Mixed and missing data: a unified treatment with latent graphical models

By Xiao Li, Jinzhu Jia and Yuan Yao

LMAM, School of Mathematical Sciences, and Center for Statistical Science,
Peking University, Peking, China 100871

seanli@pku.edu.cn,{jzjia,yuany}@math.pku.edu.cn

Summary

We propose to learn latent graphical models when data have mixed variables and missing values. This model could be used for further data analysis, including regression, classification, ranking etc. It also could be used for imputing missing values.

We specify a latent Gaussian model for the data, where the categorical variables are generated by discretizing an unobserved variable and the latent variables are multivariate Gaussian. The observed data consists of two parts: observed Gaussian variables and observed categorical variables, where the latter part is considered as partially missing Gaussian variables. We use the Expectation-Maximization algorithm to fit the model. To prevent overfitting we use sparse inverse covariance estimation to obtain sparse estimate of the latent covariance matrix, equivalently, the graphical model. The fitted model then could be used for problems including regression, classification and ranking. Such an approach is applied to a medical data set where our method outperforms the state-of-the-art methods. Simulation studies and real data results suggest that our proposed model performs better than random forest in terms of prediction error when the model is correctly specified, and is a better imputation method than hot deck imputation even if the model is not correctly specified.

Some key words: Graphical model; Expectation-Maximization algorithm; Latent Gaussian model; Missing data; Mixed data.

1. Introduction

The problem of missing data arises in many situations, especially in clinical trials and social investigations. In general settings, the explanatory variables contain both continuous variables and categorical variables (we refer this kind of data as mixed data). As such, a major difficulty of this kind of problems stems from modeling the data. It is often very hard to specify a unified model for the variables, let alone making inference and imputing missing values.

Let's look at the PCI (percutaneous coronary intervention) data provided by two hospitals in Beijing, China, which has been studied by Wang et al. (2015) from a clinical perspective. The data consists of 38 variables that are measured before one patient's heart operation and a binary variable indicating whether there is reflux after PCI. There are totally 2582 patients collected. The objective of the analysis is to determine which ones, among the 38 variables, are related to the outcome of the operation and to predict the outcome of the operation. 23 of the 39 variables are categorical variables with the levels ranging from two to six. In addition, all 39 variables have missing values and the missing proportion ranges from less than 1% to more than 70%. To analyze this data, it is crucial to propose appropriate methods for mixed and missing data.

In many cases including the PCI data describe above, it is reasonable to assume that the categorical variables and rank variables are formed by discretizing a latent variable (Skrondal & Rabe-Hesketh, 2007). For instance, in medical science, the latent variables can represent physiological conditions that exist in hidden form but are not directly measurable, and instead, they can be measured indirectly by some surrogate variables, like the outcome of an operation. Under the latent variable assumption, the latent Gaussian copula model was considered by Xue & Zou (2012) and Fan et al. (2014). This semiparametric model has many nice theoretical properties, but is not suitable for inference with missing data. The model reduces to the latent Gaussian model when the latent variables are multivariate Gaussian. Han & Pan (2012) use composite marginal likelihood function to estimate the latent parameters for the latent Gaussian model. To the best of our knowledge, there has not been any likelihood-based method for inference under the latent Gaussian model when there are missing values in the observed data.

We propose a likelihood based method along with the Expectation-Maximization algorithm to estimate the latent parameters. The fitted model is then used for classification. It also could be used for regression and ranking without any further difficulty than classification. The value of our proposed method is twofold. (1) It provides new tracks for data analysis with mixed data, whether missingness exists or not. Numerical results show that our classification model performs better than random forest in terms of prediction error, and that the latent Gaussian model is a better imputation model than hot deck imputation (Tanner, 1991). (2) Our method serves as a supplement to the estimation of Gaussian graphical model with mixed and missing data. For a review of related literature on estimation of Gaussian graphical models, see Friedman et al. (2008), Onureena et al. (2008), Cai et al. (2011),Xue et al. (2012), Lee & Hastie (2012), Fellinghauer et al. (2013) and Stadler & Buhlmann (2012). Simulations show that the inverse covariance matrix estimated by our method is close to the true covariance matrix even when missing values exist.

The rest of the paper is organized as follows. In Section 2 we introduce basic notations and formaly describe our model for mixed and missing data. In Section 3 we introduce the likelihood-based method to find the maximum likelihood estimation of parameters in our model. In Section 4, we discuss classification methods for mixed and missing data. Numerical results are given in section 5.

2. NOTATIONS AND MODEL DESCRIPTIONS

Let $X \in \mathbb{R}^{n \times p}$ be our collected data that consist both continuous and discrete variables. We view rows of X as independent and identically distributed observations. Each column of X corresponds to one random variable. We use X_j to denote one random variable and the jth column of X interchangeably. Suppose there are p_1 continuous variables and $p_2(= p - p_1)$ categorical variables. For a categorical variable X_j, let n_j be the levels of X_j. Without loss of generality, suppose that X_j takes value from $0, \ldots, n_j - 1$. For the ith row in X, denote $X_{i,con}$, $X_{i,cat}$ as the observations of the continuous and categorical variables respectively. In general we assume that X contains missing data that are missing at random. Denote X_{obs} as the collection of all observed data. Specifically, let $X_{i,con,obs}$, $X_{i,cat,obs}$, be the observed continuous and categorical variables of the ith observation.

Now we are ready to specify our model for mixed data.

DEFINITION 1 (LATENT GAUSSIAN MODEL FOR MIXED DATA). *Suppose that W is a p-dimensional Gaussian random vector with mean vector μ and covariance matrix Σ and V is a random vector of length p. We say that $V \sim LGM(\mu, \Sigma)$ if*

(1)for any continuous component j, $V_j = W_j$, and
(2)for any categorical component j, $\mu_j = 0$, $\sigma_{jj} = 1$ and

$$V_j = k \iff W_j \in (C_{j,k}, C_{j,k+1}),$$

where C_j is a fixed vector of thresholds of length $n_j + 1$ with n_j the level of categorical variable V_j.

Note that generally, $C_{j,0} = -\infty$ and $C_{j,n_j} = \infty$.

If $X \in \mathbb{R}^{n \times p}$ are random samples of $LGM(\mu, \Sigma)$, then from the definition, there exist n observations $Z_1, \ldots, Z_n$ from $N(\mu, \Sigma)$ such that (1) for any continuous component j, $X_{ij} = Z_{ij}$, and (2) for any categorical component j,

$$X_{ij} = k \iff Z_{ij} \in (C_{j,k}, C_{j,k+1}).$$

We call Z_{ij} the latent value of X_{ij}. In the next section, we propose estimation method for the latent Gaussian model. We emphasize on missing data problems, which is more difficult than parameter estimation without missing values in observed data X. Note that no matter for continuous variable or categorical variable (for example X_j), if one of the observation (for example X_{ij}) is missing the only information we have is that the latent value (Z_{ij}) lies in the interval $(-\infty, \infty)$.

3. PARAMETER ESTIMATION

Since there are a lot of parameters in this model. We use multi-stage method. First we estimate the collection of thresholds C to determine categorical variables. Then we use the Expectation-Maximization algorithm to estimate the latent mean and covariance matrix. For high-dimensional data, covariance matrix might be sparse, finally we use graphical Lasso and constrained L1-minimization to obtain the sparse covariance matrix.

3·1. *Estimating the thresholds*

The collection of thresholds C is estimated first independently as they will be needed in the estimation of μ and Σ.

We only use the jth column of X to compute the marginal likelihood of C_{jk}. It is not hard to show that the maximum marginal likelihood estimation is

$$\hat{C}_{jk} = \Phi^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} I(X_{tj} < k) \right), \text{ for } k = 0, 1, \ldots, n_j, \tag{1}$$

where Φ is the distribution function of the standard normal distribution. It is easy to check that $\hat{C}_{j0} = -\infty$ and $\hat{C}_{j,n_j} = \infty$. It is obvious that $\hat{C}_{j,k}$ is a consistent estimate of $C_{j,k}$.

3·2. *Estimating Latent Gaussian Parameters*

We use the Expectation-Maximization algorithm to estimate μ and Σ in the latent Gaussian model. To apply the Expectation-Maximization algorithm we first need to compute the expectation of complete log likelihood given observed data and current estimation. As we have noted in Section 2, the complete data here is actually Z (the latent value of X), and the observed data is $\{X_{i,con,obs}, X_{i,cat,obs}\}_{1 \leq i \leq n}$. Suppose that the current parameter estimation is $\mu^{(t)}, \Sigma^{(t)}$. Since Z_i's are independent and identically distributed samples from $N(\mu, \Sigma)$, the loglikelihood function of complete data is

$$\ell(\mu, \Sigma; Z) = C + \log(det(\Sigma)) + tr(S\Sigma^{-1}) + (\hat{\mu} - \mu)^T \Sigma^{-1} (\hat{\mu} - \mu), \tag{2}$$

where $\hat{\mu}$ and S is the sample mean and covariance matrix respectively. To compute the expected value of Equation (2) given observed data we need to compute

$$E(Z_{ij}Z_{ik}|X_{i,con,obs}, X_{i,cat,obs}; \mu^{(t)}, \Sigma^{(t)}), \tag{3}$$

for $1 \leq i \leq n$ and $1 \leq j, k \leq p$, and

$$E(Z_{ij}|X_{i,con,obs}, X_{i,cat,obs}; \mu^{(t)}, \Sigma^{(t)}), \tag{4}$$

for $1 \leq i \leq n$ and $1 \leq j \leq p$. In the E-step, to estimate term (3) and (4), we use the Gibbs sampling described in section 3·3. Once quantities (3) and (4) are obtained, maximizing $E(\ell(\mu, \Sigma; Z)|X_{obs})$ over μ and Σ is a standard procedure. The whole EM framework for mixed and missing data can now be stated as follows:

1. Start with random initial value $\mu^{(0)}, \Sigma^{(0)}$.
2. For $t = 1, \ldots, 1 \leq j, k \leq p$, let $\mu_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} E(Z_{ij}|X_{i,con,obs}, X_{i,cat,obs}; \mu^{(t)}, \Sigma^{(t)})$ and $\Sigma_{jk}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} E(Z_{ij}Z_{ik}|X_{i,con,obs}, X_{i,cat,obs}; \mu^{(t)}, \Sigma^{(t)}) - \mu_j^{(t+1)}\mu_k^{(t+1)}$.
3. Repeat (1) and (2) until convergence.

3·3. *E-step*

The previous subsection provides a framework for the Expectation-Maximization algorithm but did not show how to calculate the conditional expectations. We describe the calculations in detail in this subsection.

We use Gibbs sampling to calculate the (conditional) expectations used in the EM algorithm. Sampling from condition distributions given observed data in missing and mixed model could be viewed as sampling from a truncated multivariate normal distribution. This is because the complete data (provided that latent values are known) follows a joint multivariate normal distribution. For categorical observations, we know that latent values fall in some interval. Thus, the sampling could be view as sampling from the following conditional distribution:

$$f(W|a_j \leq W_j \leq b_j, j = 1, \ldots, p; \mu, \Sigma),$$

where W is a p-dimensional Gaussian random vector with mean vector μ and covariance matrix Σ; for missing values (no matter categorical or continuous), $a_j = -\infty$ and $b_j = \infty$; for continuous variables, $a_j = b_j$; for categorical variables, a_j and b_j are the thresholds that could be estimated via Equation (1) in Section 3·1.

Gibbs sampling from $f(W|a_j \leq W_j \leq b_j, j = 1, \ldots, p; \mu, \Sigma)$ could be done as follows. Starting from initial value $W^{(0)} = (W_1^{(0)}, \ldots W_p^{(0)})$, for $t = 1, \ldots, j = 1, \ldots, p$ we need to sample $W_j^{(t)}$ from the one-dimensional conditional density

$$f(W_j^{(t)}|W_1^{(t)}, \ldots, W_{j-1}^{(t)}, W_{j+1}^{(t-1)}, \ldots, W_p^{(t-1)}, a_j < W_j^{(t)} < b_j; \mu, \Sigma)$$
$$\propto \exp\left(-\frac{(W_j^{(t)} - \mu_j^{(t)})^2}{2(\sigma_j^{(t)})^2}\right) I(a_j < W_j^{(t)} < b_j), \tag{5}$$

where

$$\mu_j^{(t)} = E(W_j|W_1^{(t)}, \ldots, W_{j-1}^{(t)}, W_{j+1}^{(t-1)}, \mu, \Sigma), \tag{6}$$

and

$$(\sigma_j^{(t)})^2 = Var(W_j|W_1^{(t)}, \ldots, W_{j-1}^{(t)}, W_{j+1}^{(i-1)}, \mu, \Sigma). \tag{7}$$

Equations (6) and (7) could be easily obtained since W follows a joint multivatiate normal distribution with known parameters μ and Σ. We summarize the Gibbs sampling from $f(W|a_j \leq W_j \leq b_j, j = 1, \ldots, p; \mu, \Sigma)$ as in Algorithm 1. These samples will be used to calculate conditional expectations as in Equations (4) and (3).

Algorithm 1. Gibbs Sampling from truncated (conditional) normal distribution

---

Initialization. Set $W^{(0)} = (W_1^{(0)}, \ldots W_p^{(0)})$.
while not converged **do**
   Increase t by 1. t = t+1.
   for $j = 1, 2, \ldots, p$ **do**
      Compute $\mu = \mu_j^{(t)}$ and $\sigma = \sigma_j^{(t)}$ according to Equations (6) and (7) .
      Draw sample u from the uniform distribution on $[0, 1]$.
      Compute $d = \Phi(\frac{a_j - \mu}{\sigma})$ and $e = \Phi(\frac{b_j - \mu}{\sigma})$.
      Set $W_j^{(t)} = \mu + \sigma \Phi^{-1}(d + (e - d)u)$.
   end for
end while

We use the following ergodicity result to show the validity of the Gibbs sampler. It is a special case of Lemma 3 in Roberts & Smith (1994).

PROPOSITION 1 (ERGODICITY). *For* $W^{(i)} = \{W_1^{(i)}, \ldots, W_p^{(i)}\}$, $i = 1, \ldots, p$ *generated from Algorithm 1,*
(1) the density function of $W^{(i)}$ converges to $f(W|a_i < W_i < b_i, i = 1, \ldots, p; \mu, \Sigma)$;
(2) for any real-valued function g, if $E(g(W)|a_i < W_i < b_i, i = 1, \ldots, p; \mu, \Sigma) < \infty$, then as $N \to \infty$,

$$\frac{1}{N} \sum_{i=1}^{N} g(W^{(i)}) \to E(g(W)|a_i < W_i < b_i, i = 1, \ldots, p; \mu, \Sigma).$$

3·4. *Sparse inverse covariance selection*

The previous sections give a maximum likelihood estimation of parameters in the latent Gaussian model. But this is not enough, especially when the number of parameters is big. For such considerations, sparse inverse covariance selection is needed. Let $\hat{\mu}, \hat{\Sigma}$ be the maximum likelihood estimation of the mean and covariance matrix obtained by the Expectation-Maximization algorithm. Two methods could be used for sparse inverse covariance selection: graphical Lasso and constrained L1-minimization.

The Graphical lasso estimator is defined as

$$\Omega_G = \arg\min_{\Omega \geq 0} tr(S\Omega) - \log(det(\Omega)) + \lambda \sum_{i \neq j} |\Omega_{ij}| \tag{8}$$

where λ is the tuning parameter and S is the sample covariance matrix. For the latent Gaussian model, we replace S by $\hat{\Sigma}$ obtained via the likelihood-based method, and solve the following Graphical Lasso problem

$$\Omega_G = \arg\min_{\Omega \geq 0} tr(\hat{\Sigma}\Omega) - \log(det(\Omega)) + \lambda \sum_{i \neq j} |\Omega_{ij}|.$$

We refer the detail of Graphical Lasso and its algorithm to Friedman et al. (2008).

We can also use constrained L1-minimization proposed by Cai et al. (2011) to get a sparse estimation of Σ^{-1}. The constrained L1-minimization estimation is defined as

$$\Omega_C = \arg\min_{\Omega} ||\Omega||_1, \quad s.t. \quad ||\hat{\Sigma}\Omega - I||_{max} \leq \lambda \tag{9}$$

We refer the detail of this estimation and its related algorithm to Cai et al. (2011).

Using sparse inverse covariance selection (either Graphical Lasso or constrained L1-minimization), we obtain a family of covariance matrix Σ_λ indexed by the tuning parameter λ. In classification problem, λ is chosen such that the prediction error is minimized.

4. CLASSIFICATION USING LATENT GAUSSIAN MODEL

In this section we propose methods for classification using latent Gaussian model. We also point out that this model could be used for regression analysis and ranking problems. But for our purpose to deal with PCI data, we emphasize on classification problems.

For mixed data without missing value, our model can be used for classification. When there are missing values, our model can be used for both classification and imputation.

We first discuss classification problem for mixed data when there is no missing value. Suppose that the mth variable (a categorical variable) is used to indicate which class one data point belongs to. We fit a latent Gaussian model for all variables. For a specific λ, denote Σ_λ as the covariance matrix. For the ith observation in the test set, we compute

$$p_k = P(X_{im} = k|X_{i,\backslash m}; \mu, \Sigma_\lambda) = P(Z_{im} \in (C_{m,k}, C_{m,k+1})|X_{i,\backslash m}; \mu, \Sigma_\lambda)$$

via Gibbs sampling described in Section 3·3. Then X_{im} is assigned to category k that maximizes p_k. The tuning parameter λ with the smallest prediction error is chosen.

When missing values exists, the previous principle could also be applied. Denote by $\mu_\lambda, \Sigma_\lambda$ the parameters of the latent variables estimated by the training set, where λ is the tuning parameter in graphical Lasso or constrained L1-minimization method. For the ith observation in the test set, we compute

$$p_k = P(X_{im} = k|x_{i,obs\backslash m}; \mu, \Sigma_\lambda) = P(Z_{im} \in (C_{m,k}, C_{m,k})|x_{i,obs\backslash m}; \mu, \Sigma_\lambda)$$

via Gibbs sampling. Likewise, X_{im} is assigned to category k that maximizes p_k and the tuning parameter with the smallest prediction error is selected.

We could also use latent Gaussian model as an imputation model. That is, we impute the missing data in both the training set and the test set by sampling from latent Gaussian model. Then other traditional classification models, such as logistic regression and random forest, are used to do classification. We apply other methods instead of the latent Gaussian model (LGM) for classification purpose because the LGM might be misspecified, in which case, other methods could give better results.

5. NUMERICAL EVALUATION

In this section, we evaluate the performance of our classification method using both numerical experiments and real data.

5·1. *Simulation*

We set $n = 200, p = 50$. To generate the inverse covariance matrix Ω, we adopt the methods used by Meinshausen & Buhlmann (2006) and Liu et al. (2012). Set $\Omega_{jj} = 1, j = 1, 2, ..., n$ and

$\Omega_{jk} = cr_{jk}$, $j \neq k$, where c is a constant guaranteeing the positive definiteness of Ω, and r_jk are independent Bernoulli variables with $P(r_{jk} = 1) = p_0\phi(d_{jk}/\sqrt{p})$. ϕ is the density function of the standard normal distribution , $d_{jk} = ||z_j - z_k||_2$, where z_j are drawn from the uniform distribution on $[0, 1] \times [0, 1]$ and p_0 is a constant, controlling the sparsity of the true Ω. In our experiment, we set $c = 0.15$ and $p_0 = 1$. Once Ω is obtained, the covariance matrix $\Sigma = \Omega^{-1}$. We then rescale Σ by taking $\Sigma_{ij} = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$ so that each component of the Gaussian vector X_j (j =1,2,. . .,p) follows a standard normal distribution. For a categorical variable with k levels, the $k - 1$ non-trivial thresholds (the first and the last ones are $-\infty$ and $+\infty$ respectively) are drawn independently from the uniform distribution on [-1,1].

We consider the following four data generating scenarios. In each scenario, we first draw $W \in \mathbb{R}^{n \times p}$, with each row i.i.d. from $\sim N(0, \Sigma)$ and then cut a few continuous variable into categorical variables. Each point C_{jk} in the following settings is drawn independently from the uniform distribution on [-1,1].

1. Generate n observations with 49 continuous variables and one categorical (binary) variable. For $j = 1, 2 \ldots, 49$, let $X_j = W_j$ and $X_{50} = I(W_{50} > C_{50,1})$.
2. Generate n observations with 20 continuous variables and 30 categorical (binary) variables. For $j = 1, 2 \ldots, 20$, let $X_j = W_j$ and for $j = 21, 22, \ldots, 50$, $X_j = I(W_j > C_{j1})$.
3. Generate n observations with 20 Gaussian variables, 10 binary variables ,10 categorical variables with 3 levels and 10 categorical variables with 4 levels. Set $X_j = W_j$ for $j = 1, \ldots, 20$, $X_j = I(W_j > C_{j1})$ for $j = 21, ..., 30$, $X_j = I(W_j > C_{j1}) + I(Z_j > C_{j2})$ for $j = 31, ..., 40$, $V_j = I(W_j > C_{j1}) + I(W_j > C_{j2}) + I(W_j > C_{j3})$ for $j = 41, ..., 50$.
4. Generate n observations with categorical random variables only, of which 20 are binary variables ,10 are variables with 3 levels and 20 are variables with 4 levels. Set $X_j = I(W_j > C_{j1})$ for $j = 1, ..., 20$, $X_j = I(W_j > C_{j1}) + I(W_j > C_{j2})$ for $j = 21, ..., 30$, $V_j = I(W_j > C_{j2}) + I(W_j > C_{j2}) + I(W_j > C_{j3})$ for $j = 31, ..., 50$.

Now we generate missing values in each of the four settings. We use a missing at random mechanism. Ten variables are selected at random. For each observation and each selected variable, we decide whether the value is missing or not. We use $Rest$ to denote the other 40 variables that are observed completely. We use R_{ij} to denote the indicator of missingness, i.e. $R_{ij} = 1$ if X_{ij} is missing, and $R_{ij} = 0$ otherwise. We model R_{ij} using a logistic regression. The probability that X_{ij} is missing is given by

$$P(R_{ij} = 1) = \frac{\exp\{-1 + \beta \sum_{k \in Rest} x_{ik}\}}{1 + \exp\{-1 + \beta \sum_{k \in Rest} x_{ik}\}},$$

where the coefficient β is tuned such that the missing proportion for each variable is around 40%. This trick has been used in Garcia et al. (2010).

We use the EM algorithm described in Section 3·2 to find matrix Σ. Then we use graphical Lasso and constrained L1-minimization method to estimate sparse concentration matrix Ω .

In Table 1, we list the Frobenius norm and spectral norm of the matrix $\hat{\Omega} - \Omega$ in different scenarios. The tuning parameter is chosen to be the oracle parameter which minimizes the two norms mentioned above.

It can be seen from Table 1 that the loss decreases as the number of category increases. For example, all 30 categorical variables in scenario 2 are binary variables whereas 20 categorical variables in scenario 3 have more than 2 categories, and the estimation of Ω is better in scenario 3. This observation matches our intuition since a categorical variable with more levels approximates the latent variable better than one with less levels.

		Graphical Lasso	constrained L1-minimization
(1)	F	3.20(3.43)	3.21(3.47)
	S	1.20(1.27)	1.30(1.37)
(2)	F	3.96(4.23)	4.10(4.35)
	S	2.21(2.27)	2.25(2.31)
(3)	F	3.79(4.04)	3.86(4.10)
	S	2.09(2.17)	2.17(2.25)
(4)	F	3.92(4.19)	4.05(4.28)
	S	2.16(2.24)	2.23(2.31)

Table 1. *The Frobenius(F) and Spectral(S) norm of $\hat{\Omega} - \Omega$ in different data generating scenarios. Given in parenthesis are the corresponding norms when X has missing values.*

Now we examine our model's performance for classification. For complete data sets, we compare the prediction accuracy of our model with random forest, which is known to be the state-of-art method for classification problems. Let the outcome variable be the binary variable with the smallest absolute threshold value. We generate 100 data sets in each scenario and use 5-fold cross validation to compute the prediction error. The result is shown in Table 2.

	Latent Gaussian model	Random forest
(1)	32.4%(7.7%)	34.4%(6.3%)
(2)	35.9%(8.2%)	37.2%(10.2%)
(3)	34.0%(8.1%)	36.9%(7.5%)
(4)	39.3%(7.2%)	38.5%(7.0%)

Table 2. *Prediction error for different models with complete data. Standard deviations are given in parenthesis.*

For data sets with missingness, our model can be used for both classification and imputation. The following three methods will be used here fro comparasion.

1. Use latent Gaussian model as classification model.
2. Use latent Gaussian model as imputation model, and random forest as classification model. The tuning parameter with the smallest prediction error is selected and the corresponding prediction error is reported.
3. Use hot deck as imputation model, and random forest as classification model. This method is included as a benchmark for the prediction error.

We generate 100 data sets under each scenario and use 5-fold cross-validation to estimate the prediction error. For 2 and 3, we impute the missing data 100 times and calculate the mean (but not the variance) of the prediction error. The result is shown in Table 3.

It can be seen that our LGM model performs better than random Forest in most cases. Only when all variables are categorical (case 4 in Table 3), LGM performs slightly worse. The result matches our intuition since when the proportion of continuous variables are higher, data contains more information on μ and Σ. In the case of is missing data, the latent Gaussian model also proves to be a better imputation model than naive imputation methods.

5·2. *Analysis of PCI (percutaneous coronary intervention) data*

In this section, we consider the PCI (percutaneous coronary intervention) data provided by two hospitals in Beijing. The data consists of 2582 observations. For each observation, there are

	Latent Gaussian model	Latent Gaussian model+Random forest	Hot deck+Random forest
(1)	33.1%(8.5%)	35.1%(10.5%)	35.9%(9.7%)
(2)	36.6%(9.1%)	37.6%(11.5%)	37.7%(11.2%)
(3)	35.3%(8.8%)	37.8%(11.0%)	37.4%(10.6%)
(4)	39.9%(7.9%)	39.1%(10.7%)	39.7%(9.9%)

Table 3. *Prediction error for different models with missing data. Standard deviations(calculated across data sets)are given in parenthesis.*

22 categorical variables and 16 continuous variables that are measured before the operation and a binary variable indicating whether there is reflux after PCI. All 39 variables have missing data with missing proportion ranging from less than 1% to over 70%. We use our proposed method to determine the graphical model of the variables and to predict the outcome of the operation given the 38 variables measured before the operation.

First, we fit a latent Gaussian model with all observations using the Expectation-Maximazation algorithm, as described in Section 3·2. Then we use graphical Lasso to obtain a family of sparse inverse covariance matrix indexed by the tuning parameter λ. In Table 4, we listed the number of nonzero elements in the last column of the inverse covariance matrix and the new variable entering Ω under different λ. We know that when λ is big, all elements of $\hat{\Omega}$ will be zeros. When we decrease λ, some variable will enter into the estimation. We monitor the last column of $\hat{\Omega}$. Since the nonzero elements in this column mean that these variables have prediction power for the outcome. The first three non-zero entries that enter into the last column of the inverse covariance matrix correspond to the variables hypertension history, blood glucose and AP(angina pectoris). The association of these variables with the outcome of PCI was revealed both from a medical perspective and by other statistical methods (such as variable selection of logistic regression), as in Madani et al. (2013) and Dumont et al. (2006). As an illustration of the latent sparse inverse covariance matrix, we plot in Figure 1 the latent Gaussian graphical model with the tuning parameter that minimizes the test prediction error.

λ	Number of nonzero elements	New variable
0.13	1	Hypertension history
0.12	2	Blood glucose
0.10	3	AP(angina pectoris)
0.09	4	CRP
0.08	5	Admitting diagnosis

Table 4. *The number of nonzero elements in the last column of the inverse covariance matrix and the new variable entering the last column of Ω with decreasing λ.*

We now turn to predict the binary outcome of the operation. The same methods 1-3 (1. pure latent gaussian model; 2. latent gaussian model + random forest/logistic; 3. hot deck imputation + random forest/logistic) as in section 5·1 are applied, but random forest and logistic regression are both used here for complete data inference (i.e. inference after imputation). Similarly, we use 5-fold cross-validation to estimate the prediction error. For 2 and 3, we impute the missing data 100 times and calculate the mean of the prediction error. The tuning parameter with the smallest prediction error is selected and the corresponding prediction error is reported. The result is presented in Table 5. Here the latent Gaussian model is potentially (and almost certainly) misspecified. Nevertheless, we get the smallest prediction error when we use latent Gaussian model as the imputation model and random forest as the classification model.
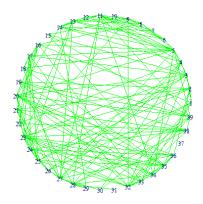
Fig. 1. The graphical model with the tuning parameter that minimizes the prediction error. The 39th variable is the binary response.

Method	Prediction Error
Latent Gaussian model	16.2%
Latent Gaussian model+Random forest	13.0%
Hot deck+Random forest	13.9%
Latent Gaussian model+logistic	20.9%
Hot deck+logistic	21.0%

Table 5. *Prediction error of different methods for the PCI data.*

REFERENCES

CAI, T., LIU, W. & LUO, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

DUMONT, C. J., KEELING, A. W., BOURGUIGNON, C., SAREMBOCK, I. J. & TURNER, M. (2006). Predictors of vascular complications post diagnostic cardiac catheterization and percutaneous coronary interventions. *Dimensions of Critical Care Nursing* **25**, 137–142.

FAN, J., LIU, H., NING, Y. & ZOU, H. (2014). High dimensional semi- parametric latent graphical model for mixed data. *arXiv preprint arXiv:1404.7236* .

FELLINGHAUER, B., BUHLMANN, P., RYFFEL, M., VON RHEIN, M. & REINHARDT, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics and Data Analysis* **64**, 132–152.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Computational statistics and data analysis. *Biostatistics* **9**, 432–441.

GARCIA, R. I., IBRAHIM, J. G. & ZHU, H. (2010). Variable selection for regression models with missing data. *Statistica Sinica* **20**, 149.

HAN, F. & PAN, W. (2012). A composite likelihood approach to latent multivariate gaussian modeling of snp data with application to genetic association testing. *Biometrics* **68**, 307–315.

LEE, J. & HASTIE, T. (2012). Learning mixed graphical models. *arXiv preprint arXiv:1205.5012* .

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. & WASSERMAN, L. (2012). High dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* **40**, 2293–2326.

MADANI, M., ALIZADEH, K., GHAZAEE, S. P., ZAVAREHEE, A., ABDI, S., SHAKERIAN, F., SALEHI, N. & FIROUZI, A. (2013). Elective percutaneous coronary intervention: the relationship between preprocedural blood glucose levels and periprocedural myocardial injury. *Texas Heart Institute Journal* **40**, 410.

MEINSHAUSEN, N. & BUHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.

ONUREENA, B., EL GHAOUI, L. & DASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research* **9**, 485–516.

ROBERTS, G. O. & SMITH, A. F. (1994). Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications* **49**, 207–216.

SKRONDAL, A. & RABE-HESKETH, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics* **34**, 712–745.

STADLER, N. & BUHLMANN, P. (2012). Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing* **22**, 219–235.

TANNER, M. A. (1991). *Tools for statistical inference*, vol. 3. Springer.

WANG, J.-W., ZHOU, Z.-Q., CHEN, Y.-D., WANG, C.-H. & ZHU, X.-L. (2015). A risk score for no reflow in patients with st-segment elevation myocardial infarction after primary percutaneous coronary intervention. *Clinical cardiology* **38**, 208–215.

XUE, L. & ZOU, H. (2012). Regularized rank-based estimation of high- dimensional nonparanormal graphical models. *The Annals of Statistics* **40**, 2541–2571.

XUE, L., ZOU, H. & CAI, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics* **40**, 1403–1429.