Adjusted Rand Index (ARI) [Hubert and Arabie, 1985] (range $[-1, 1]$) and Adjusted Mutual Information (AMI) [Vinh *et al.*, 2010] (range $[0, 1]$), higher values indicate better clustering in both metrics.

5.1 Simulations

We simulate both numerical, with only continuous valued features, and mixed datasets with dependencies.

	Numerical			Mixed		
Setting	n	p	n/p	n	p	n/p
I	100	200	0.5	100	400	0.25
II	500	200	1.5	500	1000	0.5
III	100	20	5	500	400	1.25
IV	100	10	10	1000	400	2.5
V	100	5	20	100	20	5
VI	500	20	25	100	10	10
VII	500	10	50	500	40	12.5
VIII	1000	20	50	500	20	25
IX	500	5	100	1000	20	50
X	1000	5	200	1000	10	100

Table 1: Simulation settings for numerical data (left) and mixed data (right) with varying n: number of observations, p: dimension.

Numerical Data. The number of clusters is chosen to be $G = 2$. The ratio of the number of the two cluster sizes is 1:1. Each data point in the first cluster is a product of a sample from a Multivariate Normal distribution (MVN$(0, I_p/2)$ where I_p is a $p \times p$ Identity matrix) and a sample from a F distribution (F(df1 $= p/2$, df2 $= p/2$)). Each data point in the second cluster is a product of a sample from a Multivariate Normal distribution (MVN$(2, \Delta)$ where Δ is a matrix with $(i, j)^{th}$ element $0.9^{|i-j|}$) and a sample from a Uniform distribution (Unif$(0, 1)$). We use 10 simulation settings and simulate 25 datasets for each setting as shown in table 1.

Mixed Data. We first generate two clusters, each with $n/2$ $p/2$–dimensional observations with discrete data. This data is concatenated with $p/2$–dimensional numerical data (as above) to obtain an $n \times p$ dataset. Discrete data in the two clusters is sampled from 4 different 20–category multinomial distributions P_i $(i = 1, 2, 3, 4)$ that are generated in such way that in each distribution one or two of the 20 classes are assigned a large probability, and the rest are assigned very small probability. The large probability class(es) vary from one distribution to another. Let x_{ij}^k (y_{ij}^k) denote the i–th observation of the j–th feature in the k–th cluster of the numerical (discrete) data. Dependencies between the numerical and discrete data are imposed by the following two rules. If $x_{ij}^1 < 0$ then sample y_{ij}^1 from P_1, else from P_2. If $x_{ij}^2 < 0$ then sample y_{ij}^2 from P_3, else from P_4. We use 10 simulation settings and 25 datasets are generated for each setting, as shown in table 1.

Results

Variations in n and p. Each value in the plots shown below is the average obtained over 25 runs. To avoid visual clutter, we do not plot standard error bars. We find that the variance of the algorithms across the runs were low and not significant.

Figure 1 shows the AMI of all four algorithms over 10 different simulation settings for numerical (left) and mixed (right) data. As expected, GMM performs well on numerical data but not on mixed data while CoupledMC does reasonably well with mixed data. The performance of SCENIC is comparable to that of CoupledMC and GMM. In 2 out of the 10 settings for mixed data, it outperforms all other algorithms. Our algorithm EGMCM consistently outperforms all three algorithms for numerical data. For mixed data, it is the best in 6 out of 10 settings. Similar performance of all four algorithms is observed with respect to ARI in figure 1. Algorithm EGMCM has the highest ARI among the four methods in the case of numerical data and in 8 out of 10 settings for mixed data. In numerical data, the improvement is higher for simulations III-X (large n/p), lower for simulations I, II, indicating that the method seems to work better with large n/p (in accordance with our theoretical analysis). In mixed data, the improvement over GMM is high in all cases but with more improvement in settings with large n/p.

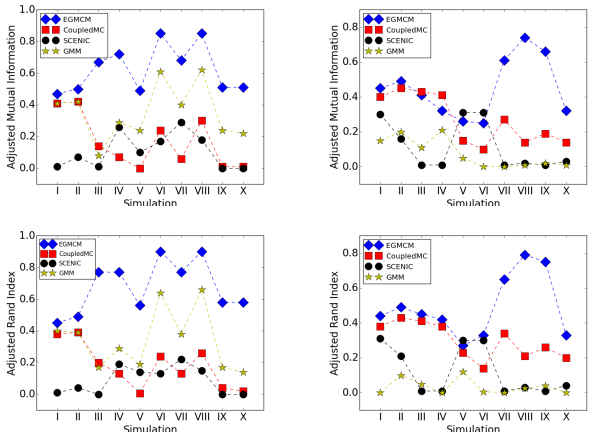

Figure 1: (Above) Average Adjusted Mutual Information (AMI), (Below) Average Adjusted Rand Index (ARI) of algorithms EGMCM, CoupledMC, SCENIC and GMM on Numerical data (left) and Mixed data (right) in ten different simulation settings.

Variations in G. We study the performance of the algorithms when the number of clusters, G, is varied for a single simulation setting: $n = 500, p = 20$. For mixed data, we use 20 numerical and 20 discrete features. Each value in the plots is the average obtained over 25 runs. Figure 2 show the AMI of all four algorithms with increasing number of clusters. The performance of all the four methods drop with increase in G. But EGMCM is consistently better than the rest for all values of G. In figure 2 we also see that the performance of EGMCM is better than the other algorithms, however the increase in ARI is not much except for $G = 2$ in the numerical case. In general, for numerical data, all four algorithms do not perform well with increasing number of clusters. On mixed data, EGMCM remains better than the other three algorithms.

5.2 Experiments on UCI Benchmark Datasets

We compare the performance of our algorithm on 10 benchmark datasets obtained from the UCI repository [Bache and Lichman, 2013]. Results, in table 3, show that EGMCM performs significantly better (p-value 0.001) than the other algo-

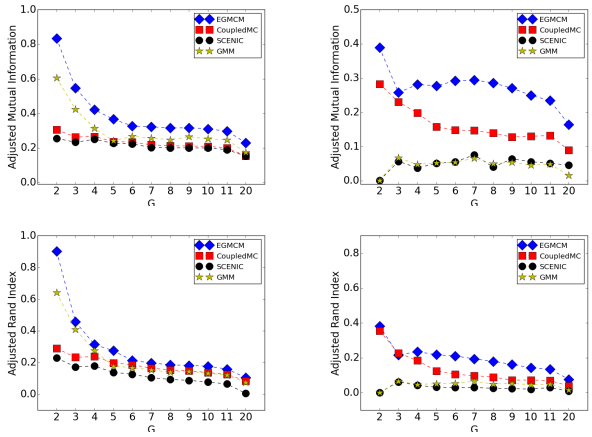

Figure 2: (Above) Average Adjusted Mutual Information (AMI), (Below) Average Adjusted Rand Index (ARI) of algorithms EGMCM, CoupledMC, SCENIC and GMM on Numerical data (left) and Mixed data (right) with varying number of clusters G.

rithms (at 1% significance level; using Friedman–post-hoc-Nemenyi test [Demšar, 2006]).

Data	Name	n	p_{num}	p_{cat}	G
A	Contraceptive Method Choice	1473	2	7	3
B	Heart Disease (Switzerland)*	303	5	8	2
C	Heart Disease (VA)*	303	5	8	5
D	Breast Cancer Wisconsin Prognostic*	286	33	0	2
E	Haberman's Survival*	306	3	0	2
F	Wine Quality	178	13	0	3
G	Parkinsons Telemonitoring	197	22	0	2
H	Climate Model Simulation Crashes	540	18	0	2
I	QSAR Bio–degradation	1055	32	9	2
J	Wholesale Customers	440	0	6	3

Table 2: Details of datasets from UCI repository used in our experiments. n: number of observations, p_{num}: number of numerical features, p_{cat}: number of discrete features, G: number of clusters. Asterisk: dataset contains missing values.

6 Discussion

We present a new clustering algorithm, EGMCM, that, using Gaussian mixture copulas (GMCM), can model a wide range of dependencies beyond those captured by meta–Gaussian distributions. We give the first semiparametric method to estimate the parameters of GMCM that can fit continuous, ordinal and binary data. We theoretically analyze the conditions required to obtain consistent and unbiased parameter estimates that, to our knowledge, is the first such analysis of an ERL approach. We empirically demonstrate that EGMCM outperforms state-of-the-art correlation clustering methods for mixed data. A limitation of EGMCM is that it is uninterpretable with nominal data.

EGMCM is more scalable than other dependency clustering methods for mixed data such as SCENIC and coupledMC. Processing of nominal data contributes to the computational expense of other correlation clustering methods but their runtime remains higher even when restricted to only ordinal and continuous data due to the expense of finding dependencies.

Data	Metric	EGMCM	CoupledMC	SCENIC	GMM
A	AMI	0.06	0.02	**0.25**	0
	ARI	0.01	0.04	**0.22**	0
B	AMI	**0.24**	0.03	0	0
	ARI	**0.26**	0.18	0	0
C	AMI	**0.23**	0.02	0	0
	ARI	**0.12**	0.02	0	0.02
D	AMI	**0.24**	0.01	0	0.01
	ARI	**0.14**	0.08	0.02	0
E	AMI	0.01	0	0	**0.06**
	ARI	0.02	0.01	0.01	**0.1**
F	AMI	**0.7**	0.34	0.01	0.17
	ARI	**0.56**	0.29	0	0.91
G	AMI	0.12	**0.21**	0	0
	ARI	0.15	**0.36**	0.04	0.01
H	AMI	**0.11**	0.01	0	0
	ARI	**0.11**	0.01	0.01	0
I	AMI	**0.15**	0.03	0.09	0
	ARI	**0.23**	0	0.02	0
J	AMI	**0.37**	0	0.02	0.03
	ARI	**0.39**	0.02	0.02	0.1

Table 3: Clustering performance of algorithms EGMCM, CoupledMC, SCENIC and GMM; UCI datasets details in table 2. Best results for each dataset (highest AMI, ARI) in **bold**.

The time complexity of EGMCM may be further improved using better sampling techniques such as those in [Kalaitzis and Silva, 2013]) and constrained covariance structures like Parsimonious Gaussian Mixture Model families [McNicholas and Murphy, 2008].

Our experiments have focussed on classification accuracy when the number of clusters is known. If unknown, the number of clusters can be inferred, for example, using the Bayesian Information Criterion that was found to be effective in the parameter estimation method for GMCM in [Bhattacharya and Rajan, 2014] for continuous data.

In our theoretical analysis we derive the conditions that ensure that the deviation in the mean, $\lambda_{gj}^{(m)}$ in the m^{th} iteration, from the initial unbiased EM estimate, tends to zero for large m. The value of $\lambda_{gj}^{(m)}$ for small m depends on the choice of initial values and variance in the original dataset. If suitable conditions for $\lambda_{gj}^{(m)}$ to be close to zero can be found for small m, then a low deviation from the original unbiased estimate of the mean can be maintained, which in turn ensures consistent and asymptotically unbiased final estimates. This follows from the fact that, if the maximum likelihood estimate of the mean lies in a small neighbourhood of the initial consistent and asymptotically unbiased estimate, then under certain regularity conditions, the estimate itself is also consistent and asymptotically unbiased [Huber, 1967]. Finding such conditions for small m remains an open problem.

The ERL function is equivalent to the distribution of the ranks for continuous data. For discrete data, the distribution of the ranks depends on the univariate marginals. By ignoring the marginal parameters, ERL loses information in the case of discrete marginals. It would be useful to have a precise characterization of this loss compared to the full likelihood involving marginal parameters. Preliminary work in this direction is in [Gu and Ghosal, 2009; Murray *et al.*, 2013; Hoff *et al.*, 2014], only for the Gaussian copula.

References

[Ahmad and Dey, 2007] Amir Ahmad and Lipika Dey. A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, 63(2):503–527, 2007.

[Bache and Lichman, 2013] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[Bhattacharya and Rajan, 2014] Sakyajit Bhattacharya and Vaibhav Rajan. Unsupervised learning using Gaussian Mixture Copula Model. In *International Conference on Computational Statistics (COMPSTAT)*, 2014.

[David and Averbuch, 2012] Gil David and Amir Averbuch. SpectralCAT: Categorical spectral clustering of numerical and nominal data. *Pattern Recognition*, 45(1):416–433, 2012.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[Elidan, 2013] Gal Elidan. Copulas in machine learning. In *Copulae in Mathematical and Quantitative Finance*, Lecture Notes in Statistics, pages 39–60. Springer, 2013.

[Genest and Neslehova, 2007] Christian Genest and Johanna Neslehova. A primer on copulas for count data. *Astin Bulletin*, 37(2):475, 2007.

[Genest *et al.*, 1995] C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995.

[Gu and Ghosal, 2009] Jiezhun Gu and Subhashis Ghosal. Bayesian ROC curve estimation under binormality using a rank likelihood. *Journal of Statistical Planning and Inference*, 139(6):2076–2083, 2009.

[Hoff *et al.*, 2014] Peter D. Hoff, Xiaoyue Niu, and Jon A. Wellner. Information bounds for Gaussian copulas. *Bernoulli*, 20(2):604–622, 2014.

[Hoff, 2007] Peter D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283, 2007.

[Huang, 1998] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

[Huber, 1967] P.J. Huber. The behaviour of maximum likelihood estimates under nonstandard conditions. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1967.

[Hubert and Arabie, 1985] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[Hunt and Jorgensen, 2011] Lynette Hunt and Murray Jorgensen. Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361, 2011.

[Jain, 2010] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[Joe, 2014] Harry Joe. *Dependence Modeling with Copulas*. CRC Press, 2014.

[Kalaitzis and Silva, 2013] Alfredo Kalaitzis and Ricardo Silva. Flexible sampling of discrete data correlations without the marginal distributions. In *NIPS*, 2013.

[Kosmidis and Karlis, 2015] Ioannis Kosmidis and Dimitris Karlis. Model-based clustering using copulas with applications. *Statistics and Computing*, pages 1–21, 2015.

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

[Murray *et al.*, 2013] Jared S. Murray, David B. Dunson, Lawrence Carin, and Joseph E. Lucas. Bayesian Gaussian copula factor models for mixed data. *Quarterly Journal of the Royal Meteorological Society*, 139(673):6982–991, 2013.

[Nityasuddhi and Bohning, 2003] Dechavudh Nityasuddhi and Dankmar Bohning. Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances. *Computational Statistics and Data Analysis*, 41:591–601, 2003.

[Plant and Böhm, 2011] Claudia Plant and Christian Böhm. Inconco: interpretable clustering of numerical and categorical objects. In *ACM SIGKDD*, 2011.

[Plant, 2012] Claudia Plant. Dependency clustering across measurement scales. In *ACM SIGKDD*, 2012.

[Rey and Roth, 2012] M. Rey and V. Roth. Copula mixture model for dependency-seeking clustering. In *ICML*, 2012.

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

[Tewari *et al.*, 2011] Ashutosh Tewari, Michael J. Giering, and Arvind Raghunathan. Parametric characterization of multimodal distributions with non-Gaussian modes. In *IEEE ICDM Workshop*, 2011.

[Vinh *et al.*, 2010] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

[Wang *et al.*, 2011] Can Wang, Longbing Cao, Mingchun Wang, Jinjiu Li, Wei Wei, and Yuming Ou. Coupled nominal similarity in unsupervised learning. In *CIKM*, 2011.

[Wang *et al.*, 2013] Can Wang, Zhong She, and Longbing Cao. Coupled attribute analysis on numerical data. In *IJCAI*, 2013.

[Wang *et al.*, 2015] Can Wang, Chi-Hung Chi, Wei Zhou, and Raymond Wong. Coupled interdependent attribute analysis on mixed data. In *AAAI*, 2015.