Computational Statistics and Data Analysis 101 (2016) 1–11

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

EM algorithm in Gaussian copula with missing data

Wei Ding[a,*], Peter X.-K. Song[b]

[a] Barclays Investment Bank, 745 7th Ave, New York, NY 10019, United States

[b] University of Michigan, MI, 48109, United States

ARTICLE INFO

Article history:
Received 7 October 2014
Received in revised form 8 September 2015
Accepted 18 January 2016
Available online 15 February 2016

Keywords:
Gaussian copula
EM algorithm
Misaligned missing data
Regression

ABSTRACT

Rank-based correlation is widely used to measure dependence between variables when their marginal distributions are skewed. Estimation of such correlation is challenged by both the presence of missing data and the need for adjusting for confounding factors. In this paper, we consider a unified framework of Gaussian copula regression that enables us to estimate either Pearson correlation or rank-based correlation (e.g. Kendall's tau or Spearman's rho), depending on the types of marginal distributions. To adjust for confounding covariates, we utilize marginal regression models with univariate location-scale family distributions. We establish the EM algorithm for estimation of both correlation and regression parameters with missing values. For implementation, we propose an effective peeling procedure to carry out iterations required by the EM algorithm. We compare the performance of the EM algorithm method to the traditional multiple imputation approach through simulation studies. For structured types of correlations, such as exchangeable or first-order auto-regressive (AR-1) correlation, the EM algorithm outperforms the multiple imputation approach in terms of both estimation bias and efficiency.

1. Introduction

Estimation of rank-based correlation is frequently required in practice to evaluate relationships between variables when they follow marginally skewed distributions. However, estimation of such correlation becomes a great challenge in the presence of missing data and with the need of adjusting for confounders. Most of recently published works on the copula models have been focused on analyzing fully observed data, e.g., Czado (2010), Joe et al. (2010), Genest et al. (2011), Masarotto and Varin et al. (2012) and Acar et al. (2012), and there is little knowledge available concerning how the analysis may be done in the presence of missing data.

In terms of handling missing data, the complete case analysis, which is often used in practice for convenience, simply discards any cases with missing values on those of the variables selected and proceeds with the analysis using standard methods. Obviously, the data attrition reduces the sample size, resulting potentially in a great loss of estimation efficiency. EM algorithm (Dempster et al., 1977) is a widely used iterative algorithm to carry out the maximum likelihood estimation in a statistical analysis with incomplete data. Multiple Imputation (Rubin, 2004) provides an alternative approach useful to deal with statistical analysis with missing values. Instead of filling in a single value for each missing value, Rubin (2004) multiple imputation procedure actually replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. When data come from skewed distributions, Hot-Deck Imputation (Andridge and Little, 2010) is also widely used, where a missing value is imputed with a randomly drawn similar record in terms of the nearest neighbor

[*] Corresponding author. Tel.: +1 7346042397.
E-mail addresses: dingwei.vivian@gmail.com, dingwei@umich.edu (W. Ding), pxsong@umich.edu (P.X.-K. Song).

http://dx.doi.org/10.1016/j.csda.2016.01.008
0167-9473/© 2016 Elsevier B.V. All rights reserved.

criterion. One caveat of Hot-Deck imputation is that it is a single imputation method, which may fail to provide desirable uncertainty associated with missing values. In addition, the number of imputed data sets is critical to obtain proper data analysis results, and a small number may lead to inappropriate inference. Some researchers have recommended 20–100 imputation data sets or even more (Graham et al., 2007), which appears computationally costly in practice. The imputation methods may become nontrivial and no longer straightforward when data distributions are skewed and adjusting for confounding factors is needed.

Multi-dimensional regression models for correlated data involve typically the specification of both correlation structures and marginal mean models that can be formulated by the classical univariate generalized linear model (GLM) (Nelder and Baker, 1972). Although the great popularity of quasi-likelihood approaches to analyzing correlated data, such as generalized estimating equation (GEE) (Liang and Zeger, 1986) and quadratic inference function (QIF) (Qu et al., 2000), a fully specified probability model with interpretable correlation structures is actually a desirable device to achieve the objective of evaluating correlations between variables. It is known that in the quasi-likelihood method correlations are treated as nuisance parameters, so that their estimation and interpretation are not of primary interest in data analysis.

In this paper we consider the Gaussian copula regression model (Song, 2000; Song et al., 2009) as the probability model for the correlated data because of the following meritorious features. First, the copula model allows us to define, evaluate and interpret correlations between variables in a full probability manner, very similar to the classical multivariate normal distribution. Second, from the copula model various types of correlations are provided to answer for different questions. For example, it provides Pearson linear correlation or rank-based nonlinear correlations (Kendall's tau or Spearman's rho), depending on if the marginal distributions are normal or skewed. Moreover these correlations may be obtained either in a form of unconditional marginal pairwise correlation, or in a form of conditional pairwise correlation. Third, the copula model has the flexibility to incorporate marginal GLMs to adjust for confounding factors, which is of practical importance. Last, the availability of the full probability model gives rise to the great ease of implementing powerful EM algorithm to handle missing data in a broad range of multi-dimensional models where the regression parameters in the mean model and the correlation parameters can be estimated simultaneously under one objective function. In such a framework, both estimation and inference are safeguarded by the well-established classical maximum likelihood theory.

It is of interest in the context of copula models to investigate and compare the two principled methods of handling missing data, EM algorithm and multiple imputation, as well as their computational complexity. Since the development of the EM algorithm is not trivial in the framework of Gaussian copula models, we propose an efficient peeling procedure to update model parameters in the M-step due to the involvement of a multi-dimensional integral. To adjust for confounding factors in the marginals, we focus on the location-scale family distribution in marginal regression models to embrace the flexibility of marginal distributions.

We compare the performance of the EM algorithm to the multiple imputation approach through simulation studies. For structured types of correlations, such as exchangeable or first-order auto-regressive (AR-1) correlation matrix, the EM algorithm method outperforms the multiple imputation approach in both aspects of estimation bias and efficiency. These two approaches perform similarly when the correlation matrix is unstructured.

This paper is organized as follows. Section 2 describes the Gaussian copula model. Together with some examples of practically useful models, Section 3 presents the details of the EM algorithm and Louis' formula (Louis, 1982) for standard error calculation. Section 4 presents simulation study, and a data analysis is included in Section 5. Section 6 provides some concluding remarks.

2. Model

The focus of this paper is on using EM algorithm in Gaussian copula to estimate of correlation with missing data. We assume that there are n partially observed subjects. For a subject, let $Y = (y_1, y_2, \ldots, y_d)'$ be a d-dimensional random vector of continuous outcomes, part of which is observed and the other part is missing. Denote by R_j as a missing data indicator, where $R_j = 0$ or 1 if the jth element y_j is missing or observed. Note that this indicator is known but varies for different subjects. Let y_{mis} be the set of variables with missing data, and y_{obs} be the set of variables with observed data of a subject.

2.1. Location-scale family distribution marginal model

Suppose $\theta = (\theta_1, \theta_2, \ldots, \theta_d)'$, where each θ_j denotes a set of marginal parameters associated with the jth ($j = 1, \ldots, d$) marginal density function, $f_j(y_j|\theta_j)$. Denote by $u_j = F_j(y_j|\theta_j)$ the marginal cumulative distribution function(CDF) corresponding to the jth margin, where F_j is a location-scale family distribution parametrized by a location parameter μ_j and a positive scale parameter σ_j, $\theta_j = (\mu_j, \sigma_j)$. More specifically, the marginal location-scale density function is given by

$$f_j(y_j|\theta_j) = \frac{1}{\sigma_j}\tilde{f}\left(\frac{y_j - \mu_j}{\sigma_j}\right), \quad j = 1, \ldots, d, \tag{1}$$

where $\tilde{f}(\cdot)$ is the standard kernel density with $\int_R y\tilde{f}(y)dy = 0$, and $\int_R y^2\tilde{f}(y)dy = 1$. In this paper, $\tilde{f}$ may be taken as a parametric or a nonparametric kernel density, and parameter μ_j or σ_j may be modeled as a function of confounding covariates.

2.2. Gaussian copula

A copula is a multivariate probability distribution in which the marginal probability distribution of each variable is uniform on $(0, 1)$. Sklar's theorem (Sklar, 1959) states that every multivariate cumulative distribution function of a continuous random vector $Y = (y_1, y_2, \ldots, y_d)'$ with marginals $F_j(y_j|\theta_j)$ can be written as $F(y_1, \ldots, y_d) = C(F_1(y_1), \ldots, F_d(y_d))$, where C is a certain copula. In this paper, Y is assumed to follow a d-dimensional distribution generated by a Gaussian copula (Song, 2000), whose density function is given by

$$f(Y|\theta, \Gamma) = c(u|\Gamma)\prod_{j=1}^{d} f_j(y_j|\theta_j), \quad u = (u_1, u_2, \ldots, u_d)' \in [0, 1]^d, \tag{2}$$

where $c(u|\Gamma) = c(u_1, \ldots, u_d|\Gamma)$, $u \in [0, 1]^d$, is the Gaussian copula density, with $u_j = F_j(y_j|\theta_j),\ i = 1, \ldots, d$, and Γ is an $d \times d$ matrix of correlation.

Let $q_j = q_j(u_j) = \Phi^{-1}(u_j)$ be the jth marginal normal quantile, where Φ is CDF of the standard normal distribution. According to Song (2007), the joint density of a Gaussian copula function $c(\cdot|\Gamma)$ takes the form:

$$c(u|\Gamma) = |\Gamma|^{-\frac{1}{2}} \exp\left\{\frac{1}{2}Q(u)^T(\mathrm{I} - \Gamma^{-1})Q(u)\right\}, \quad u \in [0, 1]^d \tag{3}$$

where $\Gamma = [\gamma_{j_1j_2}]_{d\times d}$ is the Pearson correlation matrix of $Q(u) = (q_1(u_1), \ldots, q_d(u_d))'$, and I is the $d \times d$ identity matrix. Here $|\cdot|$ denotes the determinant of a matrix. Marginally, $u_j \sim \mathrm{Uniform}(0, 1)$, and $q_j \sim \mathrm{Normal}(0, 1)$. When y_j is marginally normal distributed, matrix Γ gives the Pearson correlation matrix of Y; otherwise, Γ represents as a matrix of pairwise rank-based correlations. In fact, given a matrix Γ in Eq. (3), two types of pairwise rank-based correlations, Kendall's tau $(\left[\tau_{j_1j_2}\right]_{d\times d})$ and Spearman's rho $(\left[\rho_{j_1j_2}\right]_{d\times d})$ can be obtained as follows: $\tau_{j_1j_2} = \frac{2}{\pi}\arcsin(\gamma_{j_1j_2})$, and $\rho_{j_1j_2} = \frac{6}{\pi}\arcsin(\frac{\gamma_{j_1j_2}}{2})$ for $j_1, j_2 = 1, \ldots, d, j_1 \neq j_2$, respectively (McNeil et al., 2010).

2.3. Examples of marginal models

Among many possible marginal models, here we present two examples of marginal models to illustrate our proposed method, with or without the inclusion of covariates. These two following models are practically useful.

Example-1: marginal parametric distribution

To adjust for confounding factors in the mean marginal model, let $X_i = (1, x_i^T)^T,\ i = 1, \ldots, n$. For the jth margin, the linear model is imposed on the location parameter in Eq. (1), $\mu_{ij} = \mathrm{E}(y_{ij}|X_i) = h(X_i^T\beta_j), j = 1, \ldots, d$, where $\beta_j = (\beta_{j0}, \beta_{j1}, \ldots, \beta_{jp})'$ is a $(p + 1)$-element unknown regression vector, and h is a link function. For convenience, denote the resulting model by $Y_{ij} \sim F_j(y_j|\mu_{ij}(\beta_j), \sigma_j)$.

As an important special case, we consider $p = 0$ (no covariates), and thus $\mu_{ij} = h(\beta_{j0})$ is a common parameter for all subjects $i = 1, \ldots, n$. More generally, the marginal distribution model with the CDF $u_{ij} = F_j(y_j|\theta_j)$ may be a generalized location-scale family distribution, such as gamma distribution, of which the location parameter is 0, and the estimation procedure remains the same under a given marginal parametric distribution. This will be discussed as an example in simulation study in Section 4.1.

Example-2: Semi-parametric marginal distribution

If the type of the density function $f_j(y_j),\ j = 1, \ldots, d$ is unknown, there are several possible forms available to specify Eq. (1). In this paper, we consider an example of fully unspecified marginal distribution function $F_j(y_j)$, which will be estimated using the empirical distribution function. In this case, all the marginal parameter θ_j is absorbed into the CDF.

3. EM algorithm

Our goal is to estimate the model parameter (θ, Γ) in the presence of missing data. This may be achieved by utilizing the EM algorithm. We propose an effective peeling procedure in the EM algorithm, which serves as a core engine to speed up the calculation of M-step in the copula model. Both E-step and M-step are discussed in detail in Section 3.1, and the examples will be revisited in Section 3.2, respectively. Note that the EM algorithm assumes implicitly the missing at random (MAR) mechanism. This is because that the E-step of the algorithm requires to have identifiable conditional distributions of variables with missing data given all the other observed variables.

3.1. Expectation and maximization

Computing the likelihood of (θ, Γ) and iteratively updating the model parameter (θ, Γ) by maximizing the observed likelihood constitute the two essential procedures of the EM algorithm, corresponding respectively to the expectation step

(E-step) and the maximization step (M-step). The details of these two steps are discussed below under the setting where the forms of parametric marginal location-scale distributions are given. When these marginal distribution of forms are unspecified, we replace them by the corresponding empirical CDFs (see Example-2 above), and the resulting approximate likelihood will be used in the EM algorithm.

E-step

Denote by u_{obs} the subvector of observed margins of u and u_{mis} the subvector of margins with missing values; similarly, q_{obs} and q_{mis} denote the corresponding subvectors of transformed quantiles. Let D_{obs} and D_{mis} be the sets of indices for components with observed data and missing data, respectively. Then $D = D_{\text{obs}} \cup D_{\text{mis}}$ is the set of all indices, and $D_{\text{obs}} \cap D_{\text{mis}}$ is an empty set. Note that both D_{obs} and D_{mis} are subject-dependent, and its partition varies across subjects. Let $d_m = \dim(y_{\text{mis}}) = |D_{\text{mis}}|$.

At the E-step, the primary task is to calculate $\lambda(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, y_{\text{obs}})$ for each subject, where the pair $(\theta^{(t)}, \Gamma^{(t)})$ is the updated values of (θ, Γ) obtained from the tth iteration. For the ease of exposition, suppress index i in the following formulas. Given a subject, the λ-function $\lambda(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, y_{\text{obs}})$ is the expected value of the log likelihood function of (θ, Γ) with respect to the conditional distribution of y_{mis} given y_{obs} and $(\theta^{(t)}, \Gamma^{(t)})$:

$$\begin{aligned}\lambda(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, y_{\text{obs}}) &= \int_{R^{d_m}} \ln\{f(y|\theta, \Gamma)\} f\left(y_{\text{mis}}|y_{\text{obs}}, \theta^{(t)}, \Gamma^{(t)}\right) \mathrm{d}y_{\text{mis}} \\ &= \sum_{j\in D_{\text{obs}}} \ln\left\{f_j(y_j|\theta_j)\right\} + \int_{(0,1)^{d_m}} \ln\{c(u|\theta, \Gamma)\}\, c\left(u_{\text{mis}}|u_{\text{obs}}, \theta^{(t)}, \Gamma^{(t)}\right) \mathrm{d}u_{\text{mis}} \\ &\quad + \sum_{j\in D_{\text{mis}}} \int_0^1 \ln\left[f_j\left\{F_j^{-1}(u_j|\theta_j)|\theta_j\right\}\right] c\left(u_j|u_{\text{obs}}, \theta_j^{(t)}, \Gamma^{(t)}\right) \mathrm{d}u_j, \end{aligned} \tag{4}$$

where the right-hand side of Eq. (4) consists of three terms. The first term $\sum_{j\in D_{\text{obs}}} \ln\left\{f_j(y_j|\theta_j)\right\}$ is a sum of marginal likelihoods over those observed margins $j \in D_{\text{obs}}$, which can be evaluated directly. The second term is the observed likelihood, although it is of d_m dimension, its closed form expression can be analytically obtained. To do so, let $A = [A_{j_1 j_2}]_{d\times d} = \Gamma^{-1}$ be the precision matrix. The log copula density may be rewritten as follows:

$$\ln c(u|\theta, \Gamma) = \frac{1}{2}\ln|A| + \frac{1}{2}\sum_{j=1}^{d}(1 - A_{jj})q_j^2 - \frac{1}{2}\sum_{j_2\neq j_1}^{d} A_{j_1 j_2} q_{j_1} q_{j_2}. \tag{5}$$

It follows from Eq. (5) that

$$\begin{aligned}&\int_{(0,1)^{d_m}} \ln\{c(u|\theta, \Gamma)\} c\left(u_{\text{mis}}|u_{\text{obs}}, \theta^{(t)}, \Gamma^{(t)}\right) \mathrm{d}u_{\text{mis}} \\ &= \frac{1}{2}\ln|A| + \frac{1}{2}\sum_{j\in D_{\text{obs}}}(1 - A_{jj})q_j^2 + \frac{1}{2}\sum_{j\in D_{\text{mis}}}(1 - A_{jj})\int_R q_j^2 \phi(q_j|q_{\text{obs}}, \theta^{(t)}, \Gamma^{(t)})\mathrm{d}q_j \\ &\quad - \frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{obs}}} A_{j_1 j_2} q_{j_1} q_{j_2} - \sum_{j_1\in D_{\text{obs}}} q_{j_1} \sum_{j_2\in D_{\text{mis}}} A_{j_1 j_2}\int_R q_{j_2}\phi(q_{j_2}|q_{\text{obs}}, \theta^{(t)}, \Gamma^{(t)})\mathrm{d}q_j \\ &\quad - \frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{mis}}} A_{j_1 j_2}\int_{R^2} q_{j_1} q_{j_2}\phi_2(q_{j_1}, q_{j_2}|q_{\text{obs}}, \theta^{(t)}, \Gamma^{(t)})\mathrm{d}q_{j_1}\mathrm{d}q_{j_2} \\ &= \frac{1}{2}\ln|A| + \frac{1}{2}\sum_{j\in D_{\text{obs}}}(1 - A_{jj})q_j^2 \\ &\quad + \frac{1}{2}\sum_{j\in D_{\text{mis}}}(1 - A_{jj})\left[1 - (\Gamma_{\text{obs},j}^{(t)})^T(\Gamma_{\text{obs,obs}}^{(t)})^{-1}\Gamma_{\text{obs},j}^{(t)} + \left\{(\Gamma_{\text{obs},j}^{(t)})^T(\Gamma_{\text{obs,obs}}^{(t)})^{-1}q_{\text{obs}}^{(t)}\right\}^2\right] \\ &\quad - \frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{obs}}} A_{j_1 j_2} q_{j_1} q_{j_2} + \sum_{j_1\in D_{\text{obs}}}\sum_{j_2\in D_{\text{mis}}} A_{j_1 j_2} q_{j_1}\left\{(\Gamma_{\text{obs},j_2}^{(t)})^T(\Gamma_{\text{obs,obs}}^{(t)})^{-1}q_{\text{obs}}^{(t)}\right\} \\ &\quad - \frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{mis}}} A_{j_1 j_2}\left\{\Gamma_{j_1,j_2}^{(t)} - (\Gamma_{\text{obs},j_1}^{(t)})^T(\Gamma_{\text{obs,obs}}^{(t)})^{-1}\Gamma_{\text{obs},j_2}^{(t)}\right\} \\ &\quad - \frac{1}{2}\sum_{j_1\neq j_2\in D_{\text{mis}}} A_{j_1 j_2}\left\{(\Gamma_{\text{obs},j_1}^{(t)})^T(\Gamma_{\text{obs,obs}}^{(t)})^{-1}q_{\text{obs}}^{(t)}\right\}\left\{(\Gamma_{\text{obs},j_2}^{(t)})^T(\Gamma_{\text{obs,obs}}^{(t)})^{-1}q_{\text{obs}}^{(t)}\right\}, \end{aligned} \tag{6}$$

where $\Gamma_{\text{obs},j}$ is the jth column of Γ with observed margins, and $\Gamma_{\text{obs,obs}}$ is a submatrix of Γ, whose columns and rows are observed margins. Also, $\phi(\cdot)$ is the univariate normal density, and $\phi_2(\cdot)$ is the bivariate normal density. The third term in Eq. (4) may be rewritten as follows:

$$\sum_{j\in D_{\text{mis}}} \int_0^1 \ln\left[f_j\left\{F_j^{-1}(u_j|\theta_j)|\theta_j\right\}\right] c\left(u_j|u_{\text{obs}}, \theta_j^{(t)}, \Gamma^{(t)}\right) \mathrm{d}u_j = \sum_{j\in D_{\text{mis}}} E\left[\ln\left\{f_j\left(F_j^{-1}(u_j|\theta_j)|\theta_j\right)\right\} |y_{\text{obs}}, \theta_j^{(t)}, \Gamma^{(t)}\right], \tag{7}$$

where u_j is the CDF of normally distributed quantile q_j with mean $(\Gamma_{\text{obs},j}^{(t)})^T(\Gamma_{\text{obs,obs}}^{(t)})^{-1}q_{\text{obs}}^{(t)}$, and variance $\left\{1 - (\Gamma_{\text{obs},j}^{(t)})^T (\Gamma_{\text{obs,obs}}^{(t)})^{-1}\Gamma_{\text{obs},j}^{(t)}\right\}$, and the expectation $E(\cdot)$ may be evaluated numerically using the method of Gaussian quadratures (Abramowitz and Stegun, 1972). The observed likelihood for the full data of n subjects is expressed as:

$$\lambda(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, Y_{\text{obs}}) = \sum_{i=1}^{n} \lambda_i(\theta, \Gamma|\theta^{(t)}, \Gamma^{(t)}, y_{i,\text{obs}}), \tag{8}$$

where function $\lambda_i(\cdot)$ is given by Eq. (4). It is worth noting that Eq. (6) is of critical importance as it turns a d_m-dimensional integral a closed form expression, which ensures the E-step to be numerically feasible and stable. As a result, the evaluation of the E-step is computationally fast.

M-step

In the M-step we update parameters values by maximizing (8) with respect to θ and Γ. Following the ECM algorithm (Meng and Rubin, 1993), we will execute the M-step with several computationally simpler CM-steps. We propose a peeling procedure to facilitate the computation in the M-step, which consists of four routines given as follows.

Step M-1: updating marginal parameters

For a specific marginal parameter θ_j, we obtain its update by sequentially maximizing the observed likelihood (8) as follows, for $j = 1, \dots, d$,

$$\theta_j^{(t+1)} = \arg\max_{\theta_j} \sum_{i=1}^{n} \lambda_i(\theta_1^{(t+1)}, \dots, \theta_{j-1}^{(t+1)}, \theta_j, \theta_{j+1}^{(t)}, \dots, \theta_d^{(t)}|\Gamma^{(t)}, y_{i,\text{obs}}).$$

This optimization is carried out numerically by a quasi-Newton optimization routine available in *R* function *nlm*, and this step is computationally fast as the optimization involves only a set of low-dimensional parameters θ_j at one time.

Step M-2: updating correlation parameters

If Γ is an unstructured correlation matrix, each off-diagonal element $\gamma_{j_1j_2}$ is updated by maximizing the observed log-likelihood (8), which has a closed form expression. That is, for $j_1, j_2 = 1, \dots, d, j_1 \neq j_2$,

$$\gamma_{j_1j_2}^{(t+1)} = \frac{\sum_{i=1}^{n} q_{ij_1}^{(t)} q_{ij_2}^{(t)} 1(R_{ij_1} = 1)1(R_{ij_2} = 1)}{\sum_{i=1}^{n} 1(R_{ij_1} = 1)1(R_{ij_2} = 1)}, \tag{9}$$

where $1(\cdot)$ is an indicator function. Note that the diagonal elements $\gamma_{jj} = 1, j = 1, \dots, d$.

If Γ is a structured correlation matrix such as exchangeable or first-order auto-regressive correlation, say $\Gamma = \Gamma(\gamma)$, we update the correlation parameter γ by maximizing Eq. (8). This can be done numerically by applying *R* function *optim*(Nelder & Mead, 1965). In both cases of exchangeable and first-order auto-regressive correlations, there is only one correlation parameter involved in optimization, and the related computing is fast.

Step M-3: updating quantiles

For each subject $i = 1, \dots, n$, the quantiles are updated by the posterior mean for each margin $j = 1, \dots, d$, as follows:

$$q_{ij}^{(t+1)} = \begin{cases} \Gamma_{j,-j}^{(t+1)} \left(\Gamma_{-j,-j}^{(t+1)}\right)^{-1} \left(q_{i,-j}^{(t+1)}\right)^T, & j \in D_{i,\text{mis}} \\ \Phi^{-1}\left\{F_j\left(y_{ij}|\theta_j^{(t+1)}\right)\right\}, & j \in D_{i,\text{obs}}, \end{cases} \tag{10}$$

where $\Gamma_{j,-j}^{(t+1)}$ denotes the jth row vector of matrix $\Gamma^{(t+1)}$ without the jth element, $\Gamma_{-j,-j}^{(t+1)}$ is a submatrix of matrix $\Gamma^{(t+1)}$ without the jth row and the jth column, and $q_{i,-j}^{(t+1)}$ is the subvector of quantiles for subject i, $q_i^{(t+1)}$, with the jth element deleted. Note that the quantile updating is carried out by borrowing information from the other correlated variables via matrix $\Gamma^{(t+1)}$.

Step M-4: updating outcome values

Based on the updated parameter $\theta^{(t+1)}$ and quantiles $q_{ij}^{(t+1)}$, the outcome values are updated as follows:

$$y_{ij}^{(t+1)} = \begin{cases} F_j^{-1}\left\{\Phi\left(q_{ij}^{(t+1)}|\theta_{ij}^{(t+1)}\right)\right\}, & j \in D_{i,\text{mis}} \\ y_{ij}, \quad j \in D_{i,\text{obs}}. \end{cases} \tag{11}$$

3.2. Examples revisited

Now we revisit the examples outlined in Section 2.3 in connection to the EM algorithm.

Example-1: marginal parametric distribution

Example 1 is straightforward, and the marginal parameters and correlation parameters can be estimated by directly applying the above EM algorithm.

Example-2: semi-parametric marginal distribution

Since the marginal CDFs are no longer parametric, the step of updating marginal parameters $\theta_1, \ldots, \theta_d$ in the EM algorithm is void. At each iteration, we need to update the missing values via Step M-4 and update matrix Γ via Step M-2. In addition, quantiles q_{ij}, $j \in D_{i,\text{mis}}$ are updated by Step M-3, and consequently the uniform variates $u_{ij}, j \in D_{i,\text{mis}}$ are updated as follows,

$$u_{ij}^{(t+1)} = \frac{1}{n}\left\{\sum_{k=1}^{n} 1(q_{kj}^{(t)} < q_{ij}^{(t)}) + \frac{1}{2}\right\}, \quad \text{and} \quad q_{ij}^{(t+1)} = \Phi^{-1}\left(u_{ij}^{(t+1)}\right), \quad j = 1, 2, \ldots, d, \tag{12}$$

where the term $\frac{1}{2}$ in Eq. (12) is used to avoid $u_{ij}^{(t+1)} = 0$ leading to $q_{ij}^{(t+1)} = -\infty$, which causes numerical problem in the EM algorithm.

3.3. Standard error calculation

Louis' formula (Louis, 1982) is a well-known procedure useful to obtain standard errors of the estimates from the EM algorithm. As shown in Eq. (13), the observed Fisher Information matrix can be obtained via two information matrices. The first term in Eq. (13) is the expected full-data information matrix, while the second is the expected missing data information matrix. For the ease of exposition, suppress index i in the following formulas.

$$\begin{aligned} I(\hat{\theta}, \hat{\Gamma}) &= -\nabla^2 \ln\{f(y_{\text{obs}}|\theta, \Gamma)\}\,|_{\theta=\hat{\theta}, \Gamma=\hat{\Gamma}} \\ &= -I_{\text{full}} + I_{\text{mis}} \\ &= -\int \nabla^2 \ln\{f(y_{\text{mis}}, y_{\text{obs}}|\theta, \Gamma)\}\,|_{\theta=\hat{\theta}, \Gamma=\hat{\Gamma}} f(y_{\text{mis}}|y_{\text{obs}}, \hat{\theta}, \hat{\Gamma}) dy_{\text{mis}} \\ &\quad + \int \nabla^2 \ln\{f(y_{\text{mis}}|y_{\text{obs}}, \theta, \Gamma)\}\,|_{\theta=\hat{\theta}, \Gamma=\hat{\Gamma}} f(y_{\text{mis}}|y_{\text{obs}}, \hat{\theta}, \hat{\Gamma}) dy_{\text{mis}} \end{aligned} \tag{13}$$

where ∇^2 denotes the second order derivative with respect to the model parameters, and $(\hat{\theta}, \hat{\Gamma})$ are the estimates obtained as the final outputs of the EM algorithm. Therefore, the Fisher Information matrix is

$$I(\hat{\theta}, \hat{\Gamma}) = \sum_{i=1}^{n} I_i(\hat{\theta}, \hat{\Gamma}), \tag{14}$$

where $I_i(\hat{\theta}, \hat{\Gamma}) = -\nabla^2 l_i(\hat{\theta}, \hat{\Gamma})$, and $l_i(\hat{\theta}, \hat{\Gamma})$ is the observed log likelihood evaluated at the estimates for subject i, which can be calculated numerically via the following expression:

$$l_i(\hat{\theta}, \hat{\Gamma}) = \frac{1}{2}\ln(|\hat{A}_i|) + \frac{1}{2}\sum_{j \in D_{i,\text{obs}}}\left(1 - \hat{A}_{i,jj}\right)\hat{q}_{ij}^2 - \frac{1}{2}\sum_{j_1 \neq j_2 \in D_{i,\text{obs}}} \hat{A}_{i,j_1j_2}\hat{q}_{ij_1}\hat{q}_{ij_2} + \sum_{j \in D_{i,\text{obs}}} \ln\left\{f_j(y_{ij}|\hat{\theta}_j)\right\}. \tag{15}$$

Here $A_i = (\Gamma_i)^{-1} = [A_{i,j_1j_2}]_{d_{m,i} \times d_{m,i}}$, where Γ_i is the submatrix of Γ whose columns correspond to the observed variables in y_i for subject i, and $d_{m,i}$ counts the dimensions. By *R* function *hessian*, the Hessian function of Eq. (15) can both be numerically carried out. This provides the observed Fisher information matrix I, and moreover the asymptotic variance for $(\hat{\theta}, \hat{\Gamma})$ is $\mathrm{I}(\hat{\theta}, \hat{\Gamma})^{-1}$.

Table 1
Simulation results of correlation (Kendall's tau) parameters estimation in copula model for marginal skewed distributed data obtained by full data likelihood, EM algorithm and Imputation methods with different missing percentage. (Standard error ratio is calculated by a ratio of two standard errors between a method and the gold standard.)

%mis	Full data		Copula & EM		Multiple imputation		Hot deck imputation	
	bias($\times 10^{-2}$)	std.err	bias($\times 10^{-2}$)	std.err ratio	bias($\times 10^{-2}$)	std.err ratio	bias($\times 10^{-2}$)	std.err ratio
20%	0.10	0.0440	−0.14	1.0250	−1.78	1.0727	0.04	1.1273
	0.00	0.0402	−0.46	0.9851	−3.12	1.0995	−0.05	1.1144
	0.05	0.0434	−0.20	1.0069	−2.36	1.1060	−0.01	1.0945
30%	−0.03	0.0454	−0.30	1.0639	−2.77	1.0771	−0.13	1.1718
	−0.10	0.0416	−0.63	1.0673	−4.88	1.1370	−0.31	1.1563
	0.04	0.0442	−0.32	1.0452	−3.65	1.0905	0.01	1.1516
50%	−0.15	0.0437	−0.09	1.1716	−4.32	1.2449	−0.15	1.2792
	−0.14	0.0413	−0.55	1.1840	−7.10	1.2736	−0.46	1.3099
	−0.11	0.0428	−0.43	1.1869	−5.99	1.2453	−0.40	1.2944

3.4. *Initialization*

It is known that the quality of initial values is critical to the accuracy and efficiency of the EM algorithm. The initial parameters values $(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_d^{(0)}, \Gamma^{(0)})$ may be given by the estimates obtained from the complete case analysis, as suggested by Joe (2005). The sequential updating by the peeling algorithm enjoys numerical stability for the estimation of marginal regression parameters when the bias in the estimation of correlation parameters is asymptotically ignorable. Although theoretically the initial values may be set arbitrary, all numerical experiences have suggested that the closer initial values are to the true values, the faster the algorithm converges.

4. Simulation study

We conduct simulation experiments to evaluate and compare the performance of the EM algorithm with the multiple imputation method. In our experiments, the dimension of outcomes is set as $d = 3$, and $d_m = 1$ or 2 for different subjects. Three types of the correlation matrices Γ are considered: unstructured, exchangeable, and first-order autoregressive. Both Multiple Imputation (Little and Rubin, 2002) and Hot-deck Imputation (Andridge and Little, 2010) are included in the comparison.

Note that the R package of Multiple Imputation (R Package "*MI*") applied here is developed under multivariate normal distributions, so the skewness of the marginal distributions for outcomes may result in estimation bias. In Hot-Deck Imputation (R Package "*HotDeckImputation*"), as discussed above, each missing value is imputed by a randomly drawn similar record in terms of the nearest neighbor criterion. To adjust for confounders, Hot-Deck Imputation is adopted through the following steps. First, we run regression on the complete cases; second, impute residuals of the missing data, and then finally obtain imputed missing outcomes that will be used to run regression analysis on the "full" outcomes to yield the estimates of model parameters.

A naive approach is to use marginal data to obtain estimate of CDF $F_j(y_j|\hat{\theta}_j),\ j = 1, \ldots, d$, if F_j is a parametric model, or $\hat{F}_j(y_j),\ j = 1, \ldots, d$ by empirical CDF if F_j is nonparametric model, and make inverse-normal transformation $\hat{q}_j = \Phi^{-1}(F_j(y_j|\hat{\theta}_j))$ or $\hat{q}_j = \Phi^{-1}(\hat{F}_j(y_j))$, which are used to calculate $\mathrm{cor}(\hat{q}_{j_1}, \hat{q}_{j_2})$. Since the naive approach only uses marginal information and available data. Because it is inferior to imputation methods that replace the missing data with plausible values. So in this section, we did not include the naive approach in the comparison.

4.1. *Skewed marginal model*

We first examine the EM algorithm in the setting of the semi-parametric model discussed in Example-2, Section 2.3. In this case, only correlation parameters (Kendall's tau) are updated. To generate data, the marginal distributions are set as gamma distribution with the shape parameter $\alpha = 0.2$ and rate parameter $\beta = 0.1$, leading to the skewness 4.47. The correlation matrix Γ with $\gamma_{12} = 0.3, \gamma_{13} = 0.5, \gamma_{23} = 0.4$ is used, with the corresponding Kendall's tau being $(0.1940, 0.3333, 0.2620)$. We compare the results obtained from the full data without missingness (regarded as the gold standard) to the results obtained by the EM algorithm, Multiple Imputation, and Hot-Deck Imputation with incomplete data. The missingness percent varies from 20% to 50%. The sample size is fixed at 200, while 1000 replicates are run to draw summary statistics.

As shown in Table 1, with no surprise, in such a case of highly skewed distributions, the estimates of three Kendall's tau parameters obtained from Multiple Imputation are more biased. The estimation results from the EM algorithm and Hot-Deck Imputation are comparable, but the EM algorithm method provides smaller empirical standard errors. In both simple cases above, the EM algorithm works well.

Table 2
Simulation results concerning estimation of Pearson correlation and marginal regression parameters in the copula model for partially misaligned missing at random data obtained EM algorithm, compared with the gold standard with full data, Multiple Imputation and Hot-Deck Imputation.

Parameter	True value	Full data		Copula&EM		Multiple imputation		Hot-deck imputation	
		Estimate	std.err	Estimate	std.err	Estimate	std.err	Estimate	std.err
β_{10}	0	−0.0012	0.143	0.0596	0.1466/0.1488	−0.0012	0.143	−0.0012	0.143
β_{11}	1	1.0030	0.1424	0.9425	0.1451/0.1485	1.0030	0.1424	1.0030	0.1424
β_{12}	3	3.0003	0.0496	3.0000	0.0498/0.0528	3.0003	0.0496	3.0003	0.0496
σ_1	1	0.9975	0.051	1.0446	0.0566/0.0522	0.9975	0.051	0.9975	0.051
β_{20}	0	−0.0050	0.1435	−0.0087	0.189/0.2146	−0.0918	0.2134	−0.1405	0.2397
β_{21}	2	2.0049	0.1413	2.0098	0.1864/0.2157	2.1654	0.2053	2.1989	0.2104
β_{22}	2	2.0006	0.0514	2.0017	0.0648/0.0774	2.0015	0.0719	2.0020	0.0711
σ_2	1	0.9977	0.0493	1.0918	0.0883/0.0754	0.9836	0.0796	0.9398	0.1046
β_{30}	0	−0.0023	0.1474	0.1109	0.1915/0.1797	0.0503	0.2178	0.0424	0.2371
β_{31}	3	3.0046	0.1481	2.8902	0.1936/0.1776	2.9111	0.2155	2.9160	0.2191
β_{32}	1	1.0010	0.051	1.0005	0.0665/0.0634	0.9997	0.0745	1.0006	0.0737
σ_3	1	0.9965	0.0508	0.9379	0.0615/0.0622	0.9946	0.0752	0.9662	0.0956
γ	0.5	0.5019	0.0408	0.4951	0.0606/0.0557	0.4409	0.1212	0.4178	0.1149

4.2. Misaligned missing data

Motivated from one of our collaborative projects on a quality of life study (see the detail in Section 5), we consider a rather challenging missing data pattern in this simulation study. That concerns the so-called misaligned missingness, which refers to a situation where two correlated variables have missing values on exclusive subsets of subjects. In a completely misaligned missing case, where there is no overlap between two margins, Hot-Deck imputation fails to work, and the method of multiple imputation cannot effectively capture between-variable correlations, resulting in poor estimation of correlation parameters. However, when the correlation matrix is specified by a structured form in the Gaussian copula model, the EM algorithm is able to utilize the correlation structure for information sharing, and consequently the resulting estimation of model parameters is highly satisfactory.

The simulation setup is given as follows. Following Example-1 in Section 2.3, we include two covariates $X_1 \sim \text{Bin}(1, 0.5)$ and $X_2 \sim \Gamma(2, 1)$, and generate residuals ϵ in a linear model with $\mu_j = X^T\beta_j$, $j = 1, 2, 3$ from a tri-variate normal with the marginal $N(0, 1)$ and first-order autoregressive correlation matrix with parameter $\gamma = 0.5$.

The missing mechanism concerns missing at random (MAR) with a partially misaligned pattern with specified as follows. A tri-variate outcome $(Y_1, Y_2, Y_3)'$ is subject to be missing at random, where Y_1 is fully observed, while each of Y_2 and Y_3 has 45% missing data that are partially misaligned, with only 10% of subjects have an overlap on the observed parts of Y_2 and Y_3. The reason that a partial misalignment is considered here is to allow the Hot-Deck Imputation method possibly in the part of the comparison. The EM algorithm procedure and notations follow as discussed in Example-1, Section 2.3. The missing probability in the marginal of Y_2 is,

$$P(R_2 = 0|X_1, Y_2) = \begin{cases} 0.45, & \text{if } X_1 = 1; \\ 0.81, & \text{if } X_1 = 0, \text{ and } Y_2 > \mu_2; \\ 0.09, & \text{if } X_1 = 0, \text{ and } Y_2 < \mu_2. \end{cases}$$

The missing probability in the third marginal Y_3 is given by,

$$P(R_3 = 0|R_2) = \begin{cases} 0, & \text{if } R_2 = 0; \\ \dfrac{0.45}{1 - 0.45}, & \text{if } R_2 = 1. \end{cases} \tag{16}$$

We compare the results obtained from the EM algorithm with those from the gold standard using the full data, the multiple imputation and the Hot-Deck imputation. In addition, this comparison includes two types of standard errors: the first type is the empirical standard error in four methods, and the other type is the average of 1000 model-based standard errors obtained from Louis' formula discussed in Section 3.3, which is only provided in the EM algorithm (see Table 2).

5. Data example

Nephrotic Syndrome (NS) is a common disease in pediatric patients with kidney disease. The typical symptom of this disease is characterized by the presence of edema that significantly affects the health-related quality of life in children and adolescents. The PROMIS (Fries et al., 2005; Gipson et al., 2013) is a well-validated instrument to assess pediatric patient's quality of life. The instrument consists 7 domains, but here we only choose 3 domains with missing misalignment pattern for illustration. In the data, two QoL measures, pain and fatigue, are measured on two exclusive sets of subjects due to some logistic difficulty at the clinic; out of 226 subjects, 107 subjects have measurements of pain, but no measurements of fatigue, while the other 117 subjects have measurements of fatigue but no measurements of pain. In addition, two subjects

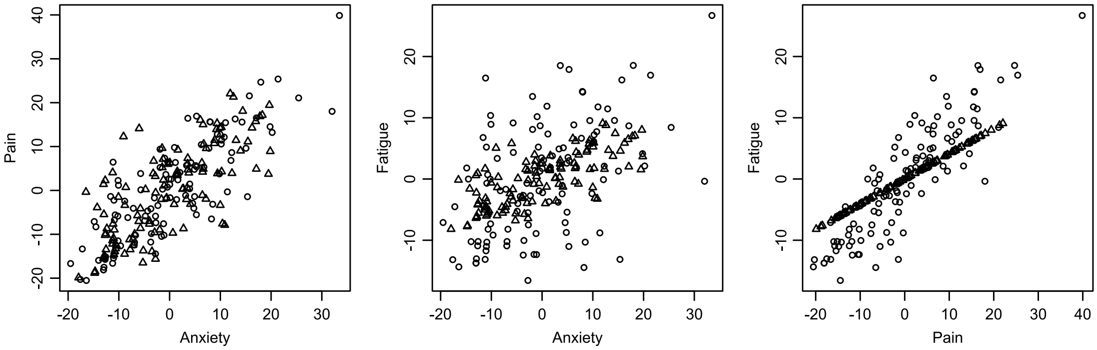

Fig. 1. Plots of observed and predicted residuals from the EM algorithm.

have neither measurements of pain nor measurements of fatigue. Interestingly, measurements of anxiety have been fully recorded on all 226 individuals with no missing data. In this case, Hot-Deck imputation does not work. We first apply the complete case univariate analysis of each QoL domain score (Y_1 = anxiety, Y_2 = pain, Y_3 = fatigue) on covariates of age, gender, edema, race (white, black, and other as reference), and estimate the linear correlation coefficient of the residuals as 0.6830 between anxiety and pain and 0.5106 between anxiety and fatigue, which turns out to be approximately the square of the correlation coefficient between anxiety and pain. This suggests us use first order autoregressive correlation for matrix Γ in the copula model.

The EM algorithm has two advantages to handle this misaligned missing data pattern. One is that we can estimate both marginal and correlations parameter adjusting for the confounders, where the information across the three QoL scores can be shared to improve efficiency. The other is the prediction of the missing QoL scores by using the correlated QoL scores together with the marginal regression models, which requires the availability of inverse correlation matrix, Γ^{-1}.

The observed data and predicted data from the EM algorithm are all shown in Fig. 1. The triangles indicate patients with missing fatigue data, and the circles correspond to patients with missing pain data. Between pain and fatigue QoL scores, outcomes have no overlap. The circles and triangles are well distributed and appear to lie in elliptical in the first two scatter plots. In the third plot, the reason that the predicted triangles appear a straight line is the use of AR-1 correlation matrix, and the shape of these points may change to another pattern when a different correlation structure is used.

In Table 3, the standard errors for the estimates obtained by the multiple imputation are calculated by the conventional method given by Little and Rubin (2002). Moreover, some findings in the results shown in Table 3 are noteworthy. First, the estimated rank-based correlations Kendall's τ and Spearman's ρ between anxiety and pain are, respectively, $\tau_{12} = 0.4805$ and $\rho_{12} = 0.6677$, between anxiety and fatigue are $\tau_{13} = 0.3110$ and $\rho_{13} = 0.4524$, and between pain and fatigue are $\tau_{23} = 0.4805$ and $\rho_{23} = 0.6677$. In addition, the estimated correlation parameter by the multiple imputation approach is clearly smaller than that obtained by the EM algorithm. This is because the key difference is that the EM algorithm makes use of the correlation structure to access the entire data, whereas the imputation method does not. Imputation methods are based on available observed information, but not on the correlation structure. Moreover, the EM algorithm provides a straightforward calculation of asymptotic standard error of the correlation parameter for inference; for example, p-value for $H_0 : \gamma = 0$ is of practical importance.

In addition, with regard to the effect of edema in pain, according to clinical information available on Mayo Clinic Website, pain is not regarded as one of key symptoms associated with edema. Both results obtained by the EM algorithm and the univariate analysis are in the agreement with this clinical information, indicating no significant effect of edema on pain score, while the multiple imputation method reports an opposite result.

6. Discussion

This paper presents a Gaussian copula framework that provides both marginal Pearson correlations, and marginal rank-based correlation in the presence of missing data. The EM-algorithm is developed and implemented to estimate both marginal parameters and correlation parameters. The proposed methodology allows to adjust for confounding factors via marginal regression models to obtain adjusted marginal correlation estimates, which are useful in practice. We propose a peeling procedure in the M-step to facilitate the computation of updating parameter values. In addition, missing values may also be updated as part of the EM-algorithm.

The EM algorithm outperforms imputation-based methods in two aspects. First, when the marginal outcomes are skewed, the classical Multiple Imputation method implemented under the multivariate normality does not work well, while the Hot-Deck Imputation approach works reasonably well. Second, when missing data patterns are fully or severely misaligned, as shown in our motivating example of the quality-of-life study, Hot-Deck Imputation approach does not work, and the multiple imputation cannot effectively utilize the correlation structure in data imputation and parameter estimation. When the correlation matrix is structured, EM algorithm can fully access the correlation structure in the Gaussian copula, and share information across different outcome variables, and therefore the resulting estimates from the EM algorithm are satisfactory.

Table 3
Estimation of correlation and marginal regression parameters in the copula model for quality of life study obtained by univariate analysis, EM algorithm and Multiple Imputation.

Outcome	Covariates	Univariate analysis		Copula&EM		Multiple imputation	
		estimate	std.err	estimate	std.err	estimate	std.err
Anxiety	Intercept	40.5406	4.1387	39.3566	4.4013	40.5406	4.1387
	Age	0.0068	0.2451	0.0879	0.2607	0.0068	0.2451
	Gender	1.8753	1.4392	1.8749	1.5306	1.8753	1.4392
	Edema	5.2453	2.1133	5.0332	2.2474	5.2453	2.1133
	White	0.4908	2.0272	0.8502	2.1558	0.4908	2.0272
	Black	4.7095	2.3491	4.5299	2.4981	4.7095	2.3491
	σ	10.33	[a]	10.9813	0.5165	10.2107	[a]
Pain	Intercept	41.3243	6.1947	36.9533	7.8865	45.1838	5.7978
	Age	0.0649	0.3575	0.3587	0.4551	−0.0029	0.3418
	Gender	0.3423	2.0304	0.1196	2.5849	0.3112	1.9516
	Edema	6.0597	3.3437	4.8594	4.2569	7.7273	3.2138
	White	1.2991	3.3331	2.8328	4.2434	−1.3024	3.1549
	Black	6.8171	3.7939	6.5658	4.8300	3.5309	3.5994
	σ	10.58	[a]	13.4695	0.8805	11.2709	[a]
Fatigue	Intercept	30.9033	5.5260	30.9030	4.6716	31.5289	4.6886
	Age	0.7043	0.3275	0.7043	0.2768	0.6351	0.2844
	Gender	0.6004	1.9962	0.6005	1.6876	0.7472	1.6722
	Edema	7.7774	2.6296	7.7774	2.2230	8.0216	2.2691
	White	3.6899	2.4766	3.6900	2.0937	3.2258	2.1573
	Black	7.1261	2.8784	7.1261	2.4334	7.0781	2.5138
	σ	9.568	[a]	8.0890	0.5504	9.2272	[a]
Correlation	γ	–	–	0.6851	0.0395	0.4001	–

[a] Unavailable in *R* function *lm* for linear regression.

Note that structured correlation (e.g., exchangeable) is seen in other families of copulas, such as Archimedean copulas, in which expansion of the EM algorithm with misaligned missing data is feasible and worth a further study.

It has been observed from both simulation studies and data analysis that the estimation of marginal parameters appears to be stable in the iterations of the EM algorithm. This may be due to (i) that the copula model provides a separable formulation for the first moment and the second moment, so that any misspecification of one model component has little effect on the other; and (ii) the initial values are given by consistent estimators of the marginal regression parameters, as suggested by Joe (2005), and such consistency is preserved when the bias in the second moment estimation is asymptotically ignorable. It is interesting to note that Segers et al. (2014) proposed a one-step estimation for correlation parameters in the Gaussian copula, which is shown to be efficient by a novel one-step adjustment, and this approach may be applied to improve the M-step of the EM algorithm to achieve estimation consistency.

In this paper we considered the EM algorithm for balanced data generated from the multivariate Gaussian copula model. We envision that the EM algorithm may be generalized to handle unbalanced data if the correlation matrix for different unbalanced data forms can be thought as of nested submatrices. This nesting property may be easily satisfied in some common correlation structures, such as exchangeable and AR-1 correlation structures that contain only one correlation parameter in submatrices of different dimensionality. In the general case of unstructured correlation matrix, unbalanced data may require a sophisticated model for correlations, and consequently the EM algorithm may become nontrivial to implement, and some alternative approaches may be worth a future exploration.

In this paper we focus on the Gaussian copula dependence model due largely to its mathematical convenience, such as its flexibility to handle an arbitrary dimension of multivariate outcomes and the separability between the marginal mean model and the copula dependence model. In effect, the EM algorithm developed in a parametric Gaussian copula framework may be sensitive to model misspecification. Copula selection is still an open problem in this field albeit some limited approaches available in the literature. This problem becomes more challenging when the dimension of variables is arbitrary, as implemented in this paper. This is because there are not many copulas available in the literature that allow to analyze data with arbitrary dimensions, and the nonparametric version may suffer the curse of dimensionality and typically require lot of data to achieve satisfactory estimation. The class of vine copulas provides many flexible formulation of copula model but its model selection is still computationally prohibited for data of relatively large dimension. Model diagnostics are required before to draw final conclusions. Several authors have proposed diagnostic methods, such as Masarotto and Varin et al. (2012); Joe (1997); Genest et al. (1995); Ané and Kharoubi (2003); Huang and Prokhorov (2014); Scaillet (2007), among others. However, how these diagnostic approaches may perform in the case of incomplete data remains unknown and is an interesting future work.

For the case of completely misaligned missingness, when the correlation matrix is unstructured, the correlation parameters are not fully identifiable. Manski (2003) introduced several approaches for partial identification problem, and Fan and Zhu (2009) developed a method to determine the bounds, within which the estimates of correlation parameters of a copula model are partially identified by a parameter set. Following the notation given in Fan and Zhu (2009), we consider

$\mu(x, y) = xy$ for the problem of covariance estimation. This function is super-modular because its cross-derivative is 1, and this function is symmetric and marginal variances are finite. Thus, according to Fan and Zhu (2009) theory, we can establish a partial identification range for the correlation parameter in the presence of misaligned missing data with the lower and upper bounds, denoted by $\gamma^L_{j_1,j_2}$ and $\gamma^U_{j_1,j_2}$. They are the lower and upper bounds of correlation parameter γ_{j_1,j_2} given by $\gamma^L_{j_1,j_2} = \left[\int_0^1 \left\{F_{j_1}^{-1}(u|\theta_{j_1})F_{j_2}^{-1}(1-u|\theta_{j_2})\right\} du - \mu_{j_1}\mu_{j_2}\right] / \sigma_{j_1}\sigma_{j_2}$, and $\gamma^U_{j_1,j_2} = \left[\int_0^1 \left\{F_{j_1}^{-1}(u|\theta_{j_1})F_{j_2}^{-1}(u|\theta_{j_2})\right\} du - \mu_{j_1}\mu_{j_2}\right] / \sigma_{j_1}\sigma_{j_2}$, where quantiles functions $F_{j_1}^{-1}$ and $F_{j_2}^{-1}$ may be estimated by available data of y_{j_1} and y_{j_2}. This direction of research is worth a thorough exploration.

Acknowledgment

This research is supported by an NSF grant (DMS # 1208939).

References

Abramowitz, M., Stegun, I.A., 1972. Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables, Vol. 55. Courier Dover Publications.

Acar, E.F., Genest, C., NešLehová, J., 2012. Beyond simplified pair-copula constructions. J. Multivariate Anal. 110, 74–90.

Andridge, R.R., Little, R.J., 2010. A review of hot deck imputation for survey non-response. Internat. Statist. Rev. 78 (1), 40–64.

Ané, T., Kharoubi, C., 2003. Dependence structure and risk measure*. J. Bus. 76 (3), 411–438.

Czado, C., 2010. Pair-copula constructions of multivariate copulas. In: Copula Theory and its Applications. Springer, pp. 93–109.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 1–38.

Fan, Y., Zhu, D., 2009. Partial identification and confidence sets for functionals of the joint distribution of potential outcomes. Tech. Rep., Working paper.

Fries, J., Bruce, B., Cella, D., 2005. The promise of promis: using item response theory to improve assessment of patient-reported outcomes. Clin. Exp. Rheumatol. 23 (5), S53.

Genest, C., Nešlehová, J., Ben Ghorbal, N., 2011. Estimators based on Kendall's tau in multivariate copula models. Aust. N. Z. J. Stat. 53 (2), 157–177.

Genest, C., Quesada Molina, J., Rodríguez Lallena, J., 1995. De l'impossibilité de construire des lois à marges multidimensionnelles données à partir de copules. C. R. Acad. Sci., Paris I 320 (6), 723–726.

Gipson, D.S., Selewski, D.T., Massengill, S.F., Wickman, L., Messer, K.L., Herreshoff, E., Bowers, C., Ferris, M.E., Mahan, J.D., Greenbaum, L.A., et al., 2013. Gaining the promis perspective from children with nephrotic syndrome: a midwest pediatric nephrology consortium study. Health Qual. Life Outcomes 11 (3).

Graham, J.W., Olchowski, A.E., Gilreath, T.D., 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev. Sci. 8 (3), 206–213.

Huang, W., Prokhorov, A., 2014. A goodness-of-fit test for copulas. Econometric Rev. 33, 751–771.

Joe, H., 1997. Multivariate Models and Multivariate Dependence Concepts, Vol. 73. CRC Press.

Joe, H., 2005. Asymptotic efficiency of the two-stage estimation method for copula-based models. J. Multivariate Anal. 94, 401–409.

Joe, H., Li, H., Nikoloulopoulos, A.K., 2010. Tail dependence functions and vine copulas. J. Multivariate Anal. 101 (1), 252–270.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.

Little, R.J., Rubin, D.B., 2002. In: Little, R.J.A., Rubin, D.B. (Eds.), Statistical Analysis with Missing Data, second ed. In: Wiley Series in Probability and Stistics, Wiley, New York, NY, p. 1. Statistical analysis with missing data, 2002.

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 226–233.

Manski, C.F., 2003. Partial Identification of Probability Distributions. Springer.

Masarotto, G., Varin, C., et al., 2012. Gaussian copula marginal regression. Electron. J. Stat. 6, 1517–1549.

McNeil, A.J., Frey, R., Embrechts, P., 2010. Quantitative Risk Management: Concepts, Techniques, and Tools. Princeton University Press.

Meng, X.-L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80 (2), 267–278.

Nelder, J.A., Baker, R., 1972. Generalized linear models. Encyclopedia Stat. Sci..

Qu, A., Lindsay, B.G., Li, B., 2000. Improving generalised estimating equations using quadratic inference functions. Biometrika 87 (4), 823–836.

Rubin, D.B., 2004. Multiple Imputation for Nonresponse in Surveys, Vol. 81. John Wiley & Sons.

Scaillet, O., 2007. Kernel based goodness-of-fit tests for copulas with fixed smoothing parameters. J. Multivariate Anal. 98, 533–543.

Segers, J., van den Akker, R., Werker, B.J., 2014. Semiparametric Gaussian copula models: Geometry and efficient rank-based estimation. Ann. Statist. 42, 1911–1940.

Sklar, M., 1959. Fonctions de Répartition à n Dimensions et Leurs Marges. Université Paris, p. 8.

Song, P.X.-K., 2000. Multivariate dispersion models generated from Gaussian copula. Scand. J. Statist. 27 (2), 305–320.

Song, P.X.-K., 2007. Correlated Data Analysis: Modeling, Analytics, and Applications. Springer.

Song, P.X.-K., Jiang, Z., Park, E., Qu, A., 2009. Quadratic inference functions in marginal models for longitudinal data. Stat. Med. 28 (29), 3683–3696.