Published in final edited form as:

J R Stat Soc Series B Stat Methodol. 2018 November ; 80(5): 899–926. doi:10.1111/rssb.12279.

An imputation–regularized optimization algorithm for high dimensional missing data problems and beyond

Faming Liang,

Purdue University, West Lafayette, USA

Bochao Jia,

University of Florida, Gainesville, USA

Jingnan Xue,

Houzz, Palo Alto, USA

Qizhai Li, and

Chinese Academy of Sciences, Beijing, People's Republic of China

Ye Luo

University of Florida, Gainesville, USA

Summary.

Missing data are frequently encountered in high dimensional problems, but they are usually difficult to deal with by using standard algorithms, such as the expectation-maximization algorithm and its variants. To tackle this difficulty, some problem-specific algorithms have been developed in the literature, but there still lacks a general algorithm. This work is to fill the gap: we propose a general algorithm for high dimensional missing data problems. The algorithm works by iterating between an imputation step and a regularized optimization step. At the imputation step, the missing data are imputed conditionally on the observed data and the current estimates of parameters and, at the regularized optimization step, a consistent estimate is found via the regularization approach for the minimizer of a Kullback–Leibler divergence defined on the pseudocomplete data. For high dimensional problems, the consistent estimate can be found under sparsity constraints. The consistency of the averaged estimate for the true parameter can be established under quite general conditions. The algorithm is illustrated by using high dimensional Gaussian graphical models, high dimensional variable selection and a random-coefficient model.

Keywords

Expectation-maximization algorithm; Gaussian graphical model; Gibbs sampler; Imputation consistency; Random-coefficient model; Variable selection

Address for correspondence: Faming Liang, Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907, USA., fmliang@purdue.edu.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplementary material for “An imputation-consistency algorithm for high-dimensional missing data problems and beyond”’.

1. Introduction

Missing data are frequently encountered in both low and high dimensional data, where low and high refer to whether the number of variables is smaller or larger than the sample size respectively. For example, microarray data are usually considered as high dimensional, where the number of genes can be much larger than the number of samples. Missing values can appear in microarray data for various reasons such as scratches on slides, spotting problems and experimental errors. In some microarray experiments, missing values can occur for more than 90% of the genes (Ouyang *et al.*, 2004). Simply deleting the samples or genes for which missing values occur can lead to a significant loss of data information. How to deal with missing data has been a long-standing problem in statistics.

For low dimensional problems, the missing data can be dealt with by using the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) or its variants. Let

Let $\mathbf{X}^{\text{obs}} = (X_1^{\text{obs}}, X_2^{\text{obs}}, \dots, X_n^{\text{obs}})$ denote the observed incomplete data, where n denotes the sample size. Let $\mathbf{X}^{\text{mis}} = (X_1^{\text{mis}}, X_2^{\text{mis}}, \dots, X_n^{\text{mis}})$ denote the missing data, and let $\mathbf{X} = (\mathbf{X}^{\text{obs}}, \mathbf{X}^{\text{mis}})$ denote the complete data. Let $\boldsymbol{\theta}$ denote the vector of unknown parameters, and let $f(\mathbf{X}|\boldsymbol{\theta})$ denote the likelihood function of the complete data. Then the maximum likelihood estimate of $\boldsymbol{\theta}$ can be determined by maximizing the marginal likelihood of the observed data,

$$f(\mathbf{X}^{\text{obs}}|\boldsymbol{\theta}) = \int f(\mathbf{X}^{\text{obs}}, \mathbf{x}^{\text{mis}}|\boldsymbol{\theta})h(\mathbf{x}^{\text{mis}}|\boldsymbol{\theta}, \mathbf{X}^{\text{obs}})d\mathbf{x}^{\text{mis}}, \quad (1)$$

where $h(\mathbf{x}^{\text{mis}}|\boldsymbol{\theta}, \mathbf{X}^{\text{obs}})$ denotes the predictive density of the missing data. The EM algorithm seeks to maximize equation (1) by iterating between the following two steps.

- a. *E-step*: calculate the expected value of the log-likelihood function with respect to the predictive distribution of the missing data given the current estimate $\boldsymbol{\theta}^{(t)}$, i.e.

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \int \log\{f(\mathbf{X}^{\text{obs}}, \mathbf{x}^{\text{mis}}|\boldsymbol{\theta})\}h(\mathbf{x}^{\text{mis}}|\boldsymbol{\theta}^{(t)}, \mathbf{X}^{\text{obs}})d\mathbf{x}^{\text{mis}}.$$

- b. *M-step*: find a value of $\boldsymbol{\theta}$ that maximizes the quantity $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, i.e. set

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

Dempster *et al.* (1977) showed that the marginal likelihood value increases with each iteration and, under fairly general conditions, it converges to a local or global maximum of the marginal likelihood. A rigorous study for the convergence is given in Wu (1983).

Both the E- and the M-steps of the algorithm can be quite complicated or even intractable. Meng and Rubin (1993) found that, in many cases, the M-step is relatively simple when conditioned on some function of the parameters under estimation. Motivated by this observation, they introduced the expectation-conditional maximization algorithm, which

replaces the M- step by a number of computationally simpler conditional maximization steps. Later, the EM algorithm was further accelerated by some other variants, such as the ‘expectation-conditional maximization either’ algorithm (Liu and Rubin, 1994; He and Liu, 2012) and the parameter expansion EM algorithm (Liu *et al.*, 1998). When the E-step is analytically intractable, Wei and Tanner (1990) introduced the Monte Carlo EM algorithm, which simulates multiple missing values from the predictive distribution $h(\mathbf{x}^{\text{mis}}|\boldsymbol{\theta}^{(t)}, \mathbf{X}^{\text{obs}})$ at the $(t + 1)$ th iteration, and then maximizes the approximate conditional expectation of the complete-data log-likelihood function

$$\hat{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \frac{1}{m} \sum_{j=1}^m \log \{f(\mathbf{X}^{\text{obs}}, \mathbf{x}_j^{\text{mis}}|\boldsymbol{\theta})\},$$

where $\mathbf{x}_1^{\text{mis}}, \dots, \mathbf{x}_m^{\text{mis}}$ denote the missing values simulated from $h(\mathbf{x}^{\text{mis}}|\boldsymbol{\theta}^{(t)}, \mathbf{X}^{\text{obs}})$. When the dimension of $\mathbf{X}^{\text{mis}}$ is high, the Monte Carlo approximation can be quite expensive. An alternative algorithm to deal with the intractable E-step is the stochastic EM (SEM) algorithm (Celeux and Diebolt, 1985). In this algorithm, the E-step is replaced by an imputation step, where the missing data are imputed with plausible values conditioned on the observed data and the current parameter estimate. At the M-step, the parameters are estimated by maximizing the likelihood function of the pseudocomplete data. Unlike the deterministic EM algorithm, the imputation step and M-step of the SEM algorithm generate a Markov chain which converges to a stationary distribution whose mean is close to the maximum likelihood estimate and variance reflects the loss of information due to the missing data (Nielsen, 2000).

Although EM and its variants work well for low dimensional problems (see McLachlan and Krishnan (2008) for an overview), essentially they fail for high dimensional problems. For high dimensional problems the maximum likelihood estimate can be non-unique or inconsistent. To address this issue, some problem-specific algorithms have been proposed; see for example, misgLasso (Stadler and Bühlmann, 2012), misPALasso (Städler *et al.*, 2014) and matrix completion algorithms (Cai *et al.*, 2010; Mazumder *et al.*, 2010). misgLasso was specifically designed for estimating Gaussian graphical models (GGMs) in the presence of missing data. Similarly to misgLasso, misPALasso also deals with multivariate Gaussian data in the presence of missing data. The matrix completion algorithm deals with large incomplete matrices, to learn a low rank approximation for a large-scale matrix with missing entries. However, lacking still is a general algorithm for high dimensional missing data problems.

This work fills the gap: we propose a general algorithm for dealing with high dimensional missing data problems. The algorithm proposed consists of two steps: an imputation step and a regularized optimization (RO) step, and is therefore called an imputation-regularized optimization (IRO) algorithm. The imputation step imputes the missing data with plausible values conditioned on the observed data and the current estimate of the parameters. The RO step finds a consistent estimate for the minimizer of the Kullback–Leibler divergence defined on the pseudocomplete data. For high dimensional problems, a certain type of regularization must be imposed on the parameter space for finding a consistent estimate.

Like the SEM algorithm, the IRO algorithm generates a Markov chain which converges to a stationary distribution. Under mild conditions, we show that the mean of the stationary distribution converges to the true value of the parameters in probability as the sample size becomes large. For low dimensional problems, the SEM algorithm can be viewed as a special case of the IRO algorithm. The IRO algorithm has strong implications for big data computing: on the basis of it, we propose a general strategy to improve Bayesian computation for big data problems. The IRO algorithm also facilitates data integration from multiple sources, which plays an important role in big data analysis. An R package accompanying this paper has been distributed to the public via the Comprehensive R Archive Network and is available from <https://cran.r-project.org/web/packages/IROmiss/index.html>.

The remainder of this paper is organized as follows. Section 2 describes the IRO algorithm with the theoretical development deferred to Appendix A. Section 3 applies the IRO algorithm to high dimensional GGMs. Section 4 applies the IRO algorithm to high dimensional variable selection. Section 5 applies the IRO algorithm to a random-coefficient model and discusses its potential use for big data problems. Section 6 concludes the paper with a brief discussion.

2. The imputation–regularized optimization algorithm

2.1. The algorithm

Let $X_1, \dots, X_n$ denote a random sample drawn from the distribution $f(x|\theta)$ (also denoted by $f_\theta(x)$ depending on convenience), where θ is a vector of parameters. Let $X_i = (X_i^{\text{obs}}, X_i^{\text{mis}})$, $i = 1, \dots, n$, where X_i^{obs} is observed and X_i^{mis} is missed. Let $\mathbf{X} = (X_1, \dots, X_n)$, $\mathbf{X}^{\text{obs}} = (X_1^{\text{obs}}, \dots, X_n^{\text{obs}})$ and $\mathbf{X}^{\text{mis}} = (X_1^{\text{mis}}, \dots, X_n^{\text{mis}})$. To indicate the dependence of the dimension of θ on the sample size n , we also write θ as θ_n and denote by $\theta_n^{(t)}$ the estimate of θ that is obtained at iteration t of the IRO algorithm. The IRO algorithm works by starting with an initial guess $\theta_n^{(0)}$ and then iterates between the imputation I- and RO steps.

- a. *I-step*: draw $\tilde{\mathbf{X}}^{\text{mis}}$ from the predictive distribution $h(\mathbf{x}^{\text{mis}}|\mathbf{X}^{\text{obs}}, \theta_n^{(t)})$ given $\mathbf{X}^{\text{obs}}$ and the current estimate $\theta_n^{(t)}$.
- b. *RO step*: on the basis of pseudocomplete data $\tilde{\mathbf{X}} = (\mathbf{X}^{\text{obs}}, \tilde{\mathbf{X}}^{\text{mis}})$, find an updated estimate $\theta_n^{(t+1)}$ which forms a consistent estimate of

$$\theta_*^{(t)} = \arg \max_{\theta} E_{\theta_n^{(t)}}[\log\{f_\theta(\tilde{\mathbf{X}})\}], \quad (2)$$

where

$$E_{\theta_n^{(t)}}[\log\{f_\theta(\tilde{\mathbf{X}})\}] = \int \log\{f(\mathbf{x}^{\text{obs}}, \tilde{\mathbf{x}}^{\text{mis}}|\theta)\} f(\mathbf{x}^{\text{obs}}|\theta_*) h(\tilde{\mathbf{x}}^{\text{mis}}|\mathbf{x}^{\text{obs}}, \theta_n^{(t)}) d\mathbf{x}^{\text{obs}} d\tilde{\mathbf{x}}^{\text{mis}}, \theta_*$$

denotes the true value of the parameters and $f(\mathbf{x}^{\text{obs}}|\boldsymbol{\theta}^*)$ denotes the marginal density function of $\mathbf{x}^{\text{obs}}$.

The RO step finds the minimizer of the Kullback–Leibler divergence from $f(\mathbf{x}^{\text{obs}}, \tilde{\mathbf{x}}^{\text{mis}}|\boldsymbol{\theta})$ to the joint density $f(\mathbf{x}^{\text{obs}}|\boldsymbol{\theta}^*)h(\tilde{\mathbf{x}}^{\text{mis}}|\mathbf{x}^{\text{obs}}, \boldsymbol{\theta}_n^{(t)})$. Intuitively, each RO step provides a drift to move the current estimate $\boldsymbol{\theta}_n^{(t)}$ towards $\boldsymbol{\theta}^*$, and the convergence will eventually happen by assuming that $n \rightarrow \infty$. For the case with a finite value of n , $\boldsymbol{\theta}_n^{(t)}$ will jump around $\boldsymbol{\theta}^*$ after convergence because of the randomness that is introduced in imputation. To find a consistent estimate of $\boldsymbol{\theta}_*^{(t)}$, an ideal objective function is $E_{\boldsymbol{\theta}_n^{(t)}}[\log\{f_{\boldsymbol{\theta}}(\tilde{\mathbf{x}})\}]$, but this cannot be directly

used as $\boldsymbol{\theta}^*$ is unknown. In practice, we can estimate the ideal objective function by the empirical mean of the log-likelihood function, i.e. $\sum_{i=1}^n \log\{f(x_i^{\text{obs}}, \tilde{x}_i^{\text{mis}}|\boldsymbol{\theta})\}/n$. Further, to accommodate the high dimensional case, for which the number of parameters can be much greater than the sample size, we impose a regularization term on the parameters, i.e. we propose to estimate $\boldsymbol{\theta}_*^{(t)}$ by

$$\boldsymbol{\theta}_{n,p}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} \left[\frac{1}{n} \sum_{i=1}^n \log\{f(x_i^{\text{obs}}, \tilde{x}_i^{\text{mis}}|\boldsymbol{\theta})\} - P_{\lambda}(\boldsymbol{\theta}) \right], \quad (3)$$

where the $\tilde{x}_i^{\text{mis}}$ s are drawn from $h(\tilde{x}^{\text{mis}}|x_i^{\text{obs}}, \boldsymbol{\theta}_n^{(t)})$, $P_{\lambda}(\boldsymbol{\theta})$ denotes the regularization or penalty function and λ denotes the regularization parameter. For this reason, we call the algorithm proposed an IRO algorithm.

Under mild conditions, theorem 1 in Appendix A establishes a uniform law of large numbers for the empirical mean of the log-likelihood function, i.e., as $n \rightarrow \infty$,

$\sum_{i=1}^n \log\{f(x_i^{\text{obs}}, \tilde{x}_i^{\text{mis}}|\boldsymbol{\theta})\}/n$ will converge to $E_{\boldsymbol{\theta}_n^{(t)}}[\log\{f_{\boldsymbol{\theta}}(\tilde{\mathbf{x}})\}]$ uniformly over the space of $\boldsymbol{\theta}$

and the space of $\boldsymbol{\theta}_n^{(t)}$. With appropriate regularization functions and regularity conditions for $E_{\boldsymbol{\theta}_n^{(t)}}[\log\{f_{\boldsymbol{\theta}}(\tilde{\mathbf{x}})\}]$, corollary 1 in Appendix A shows that $\boldsymbol{\theta}_{n,p}^{(t+1)}$ is uniformly consistent for $\boldsymbol{\theta}_*^{(t)}$

for all $t = 1, 2, \dots, T$, where T denotes the total number of iterations. Further, corollary 2 in Appendix A shows that, other than $\boldsymbol{\theta}_{n,p}^{(t+1)}$ s, any sequence of consistent estimates $\boldsymbol{\theta}_{n,g}^{(t+1)}$ can be used in the IRO algorithm, as long as $\boldsymbol{\theta}_{n,g}^{(t+1)}$ is sufficiently accurate for each t pointwise and the log-likelihood function of the pseudocomplete data satisfies certain regularity conditions. For obtaining such consistent estimates, different methods can be used, which might not explicitly maximize the objective function defined in equation (3). However, following from Jensen's inequality, they do implicitly and asymptotically maximize the objective function that is defined in equation (3). Corollary 2 greatly facilitates the use of the IRO algorithm.

Like the SEM algorithm, by simulating new independent missing values at each iteration, the sequence of estimates $\{\theta_n^{(t)}\}$ forms a time homogeneous Markov chain. The imputed values at different iterations also form a Markov chain. The two Markov chains are interleaved and share many properties, such as irreducibility, aperiodicity and ergodicity. Refer to Nielsen (2000) for more discussion on this issue. Theorem 3 and theorem 4 in Appendix A prove that the Markov chain $\{\theta_n^{(t)}\}$ has a stationary distribution and, furthermore, the mean of the stationary distribution forms a consistent estimate of θ^* .

We note that many of the assumptions that we made for proving the convergence and consistency of the IRO algorithm are quite regular. For example, we assumed that $\log\{f_{\theta}(\tilde{x})\}$ is a continuous function of θ for each $\tilde{x} \in \mathcal{X}$ and a measurable function of $\tilde{x}$ for each θ . Since we aim to address the missing data issue for a wide range of problems and it is difficult to specify the structure for each problem, we incorporate the assumptions about the parameters and structure of the problem in a metric entropy condition; see condition 2 in Appendix A. As discussed in remark 1 of Appendix A, this condition allows p to grow with n at a polynomial rate $O(n^\gamma)$ for some constant $0 < \gamma < \infty$ and allows the number of non-zero elements in θ to grow with n at a rate of $O(n^\alpha)$ for some $0 < \alpha < \frac{1}{2}$. These rates seem a little more restrictive than the exponential rate, i.e. $\log(p) = n^b$ for some constant $0 < b < 1$, conventionally sought in high dimensional regression. In the literature (see for example Raskutti *et al.* (2011)), the metric entropy condition has been used in studying the rate of convergence of the estimation for high dimensional regression.

Regarding conditions on missing data, we note that the IRO algorithm essentially works with any missing data mechanism, as long as the predictive distribution $h(\mathbf{x}^{\text{mis}} | \mathbf{X}^{\text{obs}}, \theta_n^{(t)})$ is available, well behaved and unchanged with the sample size n . Our current theory rules out the case that the missing data mechanism changes as the sample size increases, e.g. the missingness rate increases because of increased wear and tear on measurement instruments or fatigue among data subjects that are measured later in the study. Condition 3 constrains the behaviour of $h(\mathbf{x}^{\text{mis}} | \mathbf{X}^{\text{obs}}, \theta_n^{(t)})$ via some moment conditions on the log-likelihood function of the pseudocomplete data, and it implies that a high missingness rate may hurt the performance of the method.

Regarding the implementation of the algorithm, we note that the RO step is the key. However, because of the specificity of each problem, it is neither easy to specify a general regularization function nor to specify a general optimization procedure for obtaining the consistent estimates. The IRO algorithm just provides a general road map for dealing with high dimensional missing data problems, which leaves the user plenty of freedom for choosing appropriate regularization functions and optimization procedures for different problems. In this paper, the IRO algorithm is applied to learning high dimensional GGMs with missing data. How the RO step is performed for this problem will be detailed in Section 3.

Regarding the regularization function, we note that, in general, it can be specified in two ways depending on the problem under consideration. The first way is as conventional, i.e. employing some l_0 -, l_1 - or l_2 -penalties. In this way, we recommend using the same penalty function as if there are no missing data. The second way is via sure screening, which first reduces the space of $\theta_*^{(t)}$ to a low dimensional subspace and then finds a consistent estimate of $\theta_*^{(t)}$ in the low dimensional subspace by using a conventional statistical method, such as maximum likelihood, moment estimation or even penalized optimization. This is equivalent to imposing a binary-type penalty function on the original space, for which the solutions in the low dimensional subspace receives a zero penalty and those outside the subspace receive a penalty of ∞ . Such a binary-type penalty function satisfies condition 1 in Appendix A imposed for regularization functions. For low dimensional problems, the consistent estimator of $\theta_*^{(t)}$ can be obtained by solving equation (3) with a zero penalty function. In this sense, the SEM algorithm can be viewed as a special case of the IRO algorithm.

Finally, we note that, like for other Markov chains, a good initial value will accelerate the convergence of the IRO simulation. There are many ways to specify initial values for the IRO algorithm. In most examples of this paper, we started the simulation with an I-step, where all missingness values are filled by the median of the corresponding variable. This method is simple and usually works well when the missingness rate is not high. However, when we perceive that such a constant filling method does not work well, we may start the simulation with an RO step. In this case, the initial estimate $\theta_n^{(0)}$ may be obtained on the basis of the complete samples (i.e. those without missing entries) only.

2.2. An extension of the imputation-regularized optimization algorithm

As mentioned previously, the RO step is the key to the implementation of the IRO algorithm. We found that for many problems, similarly to the expectation–conditional maximization algorithm (Meng and Rubin, 1993), a consistent estimate of $\theta_*^{(t)}$ can be easily obtained with a number of conditional RO steps, i.e. we can partition the parameter θ into a number of blocks and then find a consistent estimator for each block conditioned on the current estimates of other blocks. Note that, for many problems, the partitioning of θ is natural; see, for example, the examples that are studied in Sections 4 and 5.

Suppose that $\theta = (\theta^{(1)}, \dots, \theta^{(k)})$ has been partitioned into k blocks. The imputation–conditional regularized optimization (ICRO) algorithm can be described as follows.

- a. *I-step*: draw $\tilde{\mathbf{X}}^{\text{mis}}$ from the conditional distribution $h(\mathbf{x}^{\text{mis}} | \mathbf{X}^{\text{obs}}, \theta_n^{(t,1)}, \dots, \theta_n^{(t,k)})$ given $\mathbf{X}^{\text{obs}}$ and the current estimate $\theta_n^{(t)} = (\theta_n^{(t,1)}, \dots, \theta_n^{(t,k)})$.
- b. *CRO step*: on the basis of the pseudocomplete data $\tilde{\mathbf{X}} = (\mathbf{X}^{\text{obs}}, \tilde{\mathbf{X}}^{\text{mis}})$, do the following.
 - i. Conditioned on $(\theta_n^{(t,2)}, \dots, \theta_n^{(t,k)})$, find $\theta_n^{(t+1,1)}$ which forms a consistent estimate of

$$\theta_*^{(t,1)} = \arg \max_{\theta_n^{(t,1)'}, \dots, \theta_n^{(t,k)'}} E \left[\log \left\{ f(\tilde{\mathbf{x}} \mid \theta_n^{(t,1)'}, \theta_n^{(t,2)'}, \dots, \theta_n^{(t,k)'}) \right\} \right],$$

where the expectation is taken with respect to the joint distribution of $\tilde{\mathbf{x}} = (\mathbf{x}^{\text{obs}}, \tilde{\mathbf{x}}^{\text{mis}})$ and the subscript of E gives the current estimate of θ .

- ii. Conditioned on $(\theta_n^{(t+1,1)}, \theta_n^{(t,3)}, \dots, \theta_n^{(t,k)})$, find $\theta_n^{(t+1,2)}$ which forms a consistent estimate of

$$\theta_*^{(t,2)} = \arg \max_{\theta_n^{(t,2)'}, \theta_n^{(t+1,1)}, \theta_n^{(t,2)}, \theta_n^{(t,3)}, \dots, \theta_n^{(t,k)'}} E \left[\log \left\{ f(\tilde{\mathbf{x}} \mid \theta_n^{(t+1,1)}, \theta_n^{(t,2)'}, \theta_n^{(t,3)'}, \dots, \theta_n^{(t,k)'}) \right\} \right]$$

- c. Conditioned on $(\theta_n^{(t+1,1)}, \dots, \theta_n^{(t+1,k-1)})$, find $\theta_n^{(t+1,k)}$ which forms a consistent estimate of

$$\theta_*^{(t,k)} = \arg \max_{\theta_n^{(t,k)'}, \theta_n^{(t+1,1)}, \dots, \theta_n^{(t+1,k-1)}, \theta_n^{(t,k)}} E \left[\log \left\{ f(\tilde{\mathbf{x}} \mid \theta_n^{(t+1,1)}, \dots, \theta_n^{(t+1,k-1)}, \theta_n^{(t,k)'}) \right\} \right].$$

Similarly to equation (3), the consistent estimate of each block can be obtained by maximizing a regularized conditional likelihood function, i.e. setting

$$\theta_{n,p}^{(t+1,i)} = \arg \max_{\theta_n^{(t,i)'}} \left[\frac{1}{n} \sum_{i=1}^n \log \left\{ f(x_i^{\text{obs}}, \tilde{x}_i^{\text{mis}} \mid \theta_n^{(t+1,i)}, \dots, \theta_n^{(t+1,i-1)}, \theta_n^{(t,i)'}, \theta_n^{(t,i)'}, \theta_n^{(t,i+1)}, \dots, \theta_n^{(t,k)}) \right\} \right] - P_{\lambda i}(\theta_n^{(t,i)'}) \Bigg],$$

(4)

for $i = 1, 2, \dots, k$, where $P_{\lambda i}(\cdot)$ denotes the regularization or penalty function that is used for block i . Similarly to the IRO algorithm, it is easy to see that the sequence $\{(\theta_n^{(t,1)}, \dots, \theta_n^{(t,k)})\}$ also forms a Markov chain. The convergence of the Markov chain can be studied under similar conditions to those of the IRO algorithm. Theorem 5 and theorem 6 in Appendix A show that the Markov chain $\{\theta_n^{(t)}\}$ has a stationary distribution and the mean of the stationary distribution forms a consistent estimate of θ^* .

Like the IRO algorithm, the ICRO algorithm just provides a road map for dealing with high dimensional missing data problems. How to partition the parameter vector θ into different blocks and how to obtain a consistent estimate for each block are problem dependent. In this paper, the ICRO algorithm is applied to the problem of high dimensional variable selection with missing data and the problem of random-coefficient linear model estimation. How the CRO step is performed for the two problems will be detailed in Sections 4 and 5 respectively.

3. Learning high dimensional Gaussian graphical models in the presence of missing data

GGMs have often been used in learning gene regulatory networks from microarray data; see, for example, Dobra *et al.* (2004) and Friedman *et al.* (2008). As mentioned in Section 1, missing values can appear in microarray data for many reasons. To deal with missing values in microarray data, many imputation methods, such as single-value decomposition imputation (Troyanskaya *et al.*, 2001), least square imputation (Bo *et al.*, 2004) and Bayesian principal component analysis (BPCA) imputation (Oba *et al.*, 2003) have been proposed. Since these methods impute the missing values independently of the models under consideration, they are often ineffective. Moreover, statistical inference based on the ‘one-time’ imputed data is potentially biased, because uncertainty in the missing values cannot be properly accounted for. In this section, we apply the IRO algorithm to handle missing values for microarray data. The IRO algorithm iteratively imputes missing values on the basis of the updated parameter estimate. Therefore, it overcomes the weakness of the one-time imputation methods and improves accuracy of statistical inference. For a thorough study, the algorithm proposed is also compared with the regression-tree-based multiple-imputation method (Burgette and Reiter, 2010), and the numerical results indicate the superiority of the IRO algorithm.

Let $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$ denote a microarray data set of n samples and p genes, where $\mathbf{x}_i$ is assumed to follow a multivariate Gaussian distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. According to the theory of GGMs, estimating a GGM is equivalent to identifying non-zero elements of the concentration matrix, i.e. the inverse of the covariance matrix $\boldsymbol{\Sigma}$, or to identify non-zero partial correlation coefficients for different pairs of genes. In recent years, a couple of methods have been proposed to estimate high dimensional GGMs, e.g. the graphical lasso (Yuan and Lin, 2007; Friedman *et al.*, 2008), nodewise regression (Meinshausen and Bühlmann, 2006) and ψ -learning (Liang *et al.*, 2015). However, none of them can be directly applied in the presence of missing data.

3.1. The imputation-regularized optimization algorithm

To apply the IRO algorithm to learn GGMs in the presence of missing data, we choose the ψ -learning algorithm as the estimation procedure that is used in the RO step. For GGMs, θ corresponds to the concentration matrix, which can be uniquely determined from the network structure by using the algorithm that is given in Hastie *et al.* (2009), page 634. Under mild conditions, Liang *et al.* (2015) showed that the ψ -learning algorithm provides a consistent estimator for Gaussian graphical networks. Refer to the on-line supplementary material for a brief review of the algorithm. The ψ -learning algorithm belongs to the class of

sure-screening- based algorithms. It first restricts the solution space of the GGM to a low dimensional subspace via correlation screening and then conducts GGM estimation via covariance selection (Dempster, 1972), which, by nature, is a maximum likelihood estimation method. Following from corollary 2 in Appendix A, such an estimation procedure can be used in the IRO algorithm. The resulting estimates implicitly and asymptotically maximize the objective function that is defined in equation (3). The corresponding penalty function is of binary type, which takes a value of 0 in the restricted low dimensional subspace and ∞ otherwise. Other than the ψ -learning algorithm, nodewise regression and the graphical lasso can also be used here, which both produce a consistent estimate for the Gaussian graphical network.

The Gaussian graphical network specifies the dependence between different genes, according to which missing values can be imputed. For convenience, we let $\mathbf{A} = (a_{jk})$ denote the adjacency matrix of a Gaussian graphical network, where $a_{jk} = 1$ if an edge exists between node j and node k and $a_{jk} = 0$ otherwise. For microarray data, a node corresponds to a gene. Let x_{ij} denote a missing entry, and let $\omega(j) = \{k: a_{jk} = 1\}$ denote the neighbourhood of node j . According to the faithfulness property of GGMs, conditionally on the neighbouring genes in $\omega(j)$, gene j is independent of all other genes. Therefore, x_{ij} can be imputed conditionally on the expression values of the neighbouring genes. Mathematically, we have

$$\begin{pmatrix} x_{ij} \\ \mathbf{x}_{i\omega} \end{pmatrix} \sim N \left\{ \begin{pmatrix} \mu_j \\ \boldsymbol{\mu}_\omega \end{pmatrix}, \begin{pmatrix} \sigma_j^2 & \boldsymbol{\Sigma}_{j\omega} \\ \boldsymbol{\Sigma}_{j\omega}^T & \boldsymbol{\Sigma}_{\omega\omega} \end{pmatrix} \right\}, \quad (5)$$

where $\mathbf{x}_{i\omega} = \{x_{ik}: k \in \omega(j)\}$, and $\mu_j, \boldsymbol{\mu}_\omega, \sigma_j^2, \boldsymbol{\Sigma}_{j\omega}$ and $\boldsymbol{\Sigma}_{\omega\omega}$ denote the corresponding mean and variance components. The mean and variance of x_{ij} conditionally on $\mathbf{x}_{i\omega}$ are thus given by

$$\mu_{ij|\omega} = \mu_j + \boldsymbol{\Sigma}_{j\omega} \boldsymbol{\Sigma}_{\omega\omega}^{-1} (\mathbf{x}_{i\omega} - \boldsymbol{\mu}_\omega), \sigma_{ij|\omega}^2 = \sigma_j^2 - \boldsymbol{\Sigma}_{j\omega} \boldsymbol{\Sigma}_{\omega\omega}^{-1} \boldsymbol{\Sigma}_{j\omega}^T. \quad (6)$$

As shown in Liang *et al.* (2015), for each gene, the neighbourhood size can be upper bounded by $\lceil n / \log(n) \rceil$, where $\lceil z \rceil$ denotes the smallest integer that is not smaller than z . Hence, in practice, $\sigma_j^2, \boldsymbol{\Sigma}_{j\omega}$ and $\boldsymbol{\Sigma}_{\omega\omega}$ can be directly estimated from the data. Let $\mathbf{s}_j^2, \mathbf{S}_{j\omega}, \mathbf{S}_{\omega\omega}, \bar{x}_j$ and $\bar{\mathbf{x}}_\omega$ denote the respective sample estimates of $\sigma_j^2, \boldsymbol{\Sigma}_{j\omega}, \boldsymbol{\Sigma}_{\omega\omega}, \mu_j$ and $\boldsymbol{\mu}_\omega$. Then, at each iteration, x_{ij} can be imputed by sampling from the distribution

$$X_{ij|\omega} \sim N \left\{ \bar{x}_j + \mathbf{S}_{j\omega} \mathbf{S}_{\omega\omega}^{-1} (\mathbf{x}_{i\omega} - \bar{\mathbf{x}}_\omega), \mathbf{s}_j^2 - \mathbf{S}_{j\omega} \mathbf{S}_{\omega\omega}^{-1} \mathbf{S}_{j\omega}^T \right\}. \quad (7)$$

In this way, exact evaluation of the concentration matrix can be skipped. In summary, we have the following IRO algorithm for learning GGMs in the presence of missing data.

- a. (Initialization): fill each missing entry by the median of the corresponding variable, and then iterate between the RO and I-steps.
- b. (RO step): apply the ψ -learning algorithm to learn the structure of the Gaussian graphical network.
- c. (I-step): impute missing values according to distribution (7) based on the network learned in the RO step.

The IRO algorithm outputs a series of Gaussian graphical networks. To integrate or average these networks into a single network we adopt the ψ -score averaging approach that was suggested by Liang *et al.* (2015). Let $(\psi_{ij}^{(t)})$ denote the ψ -scores at iteration t , which are obtained from ψ -partial-correlation coefficients via Fisher's transformation. Let $\bar{\psi}_{ij} = \sum_{t=1}^T \psi_{ij}^{(t)} / T$, $i, j = 1, 2, \dots, p$, $i \neq j$, denote the averaged ψ -score for gene i and gene j . Then the averaged network can be obtained by applying a multiple-hypothesis approach to threshold the averaged ψ -scores; if an averaged ψ -score is greater than the threshold value, we set the corresponding element of the adjacency matrix to 1 and to 0 otherwise. The multiple-hypothesis test can be done by using the method of Liang and Zhang (2008), which can be viewed as a generalized empirical Bayesian method (Efron, 2004). The level of significance of the multiple-hypothesis test can be specified in terms of Storey's q -value (Storey, 2002). In this paper, we set it to 0.05.

3.2. A simulated example

We consider an auto-regressive process of order 2 with the concentration matrix given by

$$C_{i,j} = \begin{cases} 0.5, & \text{if } |j-i| = 1, i = 2, \dots, p-1, \\ 0.25, & \text{if } |j-i| = 2, i = 3, \dots, p-2, \\ 1, & \text{if } i = j, i = 1, \dots, p, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

This example has been used by many researchers; see, for example, Yuan and Lin (2007), Mazumder and Hastie (2012) and Liang *et al.* (2015) to illustrate different GGM methods. In this paper, we generated multiple data sets with $n = 200$ and different values of $p = 100, 200, 300, 400$. For each combination of (n, p) , we generated 10 data sets independently, and, for each data set, we randomly deleted 10% of the observations as missing values. To evaluate the performance of the IRO algorithm, precision-recall curves were drawn by varying the threshold value of ψ -scores, where the precision is the fraction of true edges among the retrieved edges, and the recall is the fraction of true edges that have been retrieved over the total amount of true edges.

For each data set, the IRO algorithm was run for 50 iterations. Fig. 1 shows the resulting precision-recall curves for one data set with $p = 400$. For comparison, Fig. 1 also includes the precision-recall curves that are produced by misgLasso and those produced by the ψ -learning algorithm with the missing values imputed by the median filling, BPCA and regression tree (with multiple imputation) methods. The misgLasso algorithm is a

combination of the gLasso and EM algorithms, which integrates out the missing data as in the EM algorithm (see, for example, Stadler and Bühlmann (2012)) and then learns the GGM by using the gLasso algorithm. The misgLasso algorithm has been implemented in the R package spaceExt (He, 2011). The BPCA method has been implemented in the R-package pcaMethods (Stacklies *et al.*, 2007) and the regression tree method has been implemented in the R package MICE (van Buuren and Groothuis-Oudshoorn, 2011). In the regression tree method, for each data set, the missing values were imputed 20 times and, as the IRO-Ave curve, the MIRegTree curve that is reported in Fig. 1 was obtained with the ψ -scores averaged over the 20 pseudocomplete data sets. Refer to Fig. 1 of the on-line supplementary material for the precision-recall curves for other values of p .

Table 1 compares the averaged areas (over 10 data sets) under the precision-recall curves produced by various methods. The comparison indicates that IRO-Ave outperforms all others for this example. It is interesting that, although IRO-Last is also based on one-time imputation, it significantly outperforms the median filling, regression tree (with multiple imputation) and BPCA methods. The misgLasso algorithm does not work well for this example. This inferiority is not due to the EM algorithm, but the gLasso algorithm which does not work well for this example. This is consistent with Liang *et al.* (2015), where it was shown that the ψ -learning algorithm works much better than gLasso for the complete-data version of this example.

3.3. Yeast cell expression data

Using deoxyribonucleic acid microarrays, Gasch *et al.* (2000) measured the changes in gene expression levels in the yeast *Saccharomyces cerevisiae* as cells responded to different environmental changes, such as temperature shocks, hydrogen peroxide and nitrogen source depletion. The whole data set is available from <http://genome-www.stanford.edu/yeast-stress/>, consisting of 173 samples and 6152 known or predicted yeast genes and having a missingness rate of 3.01%.

In this example, our goal is to study the regulatory relationships between yeast genes. We considered 1000 genes with expression levels having the largest variation over the samples. The missingness rate for the subset data is also 3.01%. On the basis of the subset data, the IRO algorithm produced an approximate scale-free gene regulatory network which is reported in the on-line supplementary material. From the network, a few hub genes can be identified, and the top three of them are ARB1, HXT5 and XBP1. A gene is called a hub gene if it has strong connectivity to other genes. We find that the top three hub genes are all included in the transcriptional factor (TF)-gene direct regulatory network (Yang *et al.*, 2014), which is available from the yeast transcriptional regulatory pathway database at <http://cosbi3.ee.ncku.edu.tw/YTRP>. The yeast transcriptional regulatory pathway database has deposited all possible transcriptional regulatory pathways for each TF-gene regulatory pair inferred from the TF perturbation experiment and can be somehow considered as the 'ground truth' for gene regulatory relationships. For comparison, we have applied the BPCA, median filling, regression tree (with multiple imputation) and misgLasso methods to this example with the top three rule employed in identifying hub genes. The misgLasso algorithm always aborted for this data set, which may be due to some computational issues

that are involved in the package. Among the top three hub genes identified by the BPCA method, only one gene is included in the TF-gene direct regulatory network from the yeast transcriptional regulatory pathway database. For the median filling and regression tree methods, this number is 0 and 1 respectively. This comparison indicates the superiority of the IRO algorithm over the competitive methods in learning true gene regulatory relationships for the data set. Refer to the supplementary material for details.

4. High dimensional variable selection in presence of missing data

This problem is also motivated by microarray data analysis, but the goal has been shifted to selection of genes that are relevant to a particular phenotype. To be more general, we let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ denote the response vector for n observations, and let $\mathbf{X} = (X_1, \dots, X_n)^T$ denote the matrix of covariates, where each X_i is a p -dimensional vector and p can be much larger than n (also known as small n -large p). The response variable and covariates are linked through the regression

$$\mathbf{Y} = (\mathbf{1}_n, \mathbf{X})\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (9)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ denotes the vector of regression coefficients, and $\boldsymbol{\epsilon} \sim N(0, \sigma_e^2 \mathbf{I}_n)$ denotes the vector of random errors.

Variable selection for model (9) with complete data has been extensively studied in the recent literature. Methods have been developed from both frequentist and Bayesian perspectives; see, for example, Tibshirani (1996), Johnson and Rossell (2012) and Song and Liang (2015a). For incomplete data, Garcia *et al.* (2010) proposed to conduct variable selection by maximizing the penalized likelihood function of the incomplete data. However, when p is large and the covariates X_i are generally correlated, the incomplete-data likelihood function can be intractable, rendering failure of their method. Zhao and Long (2016) showed through numerical studies that for the high dimensional data the standard multiple-imputation approach performs poorly, whereas the imputation method based on the Bayesian lasso often works better. However, since the Bayesian lasso tends to overshrink the non-zero regression coefficients, its consistency in variable selection is difficult to justify when p is much greater than n (Castillo *et al.*, 2015). Quite recently, Long and Johnson (2015) proposed to combine Bayesian lasso imputation and stability selection (Meinshausen and Bühlmann, 2010). Again, the consistency of this method is difficult to justify because of the inconsistency of the Bayesian lasso.

4.1. The imputation–conditional regularized optimization algorithm

In what follows, we consider a general setting of model (9), where the covariates follow a multivariate Gaussian distribution $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Under this setting, the parameter vector $\boldsymbol{\theta}$ consists of three natural blocks $\boldsymbol{\beta}$, σ_e^2 and the concentration matrix $\mathbf{C} = \boldsymbol{\Sigma}^{-1}$. Since n has been assumed to be smaller than p , we further assume sparsity for both the regression coefficients $\boldsymbol{\beta}$ and the concentration matrix $\mathbf{C}$.

To apply the ICRO algorithm to this problem, we employ the sure independence screening (SIS)-minimax concave penalty (MCP) algorithm for estimating β , i.e. the variables are first subject to an SIS procedure, and then the survived variables are selected by using the MCP method (Zhang, 2010). This algorithm has been implemented in the R package SIS, and the resulting estimate maximizes the regularized conditional likelihood function as defined in expression (4), where the regularization function is given by the MCP penalty (Zhang, 2010) in the restricted subspace and ∞ otherwise. Given the estimate of β , σ_e^2 can be estimated by $\hat{\sigma}_e^2 = \sum_{i=1}^n \hat{\epsilon}_i^2 / (n - |\hat{\beta}| - 1)$, where $\hat{\epsilon}_i$ denotes the residual of sample i , and $|\hat{\beta}|$ denotes the number of non-zero elements that are included in the estimate $\hat{\beta}$. Given the consistency of $\hat{\beta}$, the consistency of $\hat{\sigma}_e^2$ can be easily justified, for which the corresponding penalty function is

0. To estimate the concentration matrix C , we employ the ψ -learning algorithm. As mentioned previously, the ψ -learning algorithm provides a consistent estimate for the Gaussian graphical network, based on which a consistent estimate of the concentration matrix can be uniquely determined by the algorithm given in Hastie *et al.* (2009), page 634. As discussed in Section 3, the ψ -learning algorithm implicitly and asymptotically maximizes the regularized conditional likelihood function as defined in expression (4) with a binary-type penalty function. Note that the SIS-MCP algorithm does not make use of the dependence between the covariates. Given the structure of the ICRO algorithm, some other variable selection algorithms which have made use of the dependence between the covariates, e.g. Yu and Liu (2016), can also be applied here.

Next, we consider the imputation step. Suppose that the value of x_{hk} is missed in $\mathbf{X}$. Section 4 of the on-line supplementary material presents the conditional distributions of X_{hk} given $\mathbf{Y}$ and the remaining elements of $\mathbf{X}$ under various scenarios. On the basis of the conditional distributions, x_{hk} can be easily imputed by sampling from the respective sampled conditional distributions. Here the sampled conditional distribution refers to the distribution with its population parameters replaced by their respective estimates calculated from samples. For example, β_s are replaced by their SIS-MCP estimates and $\hat{\sigma}_e^2$ is replaced by $\hat{\sigma}_e^2$. In summary, the ICRO algorithm works as follows.

- a. (Initialization): fill each missing entry of $\mathbf{X}$ by the median of the corresponding variable, and then iterate between the CRO and I-steps.
- b. (CRO-step):
 - i. apply the SIS-MCP algorithm to estimate the regression coefficients β ;
 - ii. estimate $\hat{\sigma}_e^2$ conditionally on the estimate of β ;
 - iii. apply the ψ -learning algorithm to learn the structure of the Gaussian graphical network.
- c. (I-step): impute missing values according to the conditional distributions (given in the online supplemental material) on the basis of the regression model and network structure learned in the CRO step.

4.2. A simulated example

The data sets were simulated from model (9) with $n = 100$ and $p = 200$ and $p = 500$. The covariates $\mathbf{X}$ were generated under two settings:

- a. the covariates are mutually independent, where $\mathbf{x}_i \sim N(0, 2I_n)$ for $i = 1, \dots, n$;
- b. the covariates are generated according to the concentration matrix (8).

For both settings, we set $(\beta_0, \beta_1, \dots, \beta_5) = (1, 1, 2, -1.5, -2.5, 5)$ and $\beta_6 = \dots = \beta_p = 0$, and random error $\epsilon \sim N(0, I_n)$. For each pair of (n, p) , we simulated 10 data sets independently. For each data set, we considered two missingness rates, randomly deleting 5% and 10% entries of $\mathbf{X}$ to create missingness data. The performance of the various methods was measured by using three criteria:

$$\text{err}_{\beta}^2 = \|\hat{\beta} - \beta\|^2,$$

$$\text{fsr} = |\mathbf{s}^*|/|\mathbf{s}|,$$

$$\text{nsr} = |\mathbf{s}^* \setminus \mathbf{s}|/|\mathbf{s}^*|,$$

where ' $\|\cdot\|$ ' denotes the Euclidean norm, $\hat{\beta}$ denotes the estimate of β , $\mathbf{s}^*$ denotes the set of true covariates and $\mathbf{s}$ denotes the set of selected covariates.

The ICRO algorithm was first applied to this example with the results summarized in Table 2 and 3. For each data set, the algorithm was run for 30 iterations. For variable selection, we kept only the variables that appeared five or more times in the last 10 iterations. For estimation of β , we averaged the estimates of β that were obtained in the last 10 iterations. For comparison, we also tried the median filling, BPCA and regression tree (with multiple imputation) methods. As explained previously, the median filling method fills each missing value by the median of the corresponding variable, and BPCA imputes the missing values on the basis of the principal component regression. The regression tree method performs multiple imputation via sequential regression trees: as in ICRO, we kept only the variables that appeared five or more times in 10 imputations for variable selection, and averaged the estimates of β that were obtained in all 10 imputations for estimation of β .

The comparison indicates that the ICRO algorithm works extremely well for this example. For the case of independent covariates, its results are almost as good as those obtained from the complete data. In both cases, the ICRO algorithm significantly outperforms the other imputation methods.

As an extension of this example, we also considered one case in which the missing data are generated under the mechanism of missingness at random. To mimic this mechanism, we deleted each entry of $\mathbf{X}$ with a probability proportional to the absolute value of the mean of its conditional distribution; see (equation 2) of the on-line supplementary material, i.e. the

propensity for each data point to be missing is not related to the missing data but is related to some of the observed data. Table 2 of the supplementary material summarizes the results from the median filling, BPCA, regression tree (with multiple imputation) and ICRO methods. The comparison indicates that, under this data missingness mechanism, ICRO can still work well and significantly outperform the competitors.

4.3. Expression data for Bardet–Biedl syndrome

We analysed one real gene expression data set about Bardet–Biedl syndrome (Scheetz *et al.*, 2006). The complete data set contains 120 samples, where the expression level of the gene TRIM32 works as the response variable and the expression levels of 200 other genes work as the predictors. The data set is available in the R package flare. We generated 10 incomplete data sets from the complete data set by randomly deleting 5% of the observations. For each incomplete data set, we ran the ICRO algorithm for 30 iterations and averaged the estimates of β obtained in the last 10 iterations as the final estimate. For comparison, the median filling, BPCA and regression tree (with multiple imputation) methods were also applied to this example. Table 4 summarizes the estimation errors of $\hat{\beta}$ (with respect to β_c , the estimate of β from the complete data) produced by the four methods for 10 incomplete data sets.

We have also explored the results of variable selection. The complete-data model selects five variables: v.153, v.180, v.185, v.87 and v.200. For the ICRO, median filling, BPCA and regression tree models, we count the selection frequency of each variable for the 10 incomplete data sets. For the ICRO and regression tree methods, the top five variables in selection frequency are v.153, v.185, v.180, v.87 and v.200, which are the same (ignoring the order) as for the complete-data model. For the median filling models, the top five variables are v.153, v.185, v.62, v.200 and v.54. For the BPCA models, the top five variables are v.153, v.87, v.185, v.62 and v.200. Both the results of β -estimation and of variable selection indicate the superiority of ICRO over the other imputation methods.

5. A random-coefficient linear model

To illustrate the use of the ICRO algorithm further, we consider a random-coefficient linear model. Such a model often arises, for instance, in recommendation systems where customers rate different items, e.g. products or service. Specifically, we simulate the data from the model

$$y_{ij} = \mathbf{x}_{ij}^T \beta + \mathbf{z}_i^T \lambda_i + \mathbf{w}_j^T \gamma_j + e_{ij}, e_{ij} \sim N(0, \sigma^2), \quad \lambda_i \sim N(0, \Lambda), \quad \gamma_j \sim N(0, \Gamma), \quad (10)$$

where y_{ij} represents the response for customer i on item j . Assume that there are I customers and each customer responds to J items. Thus, the data set consists of a total of $n = IJ$ observations. The vector $\mathbf{x}_{ij}$ represents the covariates that characterize the customers and items, e.g. how and how long the customer has purchased the item, $\mathbf{z}_i$ represents customer-specific covariates such as gender, education and demographics, and $\mathbf{w}_j$ represents item-specific covariates, e.g. the manufacturer and category of the item. The vector λ_i represents the customer-specific (random) coefficients and γ_j represents the item-specific (random)

coefficients. This model can be easily extended to the case where each customer responds to only a subset of items. For this model, we treat the random coefficients λ_i and γ_j as missing data and are interested in estimation of β . For simplicity, we assume that β is low dimensional, although the whole data set can be big when I and/or J become large. Under this assumption, the ICRO algorithm is essentially reduced to the SEM algorithm for this example. Instead of using the ICRO algorithm in this straightforward way, we propose to use it under the Bayesian framework. This extends the applications of the ICRO algorithm to Bayesian computation.

To conduct Bayesian analysis for the model, we assume the following semiconjugate priors:

$$\left. \begin{aligned} \beta &\sim N(\mu_\beta, \Sigma_\beta), \\ \sigma^2 &\sim \text{IG}(a, b), \\ \Lambda &\sim \text{IW}(\rho_\Lambda, \mathbf{R}_\Lambda), \\ \Gamma &\sim \text{IW}(\rho_\Gamma, \mathbf{R}_\Gamma), \end{aligned} \right\} \quad (11)$$

where $\text{IG}(\cdot, \cdot)$ denotes the inverse gamma distribution, $\text{IW}(\cdot, \cdot)$ denotes the inverse Wishart distribution and $\mu_\beta, \Sigma_\beta, a, b, \rho_\Lambda, \mathbf{R}_\Lambda, \rho_\Gamma$ and $\mathbf{R}_\Gamma$ are hyperparameters to be specified by the user. Each of these priors is individually conjugate to the normal likelihood function, given the other parameters, although the joint prior is not conjugate. The resulting full conditional posterior distributions and their modes are derived in section 5 of the on-line supplementary material. Since, under the low dimensional setting, the mode of the full conditional posterior distribution provides a consistent estimator for the corresponding parameter, the ICRO algorithm can work as follows.

- a. (Initialization): initialize the λ_s , γ_s and all parameters by some random numbers.
- b. (CRO step): estimate the parameters β , Λ , Γ and σ^2 by the mode of their respective full conditional posterior distributions, for which the log-prior density plays the role of the penalty function that is imposed in expression (4).
- c. (I-step): impute the values of the λ_s and γ_s according to their respective full conditional posterior distributions.

Under the Bayesian framework, the ICRO algorithm works in a similar way to the Gibbs sampler except that it replaces posterior samples of the parameters by their respective full conditional posterior modes. Also, in this case, it is reduced to a hybrid of data augmentation (Tanner and Wong, 1987) and iterative conditional modes (Besag, 1974) by using imputation for the missing data and conditional modes for the parameters. However, the ICRO algorithm offers more, whose consistency step allows it to conduct parameter estimation based on subsamples only, and this can create great savings in computation for big data problems. For high dimensional problems, if the choice of prior distributions ensures posterior consistency, then the above algorithm can still be employed.

Fig. 3 of the on-line supplementary material compares the sampling path and auto-correlations of the ICRO and Gibbs samples. As expected, the comparison shows that the ICRO algorithm can converge faster than the Gibbs sampler and, in addition, the samples that are generated by the ICRO algorithm tend to have smaller variations than those by the Gibbs sampler. Note that the accuracy of the ICRO estimates is achieved at the price that we sacrifice the variance information that is contained in the posterior samples. As pointed out by Nielsen (2000), the variance of the ICRO samples reflects the loss of information due to the missing data. However, for the random-coefficient example, the variance information can be obtained from the full conditional posterior distributions (given in the on-line supplementary material) by simply plugging the parameter estimates into their variances. This observation suggests a general strategy to improve simulations of the Gibbs sampler: at each iteration, we need to draw samples for only the components for which the posterior variance is not analytically available and also of interest to us, and the other components can be replaced by the mode of the respective full conditional posterior distributions. As aforementioned, the mode can be found with a subset of samples, which can be much cheaper than sampling from the full data conditional posterior. We expect that the strategy proposed can significantly facilitate Bayesian computation for big data problems. A further study of this proposed strategy will be reported elsewhere.

6. Discussion

In this paper, we have proposed the IRO algorithm, as a general algorithm for dealing with high dimensional missing data problems. Under quite general conditions, we show that the IRO algorithm can lead to a consistent estimate for the parameters. We have also extended the IRO algorithm to the case of multiple block parameters, which leads to the ICRO algorithm. We illustrate the proposed algorithms by using high dimensional GGMs, high dimensional variable selection and a random-coefficient model.

Like the EM algorithm for low dimensional data, we expect that the IRO or ICRO algorithm can have many applications for high dimensional data. With the IRO or ICRO algorithm, many problems can be much simplified, e.g. variable selection for high dimensional mixture regression (Khalili and Chen, 2007) and variable selection for high dimensional mixed effect models (Fan and Li, 2012). For the former, the group index of each sample can be treated as missing data, and then the IRO algorithm can be applied: the I-step assigns the samples to different groups and the RO step conducts variable selection for each group separately. For the latter, at each iteration of the IRO algorithm, it is reduced to variable selection for a high dimensional regression with fixed designs given the imputed random effects.

To assess the convergence of IRO simulations, we recommend the Gelman-Rubin statistic (Gelman and Rubin, 1992). Since the IRO algorithm usually converges very fast, we do not recommend a long run. Our experience shows that 20 iterations have often been sufficiently long to produce a stable parameter estimate. Because of the Markov chain Monte Carlo nature of IRO simulations, we recommend the averaging method for parameter estimation and allow a burn-in period before collecting the samples for averaging. In Section 3.2, we reported the results based on the last iteration, which is just to compare with other one-time imputation methods. Theoretically, the variation in the IRO estimates that are collected at

different iterations reflects the missing data information. Hence, in addition to parameter estimates, one might report their variance over iterations. However, as this information is often not of interest to us, we choose not to report them in the paper.

When a large amount of noise is brought into the system through missing data, the IRO or ICRO algorithm may fail to work as implied by condition 3 in Appendix A. When the sample size is small, consistent estimators may not be adequately accurate. Our numerical experience shows that, in this case, the IRO or ICRO algorithm might not significantly outperform onetime imputation methods, such as median filling and BPCA, as there is not much information to use for improving imputation during iteration. In general, when the sample size increases, the IRO or ICRO algorithm can significantly outperform one-time imputation methods.

Regarding statistical inference for parameters, we note that, theoretically, it can be done in general scenarios, not limited to the constraint that the posterior distribution of the parameters has a closed form. For example, for high dimensional regression, if the lasso algorithm is employed as the consistent estimator in the IRO algorithm, then the inference for θ , e.g. constructing confidence intervals for each component of θ , can be done by using the desparsified method (Zhang and Zhang, 2014; van de Geer *et al.*, 2014). Let $V(\mathbf{x}^{\text{obs}}, \tilde{\mathbf{x}}_t^{\text{mis}}, \theta_n^{(t)})$ denote an uncertainty assessment statistic that is obtained at iteration t . Assume that $V(\mathbf{x}^{\text{obs}}, \tilde{\mathbf{x}}_t^{\text{mis}}, \theta)$ is a Lipschitz function with respect to θ and is integrable. Then, by corollary 3 or corollary 4 in Appendix A (depending on whether IRO or ICRO is being used), we shall be able to obtain an uncertainty assessment for θ by averaging $V(\mathbf{x}^{\text{obs}}, \tilde{\mathbf{x}}_t^{\text{mis}}, \theta_n^{(t)})$ along the IRO or ICRO chain. However, how to obtain the uncertainty assessment statistic $V(\mathbf{x}^{\text{obs}}, \tilde{\mathbf{x}}_t^{\text{mis}}, \theta_n^{(t)})$ for general high dimensional problems is beyond the scope of this paper.

As a variant of the ICRO algorithm, we note that the I-step can be replaced by an E-step if it is available. In this case, the RO step maximizes the objective function

$$\max_{\theta} E_{\theta^*} [Q(\mathbf{x}^{\text{obs}}, \theta | \theta_n^{(t)})], \quad (12)$$

where the Q -function is as defined in the EM algorithm, and E_{θ^*} denotes expectation with respect to the true distribution $\pi(\mathbf{x}^{\text{obs}} | \theta^*)$. Suppose that θ has been partitioned into a few blocks $\theta = (\theta^{(1)}, \dots, \theta^{(k)})$. To solve problem (12), the ICRO algorithm is reduced to a block-wise consistency algorithm, which finds consistent estimates iteratively for the parameters of each block conditioned on the current estimates of the parameters of other blocks. The block-wise consistency algorithm is closely related to, but more flexible than, the co-ordinate ascent algorithm (Tseng, 2001; Tseng and Yun, 2009). The co-ordinate ascent algorithm iteratively finds the exact maximizer for each block conditioned on the current estimates of the parameters of other blocks. Under appropriate conditions, such as contraction as in condition 11' in Appendix A and uniform consistency of the $\theta_n^{(t)}$ s as shown

in theorem 2, we shall be able to show that the paths of the two algorithms will converge to the same point. This will be explored elsewhere.

The ICRO algorithm has strong implications for big data computing. On the basis of the ICRO algorithm, we have proposed a general strategy to improve Bayesian computation under big data scenarios, i.e. we can replace posterior samples by posterior modes in Gibbs iterations to accelerate simulations, where the posterior modes can be calculated with a subset of samples. In addition, the IRO or ICRO algorithm facilitates data integration from multiple sources when missing data are present. With the IRO or ICRO algorithm, the problem of data integration for incomplete data is converted into a problem for complete data and thus many of the existing meta-analysis methods can be conveniently applied for inference. This is very important for big data analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Liang's research was supported in part by grants DMS-1612924, DMS/NIGMS R01-GM117597 and NIGMS R01-GM126089. The authors thank the Joint Editor, Associate Editor and three referees for their constructive comments which have led to a significant improvement in this paper.

Appendix A

A.1. Proof of consistency of $\theta_n^{(t+1)}$

Define Θ_n as the parameter space of θ , where the subscript n indicates the dependence of the dimension of θ on the sample size n . Without possible confusion, we shall refer to Θ_n as Θ .

In addition, we let $\theta_n^T = \{\theta_n^{(1)}, \dots, \theta_n^{(T)}\}$ denote a path of θ_n in the IRO algorithm, which can be considered as an arbitrary subset of Θ_n with T elements (replicates are allowed). Let

$\tilde{x} = (x^{\text{obs}}, \tilde{x}^{\text{mis}})$ and define

$$G_n(\theta \mid \theta_n^{(t)}) = E_{\theta^{(t)}}[\log\{f_{\theta}(\tilde{x})\}] = \int \log\{f_{\theta}(\tilde{x})\} f(x^{\text{obs}} \mid \theta^*) h(\tilde{x}^{\text{mis}} \mid \theta_n^{(t)}) d\tilde{x}, \quad \hat{G}_n(\theta \mid \tilde{x}, \theta_n^{(t)}) = \frac{1}{n} \sum_{i=1}^n \log\{f(x_i^{\text{obs}}, \tilde{x}_i^{\text{mis}} \mid \theta)\},$$

$$\tilde{G}_n(\theta \mid \theta_n^{(t)}) = \frac{1}{n} \sum_{i=1}^n \int \log\{f(x_i^{\text{obs}}, \tilde{x}^{\text{mis}} \mid \theta)\} h(\tilde{x}^{\text{mis}} \mid x_i^{\text{obs}}, \theta_n^{(t)}) d\tilde{x}^{\text{mis}} = \frac{1}{n} \sum_{i=1}^n q(x_i^{\text{obs}}).$$

(13)

Our first goal is to show that the following uniform law of large numbers holds for any T such that $\log(T) = o(n)$:

$$\sup_{\theta_n^{(t)} \in \Theta_n} \sup_{\theta \in \Theta_n} |\hat{G}_n(\theta|\tilde{\mathbf{x}}, \theta_n^{(t)}) - G_n(\theta|\theta_n^{(t)})| \xrightarrow{p} 0, \quad (14)$$

where $\xrightarrow{p}$ denotes convergence in probability. To achieve this goal, we need the following conditions.

Condition 1. $\log\{f_{\theta}(\tilde{x})\}$ is a continuous function of θ for each $\tilde{x} \in \mathcal{X}$ and a measurable function of $\tilde{x}$ for each θ .

Condition 2 (conditions for Glivenko–Cantelli theorem).

- a. There is a function $m_n(\tilde{x})$ such that $\sup_{\theta \in \Theta_n, \tilde{x} \in \mathcal{X}} |\log\{f_{\theta}(\tilde{x})\}| \leq m_n(\tilde{x})$.
- b. Define $\tilde{m}_n(x^{\text{obs}}, \theta_n^{(t)}) = \int m_n(\tilde{x}) h(\tilde{x}^{\text{mis}} | x^{\text{obs}}, \theta_n^{(t)}) d\tilde{x}^{\text{mis}}$. Assume that there exists $m_n^*(x^{\text{obs}})$ such that $0 \leq \tilde{m}_n(x^{\text{obs}}, \theta_n^{(t)}) \leq m_n^*(x^{\text{obs}})$ for all $\theta_n^{(t)}, E[m_n^*(x^{\text{obs}})] < \infty$ and $\sup_{n \in \mathbb{Z}} E[m_n^*(x^{\text{obs}}) \mathbf{1}\{m_n^*(x^{\text{obs}}) \geq \zeta\}] \rightarrow 0$ as $\zeta \rightarrow \infty$. In addition, $\sup_{n \geq 1} \sup_{x \in \mathcal{X}, \theta \in \Theta_n} \int m_n(\tilde{x}) \mathbf{1}\{m_n(\tilde{x}) > \zeta\} h(\tilde{x}^{\text{mis}} | x, \theta) d\tilde{x}^{\text{mis}} \rightarrow 0$ as $\zeta \rightarrow \infty$.
- c. Define $\mathcal{F}_n = \left\{ \int \log\{f(x^{\text{obs}}, \tilde{x}^{\text{mis}} | \theta)\} h(\tilde{x}^{\text{mis}} | x^{\text{obs}}, \theta_n^{(t)}) d\tilde{x}^{\text{mis}} | \theta, \theta_n^{(t)} \in \Theta_n \right\}$ and $\mathcal{G}_{n,M} = \left\{ q \mathbf{1}\{m_n^*(x^{\text{obs}}) \leq M\} q \in \mathcal{F}_n \right\}$. Suppose that, for every ϵ and $M > 0$, the metric entropy $\log N[\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)] = o_p^*(n)$, where $\mathbb{P}_n$ is the empirical measure of x^{obs} , and $N[\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)]$ is the covering number with respect to the $L_1(\mathbb{P})$ -norm.

Condition 3 (conditions for imputed data). Define

$Z_{t,i} = \log\{f(x_i^{\text{obs}}, \tilde{x}_i^{\text{mis}} | \theta)\} - \int \log\{f(x_i^{\text{obs}}, \tilde{x}^{\text{mis}} | \theta)\} \times h(\tilde{x}^{\text{mis}} | x_i^{\text{obs}}, \theta_n^{(t)}) d\tilde{x}^{\text{mis}}$. Suppose that, for any θ and $\theta_n^{(t)} \in \Theta_n$, $E[Z_{t,i}]^m \leq m! M_b^{m-2} v_i/2$ for every $m \geq 2$ and some constants $M_b > 0$ and $v_i = O(1)$, i.e. $Z_{t,i}$ are subexponential random variables.

Theorem 1. Assume conditions 1–3; then law (14) holds for any T such that $\log(T) = o(n)$.

Proof. By the definitions in expression (13), we have the decomposition

$$\hat{G}_n(\theta|\tilde{\mathbf{x}}, \theta_n^{(t)}) - G_n(\theta|\theta_n^{(t)}) = \{\hat{G}_n(\theta|\tilde{\mathbf{x}}, \theta_n^{(t)}) - \tilde{G}_n(\theta|\theta_n^{(t)})\} + \{\tilde{G}_n(\theta|\theta_n^{(t)}) - G_n(\theta|\theta_n^{(t)})\}, \quad (15)$$

which consists of two terms; the first term comes from imputation of missing data, and the second term comes from the observed data.

First, we show that the second term of equation (15) converges to 0 uniformly, following the proof of theorem 2.4.3 of van der Vaart and Wellner (1996). By the symmetrization lemma 2.3.1 of van der Vaart and Wellner (1996), measurability of the class $\mathcal{F}_n$ and Fubini's theorem,

$$\begin{aligned} E^* \left[\sup_{\theta, \theta_n^{(t)} \in \Theta_n} \left| \tilde{G}_n(\theta | \theta_n^{(t)}) - G_n(\theta | \theta_n^{(t)}) \right| \right] &\leq 2E_{x \text{ obs}} \left[E_{\epsilon} \left[\sup_{q(x) \in \mathcal{F}_n} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i q(x_i^{\text{obs}}) \right\| \right] \right] \\ &\leq 2E_{x \text{ obs}} \left[E_{\epsilon} \left[\sup_{q(x) \in \mathcal{G}_{n,M}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i q(x_i^{\text{obs}}) \right\| \right] + 2E^* \left[m_n^*(x^{\text{obs}}) \mathbf{1} \{ m_n^*(x^{\text{obs}}) > M \} \right] \right], \end{aligned}$$

where ϵ_i are independent and identically distributed Rademacher random variables with $P(\epsilon_i = 1) = P(\epsilon_i = -1) = \frac{1}{2}$, and E^* denotes the outer expectation.

By condition 2, $2E^* \left[m_n^*(x^{\text{obs}}) \mathbf{1} \{ m_n^*(x^{\text{obs}}) > M \} \right] \rightarrow 0$ for sufficiently large M . To prove convergence in mean, it suffices to show that the first term converges to 0 for fixed M . Fix $x_1^{\text{obs}}, \dots, x_n^{\text{obs}}$, and let $\mathcal{H}$ be an ϵ -net in $L_1(\mathbb{P}_n)$ over $\mathcal{G}_{n,M}$; then

$$E_{\epsilon} \left[\sup_{q(x) \in \mathcal{G}_{n,M}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i q(x_i^{\text{obs}}) \right\| \right] \leq E_{\epsilon} \left[\sup_{q(x) \in \mathcal{H}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i q(x_i^{\text{obs}}) \right\| \right] + \epsilon.$$

The cardinality of $\mathcal{H}$ can be chosen equal to $N\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\}$. Bounding the L_1 -norm on the right-hand side by the Orlicz norm ψ_2 and using the maximal inequality (lemma 2.2.2 of van der Vaart and Wellner (1996)) and Hoeffding's inequality, it can be shown that

$$E_{\epsilon} \left[\sup_{q(x) \in \mathcal{G}_{n,M}} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i q(x_i^{\text{obs}}) \right\| \right] \leq K \sqrt{(1 + \log[N\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\}])} \sqrt{(6/n)M + \epsilon} \xrightarrow{P^*} \epsilon, \quad (16)$$

where K is a constant, and P^* denotes outer probability. It has been shown that the left-hand side of inequality (16) converges to 0 in probability. Since it is bounded by M , its expectation with respect to $x_1^{\text{obs}}, \dots, x_n^{\text{obs}}$ converges to 0 by the dominated convergence theorem.

This concludes the proof that $\sup_{\theta_n^{(t)} \in \Theta_n} \sup_{\theta \in \Theta_n} |G_n(\theta | \theta_n^{(t)}) - G_n(\theta | \theta_n^{(t)})| \xrightarrow{P} 0$ in mean.

Further, by Markov inequality, we conclude that

$$\sup_{\theta_n^{(t)} \in \Theta_n} \sup_{\theta \in \Theta_n} |\tilde{G}_n(\theta|\theta_n^{(t)}) - G_n(\theta|\theta_n^{(t)})| \xrightarrow{p} 0. \quad (17)$$

To establish the uniform convergence of the first term of equation (15), we fix $x_1^{\text{obs}}, \dots, x_n^{\text{obs}}$.

By condition 3, $n\{\hat{G}_n(\theta|\tilde{x}, \theta_n^{(t)}) - \tilde{G}_n(\theta|\theta_n^{(t)})\} = Z_{t,1} + Z_{t,2} + \dots + Z_{t,n}$. By Bernstein's inequality,

$$P\left\{n\left|\hat{G}_n(\theta|\tilde{x}, \theta_n^{(t)}) - \tilde{G}_n(\theta|\theta_n^{(t)})\right| > z\right\} = P\left\{|Z_{t,1} + Z_{t,2} + \dots + Z_{t,n}| > z\right\} \leq 2\exp\left(-\frac{1}{2} \frac{z^2}{v + M_b z}\right),$$

for $v = v_1 + \dots + v_n$. By lemma 2.2.10 of van der Vaart and Wellner (1996), for Orlicz norm ψ_1 , we have

$$\left\| \sup_{\theta \in \Theta_n, t=1,2,\dots,T} n\left\{\hat{G}_n(\theta|\tilde{x}, \theta_n^{(t)}) - \tilde{G}_n(\theta|\theta_n^{(t)})\right\} \right\|_{\psi_1} \leq \epsilon + K\left(M_b \log\left[1 + TN\left\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\right\}\right]\right) + \sqrt{v \log\left[1 + TN\left\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\right\}\right]},$$

for a constant K and any $\epsilon > 0$. By condition, part 2(c), and the condition $\log(T) = \alpha(n)$,

$$\left\| \sup_{\theta \in \Theta_n, t=1,2,\dots,T} n\left\{\hat{G}_n(\theta|\tilde{x}, \theta_n^{(t)}) - \tilde{G}_n(\theta|\theta_n^{(t)})\right\} \right\|_{\psi_1} \leq \epsilon + K\left(M_b \log\left[1 + TN\left\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\right\}/n\right]\right) + \sqrt{(v/n) \log\left[1 + TN\left\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\right\}/n\right]} \xrightarrow{P^*} \epsilon.$$

Therefore,

$$\sup_{\theta \in \Theta_n, t=1,2,\dots,T} |\hat{G}_n(\theta|\tilde{x}, \theta_n^{(t)}) - \tilde{G}_n(\theta|\theta_n^{(t)})| \xrightarrow{p} 0. \quad (18)$$

The theorem can then be concluded by combining results (17) and (18).

Remark 1 (on the metric entropy condition). Assume that all elements in $\bigcup_n \mathcal{F}_n$ are uniformly Lipschitz with respect to the L_1 -norm. Then the metric entropy $\log\left[N\left\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\right\}\right]$ can be measured on the basis of the parameter space Θ . Since the functions in $\mathcal{G}_{n,M}$ are all bounded, the corresponding parameter space $\Theta_{n,M}$ can be contained in an L_1 -ball by the continuity of $\log\{f(\tilde{x}|\theta)\}$ in θ . Further, we assume that the diameter of the L_1 -ball or the space $\Theta_{n,M}$ grows at a rate of $O(n^\alpha)$ for some $0 \leq \alpha < \frac{1}{2}$; then $\log\left[N\left\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\right\}\right] = O\left\{n^{2\alpha \log(p)}\right\}$ holds, which allows p to grow at a polynomial rate of

$O(n^\gamma)$ for some constant $0 < \gamma < \infty$. The increased diameter accounts for the conventional assumption that the size of the true model grows with the sample size n . Refer to Vershynin (2015) for more discussion on this issue. Similar conditions on metric entropy have been used in the literature of high dimensional statistics. For example, Raskutti *et al.* (2011) studied minimax rates of estimation for high dimensional linear regression over L_q -balls.

Remark 2 (on the condition of T). Since the imputation step draws random data at each iteration t , there is no way to show uniform convergence of $\theta_n^{(t+1)}$ to $\theta_*^{(t)}$ over all possible $\theta_n^{(t)} \in \Theta_n$. However, we can prove that the consistency results hold for any sequence of $\theta_n^{(1)}, \dots, \theta_n^{(T)}$ with T being not too large compared with $\exp(n)$. This is enough for theorems 3–6. To justify this, we may consider the case that the dimension of Θ_n grows with n at a rate of $p = O(n^\gamma)$ for a constant $\gamma > 0$, say $\gamma = 5$. Then it is easy to see that, when $n > 13$, the ratio T/p has an order of

$$O\{\exp(n)/p\} = O[\exp\{n - \gamma \log(n)\}] > O\{\exp(0.1n)\} > O(p^{100}),$$

which implies that essentially there is no constraint on the setting of T . For Markov chain Monte Carlo simulations, the number of iterations is often set to a low order polynomial of p for a given set of observations.

For any $\theta_n^{(t)} \in \Theta_n^T$, we define $\theta_n^{(t+1)}$. We would like to establish the uniform consistency of $\theta_n^{(t+1)}$ with respect to t , i.e.

$$\sup_{t \in \{1, 2, \dots, T\}} \left\| \theta_n^{(t+1)} - \theta_*^{(t)} \right\|_p \rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (19)$$

To achieve this goal, we assume the following condition.

Condition 4. For each $t = 1, 2, \dots, T$, $G_n(\theta | \theta_n^{(t)})$ has a unique maximum at $\theta_*^{(t)}$; for any $\epsilon > 0$, $\sup_{\theta \in \Theta_n \setminus B_t(\epsilon)} G_n(\theta | \theta_n^{(t)})$ exists, where $B_t(\epsilon) = \{\theta \in \Theta_n : \|\theta - \theta_*^{(t)}\| < \epsilon\}$. Let

$$\delta_t = G_n(\theta_*^{(t)} | \theta_n^{(t)}) - \sup_{\theta \in \Theta_n \setminus B_t(\epsilon)} G_n(\theta | \theta_n^{(t)}),$$

$$\delta = \min_{t \in \{1, 2, \dots, T\}} \delta_t > 0.$$

The existence of $\sup_{\theta \in \Theta_n \setminus B_t(\epsilon)} G_n(\theta | \theta_n^{(t)})$ can be easily satisfied if Θ_n is restricted to a compact set, which implies that $\Theta_n \setminus B_t(\epsilon)$ is also a compact set and thus the supremum is achievable. This condition can also be satisfied by assuming that Θ_n is convex and, for each i , $\theta_*^{(i)}$ is in the interior of Θ_n and $G_n(\theta | \theta_n^{(i)})$ is concave in θ .

Theorem 2. Assume that conditions 1–4 hold; then the maximum pseudocomplete-data likelihood estimate $\theta_n^{(t+1)}$ is uniformly consistent to $\theta_*^{(t)}$ over $t = 1, 2, \dots, T$, i.e. expression (19) holds.

Proof. Since both $\hat{G}_n(\theta|\tilde{x}, \theta_n^{(t)})$ and $G_n(\theta|\theta_n^{(t)})$ are continuous in θ as implied by the continuity of $\log\{f_\theta(\tilde{x})\}$, the remaining part of the proof follows from lemma 1 by setting the penalty function $P_{\lambda_n}(\theta) = 0$ for all $\theta \in \Theta_n$.

Lemma 1. Consider a sequence of functions $Q_t(\theta, \mathbf{X}_n)$ for $t = 1, 2, \dots, T$. Suppose that the following conditions are satisfied.

Condition 5. For each t , $Q_t(\theta, \mathbf{X}_n)$ is continuous in θ and there is a function $Q_t^*(\theta)$, which is continuous in θ and uniquely maximized at $\theta_*^{(t)}$.

Condition 6. For any $\epsilon > 0$, $\sup_{\theta \in \Theta_n \setminus B_t(\epsilon)} Q_t^*(\theta)$ exists, where $B_t(\epsilon) = \{\theta: \|\theta - \theta_*^{(t)}\| < \epsilon\}$; let $\delta_t = Q_t^*(\theta_*^{(t)}) - \sup_{\theta \in \Theta_n \setminus B_t(\epsilon)} Q_t^*(\theta)$, $\delta = \min_{t \in \{1, 2, \dots, T\}} \delta_t > 0$.

Condition 7. $\sup_{t \in \{1, 2, \dots, T\}} \sup_{\theta \in \Theta_n} |Q_t(\theta, \mathbf{X}_n) - Q_t^*(\theta)| \rightarrow_p 0$ as $n \rightarrow \infty$.

Condition 8. The penalty function $P_{\lambda_n}(\theta)$ is non-negative and converges to 0 uniformly over the set $\{\theta_*^{(t)}: t = 1, 2, \dots, T\}$ as $n \rightarrow \infty$, where λ_n is a regularization parameter and its value can depend on the sample size n . Let $\theta_n^{(t)} = \arg \max_{\theta \in \Theta_n} \{Q_t(\theta, \mathbf{X}_n) - P_{\lambda_n}(\theta)\}$. Then uniform convergence holds, i.e. $\sup_{t \in \{1, 2, \dots, T\}} \|\theta_n^{(t)} - \theta_*^{(t)}\| \rightarrow_p 0$.

Proof. Consider the two events

- a. $\sup_{t \in \{1, 2, \dots, T\}} \sup_{\theta \in \Theta_n \setminus B_t(\epsilon)} |Q_t(\theta, \mathbf{X}_n) - Q_t^*(\theta)| < \delta/2$ and
- b. $\sup_{t \in \{1, 2, \dots, T\}} \sup_{\theta \in \Theta_n} |Q_t(\theta, \mathbf{X}_n) - Q_t^*(\theta)| < \delta/2$.

From event (a), we can deduce that, for any $t \in \{1, 2, \dots, T\}$ and any $\theta \in \Theta_n \setminus B_t(\epsilon)$, $Q_t(\theta, \mathbf{X}_n) < Q_t^*(\theta) + \delta/2 \leq Q_t^*(\theta_*^{(t)}) - \delta_t + \delta/2 \leq Q_t^*(\theta_*^{(t)}) - \delta/2$. Therefore, $Q_t(\theta_*^{(t)}, \mathbf{X}_n) - P_{\lambda_n}(\theta_*^{(t)}) < Q_t^*(\theta_*^{(t)}) - \delta/2 - o(1)$ by condition 8.

From event (b), we can deduce that, for any $t \in \{1, 2, \dots, T\}$ and any $\theta \in B_t(\epsilon)$, $Q_t(\theta, \mathbf{X}_n) > Q_t^*(\theta) - \delta/2$ and $Q_t(\theta_*^{(t)}, \mathbf{X}_n) > Q_t^*(\theta_*^{(t)}) - \delta/2$. Therefore, $Q_t(\theta_*^{(t)}, \mathbf{X}_n) - P_{\lambda_n}(\theta_*^{(t)}) > Q_t^*(\theta_*^{(t)}) - \delta/2 - o(1)$ by condition 8.

If both events hold simultaneously, then we must have $\theta_n^{(t)} \in B_t(\epsilon)$ for all $t \in \{1, 2, \dots, T\}$ as $n \rightarrow \infty$. By condition 7, the probability that both events hold tends to 1. Therefore,

$$P\left\{\theta_n^{(t)} \in B_t(\epsilon) \text{ for all } t = 1, 2, \dots, T\right\} \rightarrow 1,$$

which concludes the lemma. $\square$

Theorem 2 establishes the consistency of $\theta_n^{(t+1)}$ with respect to $\theta_*^{(t)}$ for each $t = 1, 2, \dots, T$.

However, in the small n -large p scenario, $\theta_n^{(t+1)}$ is not well defined. For this reason, a sparsity constraint needs to be imposed on θ . For example, we can apply a regularization method to obtain an estimate of $\theta_*^{(t)}$, i.e. we can define

$$\theta_{n,p}^{(t+1)} = \arg \max_{\theta \in \Theta_n} \left\{ \widehat{G}_n(\theta) \left| \tilde{\mathbf{x}}, \theta_n^{(t)} \right. - P_{\lambda_n}(\theta) \right\}, \quad (20)$$

where the penalty function $P_{\lambda_n}(\theta)$ constrains the sparsity of the solution. Make the following assumption.

Condition 9. The penalty function $P_{\lambda_n}(\theta)$ is non-negative, ensures the existence of $\theta_{n,p}^{(t+1)}$ for all $n \in \mathbb{N}$ and $t = 1, 2, \dots, T$ and converges to 0 uniformly over the set $\{\theta_*^{(t)} : t = 1, 2, \dots, T\}$ as $n \rightarrow \infty$.

Corollary 1. If conditions 1–4 and 9 hold, then the regularization estimator $\theta_{n,p}^{(t+1)}$ in definition (20) is uniformly consistent to $\theta_*^{(t)}$ over $t = 1, 2, \dots, T$, i.e.

$$\sup_{t \in \{1, 2, \dots, T\}} \left\| \theta_{n,p}^{(t+1)} - \theta_*^{(t)} \right\| \xrightarrow{p} 0 \text{ as } n \rightarrow \infty.$$

The proof of corollary 1 follows the proof of lemma 1 directly.

Take the high dimensional regression as an example. If we allow p to grow with n at the rate $p = O(n^\gamma)$ for some constant $\gamma > 0$, allow the size of $\beta_*^{(t)}$ for all t to grow with n at the rate

$O(n^\alpha)$ for some constant $0 < \alpha < \frac{1}{2}$, choose $\lambda_n = O[\sqrt{\log(p)/n}]$ and set

$$P_{\lambda_n}(\theta) = \lambda_n \sum_{i=1}^p c_{\lambda_n}(|\theta_i|), \text{ where } c_{\lambda_n}(\cdot) \text{ is set in the form of smoothly clipped absolute}$$

deviation (Fan and Li, 2001) or MCP (Zhang, 2010) penalties, then condition 9 is satisfied.

For both the smoothly clipped absolute deviation and MCP penalties, $c_{\lambda_n}(|\theta_i|)$ equals 0 if $\theta_i =$

0 and is bounded by a constant otherwise. Similarly, if the beta-min assumption holds, i.e.

there is a constant $\beta_{\min} > 0$ such that $\min_{j \in S^*} |\beta_{*j}| \geq \beta_{\min}$, where $S^* = \{j: \beta_{*j} \neq 0\}$ denotes

the index set of non-zero regression coefficients, then the reciprocal lasso penalty (Song and

Liang, 2015b) also satisfies condition 8. Note that, if $\Theta = \mathbb{R}^p$, the lasso penalty does not satisfy condition 9, which is unbounded. This explains why the lasso estimate is unbiased even as $n \rightarrow \infty$. However, if Θ_n is restricted to a bounded space, then the lasso penalty also satisfies condition 9.

Alternatively to regularization methods, one may first restrict the space of $\theta_*^{(t)}$ to some low dimensional subspace through sure screening and then find a consistent estimate in the subspace by using conventional statistical methods, such as maximum likelihood, moment estimation or even regularization. Both the ψ -learning (Liang *et al.*, 2015) and the SIS (Fan and Lv, 2008; Fan and Song, 2010) methods belong to this class. For ψ -learning, after correlation screening (based on the pseudocomplete data), the remaining network structure estimation procedure is essentially the same as the covariance selection method (Dempster, 1972) which, by nature, is a maximum likelihood estimation method. It is interesting that the sure-screening-based methods can be viewed as a special subclass of regularization methods, for which the solutions in the low dimensional subspace receive a zero penalty, and those outside the subspace receive a penalty of ∞ . It is easy to see that such a binary-type penalty function satisfies condition 9.

Both the regularization and the sure-screening-based methods are constructive. In what follows, we give a proof for the use of general consistent estimation procedures in the IRO algorithm. Let $\theta_{n,g}^{(t-1)}$ denote the estimate of $\theta_*^{(t)}$ produced by such a general consistent estimation procedure at iteration $t+1$. Corollary 2 shows that, if $\theta_{n,g}^{(t+1)}$ is sufficiently accurate for each t (pointwise) and the log-likelihood function of the pseudocomplete data satisfies some moment conditions, then the estimation procedure can be used in the IRO algorithm. Therefore, by its maximum likelihood estimation nature in the subspace, the use of the ψ -learning algorithm in the IRO algorithm can also be justified by corollary 2.

Condition 10 (conditions for general consistent estimate $\theta_{n,g}^{(t)}$). Assume that, for each $t = 1, 2, \dots, T$, $\theta_{n,g}^{(t+1)} - \theta_*^{(t)} = O_p(1/\sqrt{n})$ (pointwise) and the Hessian matrix $\partial^2 G_n(\theta|\tilde{x}, \theta_n^{(t)})/\partial\theta\partial\theta'$ is bounded in a neighbourhood of $\theta_*^{(t)}$; let

$$Z'_{t,i} = \log \left\{ f(x_i^{\text{obs}}, \tilde{x}_i^{\text{mis}} | \theta_{n,g}^{(t+1)}) \right\} - \int \log \left\{ f(x^{\text{obs}}, \tilde{x}^{\text{mis}} | \theta_{n,g}^{(t+1)}) \right\} f(x^{\text{obs}} | \theta_*) h(\tilde{x}^{\text{mis}} | x_i^{\text{obs}}, \theta_n^{(t)}) dx^{\text{mis}} dx^{\text{obs}};$$

then $E|Z'_{t,i}|^m \leq m! \tilde{M}_b^m - 2\tilde{v}_i/2$ for every $m \geq 2$ and some constants $\tilde{M}_b > 0$ and $\tilde{v}_i = O(1)$.

Corollary 2. Assume conditions 1–4 and 10. Then $\theta_{n,g}^{(t+1)}$ is uniformly consistent to $\theta_*^{(t)}$ over $t = 1, 2, \dots, T$, i.e. $\sup_{t \in \{1, 2, \dots, T\}} \|\theta_{n,g}^{(t+1)} - \theta_*^{(t)}\| \xrightarrow{p} 0$ as $n \rightarrow \infty$.

Proof. Applying Taylor series expansion to $G_n(\theta|\theta_n^{(t)})$ at $\theta_*^{(t)}$, we obtain

$G_n(\theta_{n,g}^{(t+1)}|\theta_n^{(t)}) - G_n(\theta_*^{(t)}|\theta_n^{(t)}) = O_p(1/n)$, following from condition 10 and condition 4 that $G_n(\theta|\theta_n^{(t)})$ is maximized at $\theta_*^{(t)}$. Therefore,

$$n\left\{\widehat{G}_n(\theta_{n,g}^{(t+1)}|\tilde{\mathbf{x}}, \theta_n^{(t)}) - G_n(\theta_*^{(t)}|\theta_n^{(t)})\right\} = Z'_{t,1} + \dots + Z'_{t,n} + n\left\{G_n(\theta_{n,g}^{(t+1)}|\theta_n^{(t)}) - G_n(\theta_*^{(t)}|\theta_n^{(t)})\right\} = Z'_{t,1} + \dots + Z'_{t,n} + \epsilon_n$$

where $\epsilon_n = O_p(1)$, and

$$P\left\{n\left|\widehat{G}_n(\theta_{n,g}^{(t+1)}|\tilde{\mathbf{x}}, \theta_n^{(t)}) - G_n(\theta_*^{(t)}|\theta_n^{(t)})\right| > nz\right\} \leq P\left(|Z'_{t,1} + \dots + Z'_{t,n}| > nz - |\epsilon_n|\right). \quad (21)$$

By Bernstein's inequality,

$$P\left(|Z'_{t,1} + \dots + Z'_{t,n}| > nz - |\epsilon_n|\right) \leq 2\exp\left\{-\frac{1}{2} \frac{(z - |\epsilon_n|/n)^2}{\tilde{v}' + \tilde{M}'_b(z - |\epsilon_n|/n)}\right\}, \quad (22)$$

for $\tilde{v} \geq (\tilde{v}_1 + \dots + \tilde{v}_n)/n^2$ and $\tilde{M}'_b = \tilde{M}_b/n$. Applying Taylor series expansion to the right-hand side of in equality (22) at z and combining with in equality (21) lead to

$$P\left\{n\left|\widehat{G}_n(\theta_{n,g}^{(t+1)}|\tilde{\mathbf{x}}, \theta_n^{(t)}) - G_n(\theta_*^{(t)}|\theta_n^{(t)})\right| > z\right\} \leq K\exp\left\{-\frac{1}{2} \frac{z^2}{\tilde{v}' + \tilde{M}'_bz}\right\}, \quad (23)$$

where $K = 2 + (3/\tilde{M}'_b)O_p(1/n) = 2 + (3/\tilde{M}_b)O_p(1)$, since the derivative

$$\left|d\left\{z^2/(\tilde{v}' + \tilde{M}'_bz)\right\}/dz\right| \leq 3/\tilde{M}'_b.$$

As in the proof of theorem 1, by applying lemma 2.2.10 of van der Vaart and Wellner (1996), we can prove

$$\sup_{\theta_n^{(t)} \in \Theta_n, t \in \{1, 2, \dots, T\}} \left|\widehat{G}_n(\theta_{n,g}^{(t+1)}|\tilde{\mathbf{x}}, \theta_n^{(t)}) - G_n(\theta_*^{(t)}|\theta_n^{(t)})\right| \xrightarrow{p} 0. \quad (24)$$

As implied by the proof of lemma 2.2.10 of van der Vaart and Wellner (1996), result (24) holds for a general constant K in inequality (22). Then, by condition 4, we must have the uniform convergence that $\theta_{n,q}^{(t+1)} \in B_t(\epsilon)$ for all t as $n \rightarrow \infty$, where $B_t(\epsilon)$ is as defined in condition 4. This statement can be proved by contradiction as follows.

Assume that $\theta_{n,q}^{(i+1)} \notin B_i(\epsilon)$ for some $i \in \{1, 2, \dots, T\}$. By the uniform convergence that was established in theorem 1, $\left| \hat{G}_n(\theta_{n,g}^{(i+1)} | \tilde{\mathbf{x}}, \theta_n^{(i)}) - G_n(\theta_{n,g}^{(i+1)} | \theta_n^{(i)}) \right| = o_p(1)$. Further, by condition 4 and the assumption that $\theta_{n,g}^{(i+1)} \notin B_i(\epsilon)$,

$$\begin{aligned} & \left| \hat{G}_n(\theta_{n,g}^{(i+1)} | \tilde{\mathbf{x}}, \theta_n^{(i)}) - G_n(\theta_{n,g}^{(i+1)} | \theta_n^{(i)}) \right| \geq \left| G_n(\theta_{n,g}^{(i+1)} | \theta_n^{(i)}) - G_n(\theta_{n,g}^{(i+1)} | \theta_n^{(i)}) \right| \\ & - \left| \hat{G}_n(\theta_{n,g}^{(i+1)} | \tilde{\mathbf{x}}, \theta_n^{(i)}) - G_n(\theta_{n,g}^{(i+1)} | \theta_n^{(i)}) \right| \geq \delta - o_p(1), \end{aligned}$$

which contradicts the uniform convergence that was established in result (24). This concludes the proof. $\square$

Remark 3 (on the accuracy of $\theta_{n,g}^{(t)}$ s). Condition 10 restricts the consistent estimates to those having a distance to the true parameter point of the order $O_p(1/\sqrt{n})$. Such a condition can be satisfied by some estimation procedures in the low dimensional subspace, e.g. maximum likelihood, for which both the variance and the bias are often of the order $O(1/n)$ (Firth, 1993) and therefore the root-mean-squared error is of the order $O(1/\sqrt{n})$. To make the result of corollary 2 more general to include more estimation procedures, we can relax this order to $\theta_{n,g}^{(t+1)} - \theta_{n,g}^{(t)} = O_p(n^{-1/4})$, if we would like to relax the order of T to $\log(T) = o(\sqrt{n})$ and the order of metric entropy to $\log[N\{\epsilon, \mathcal{G}_{n,M}, L_1(\mathbb{P}_n)\}] = o_p^*(\sqrt{n})$. As mentioned in remarks 1 and 2, both the order of T and the order of metric entropy are technical conditions and relaxing them to the order of $O(\sqrt{n})$ will not restrict much the applications of the IRO algorithm. The proof for this relaxation is straightforward, following the proof of corollary 2.

A.2. Proof of ergodicity of the Markov chain $\{\theta_n^{(t)}\}$

Although the IRO algorithm is different from the SEM algorithm in the $\theta_n^{(t)}$ updating step, the Markov chains $\{\theta_n^{(t)}\}$ that are induced by the two algorithms share some similar properties as well as similar proofs. The following two lemmas, lemma 2 and lemma 3, can be proved in the same way as in Nielsen (2000), and thus the proofs have been omitted.

Lemma 2. The Markov chain $\{\theta_n^{(t)}\}$ is irreducible and aperiodic.

Lemma 3. If condition 1 holds, then the Markov chain has the weak Feller property, and any compact subsets of Θ are small.

If condition 1 holds and Θ_n is restricted to a compact set, then the Markov chain $\{\theta_n^{(t)}\}$ is ergodic. Here we would like to establish the ergodicity of the Markov chain $\{\theta_n^{(t)}\}$ under a more general scenario $\Theta_n = \mathbb{R}^p$. This can be done by verifying a drift condition. Similarly to Nielsen (2000), we choose the negative log-likelihood function of the observed data as the drift function, motivated by the drift in the EM algorithm towards high density areas.

Theorem 3. If conditions 1–3 hold, then $\{\theta_n^{(t)}\}$ is almost surely ergodic for sufficiently large n .

Proof. Let $v(\theta) = C - (1/n)\log f\{x_1^{\text{obs}}, \dots, x_n^{\text{obs}} | \theta\}$, where C denotes a constant such that $v(\theta) \geq 0$ for all $\theta \in \Theta_n$. Since $v(\theta)$ is non-negative, it can be used to build the drift condition. Define

$$\begin{aligned} \Delta v(\theta) &= E_h[v(\theta_n^{(t+1)}) - v(\theta_n^{(t)})] = E_h\left[\frac{1}{n}\log\{f(x^{\text{obs}} | \theta_n^{(t)})\} - \frac{1}{n}\log\{f(x^{\text{obs}} | \theta_n^{(t+1)})\}\right] \\ &= E_h\left[\frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t)})\} - \frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t+1)})\}\right] - E_h\left[\frac{1}{n}\log\{h(\tilde{x}^{\text{mis}} | x^{\text{obs}}, \theta_n^{(t)})\} - \frac{1}{n}\log\{h(\tilde{x}^{\text{mis}} | x^{\text{obs}}, \theta_n^{(t+1)})\}\right] \\ &= \text{I} + \text{II}, \end{aligned}$$

where E_h refers to the expectation with respect to the predictive distribution $h(\tilde{x}^{\text{mis}} | x^{\text{obs}}, \theta_n^{(t)})$.

First, we consider the negative of part I, which can be decomposed as

$$\begin{aligned} -\text{I} &= E_h\left[\frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t+1)})\} - \frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t)})\}\right] \\ &= E_h\left[\frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t+1)})\} - \frac{1}{n}\log\{f(\tilde{x} | M(\theta_n^{(t)}))\}\right] + E_h\left[\frac{1}{n}\log\{f(\tilde{x} | M(\theta_n^{(t)}))\} - \frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t)})\}\right], \end{aligned}$$

where the function $M(\theta)$ is defined by

$$M(\theta) = \arg \max_{\theta'} E_{\theta}[\log\{f(\tilde{x} | \theta')\}] = \arg \max_{\theta'} \int f(x^{\text{obs}}, \tilde{x}^{\text{mis}} | \theta') f(x^{\text{obs}} | \theta^*) h(\tilde{x}^{\text{mis}} | x^{\text{obs}}, \theta) d\tilde{x}^{\text{mis}} dx^{\text{obs}}. \quad (25)$$

By the uniform law of large numbers that was established in theorem 1, we have

$$\frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t+1)})\} - \frac{1}{n}\log\{f(\tilde{x} | M(\theta_n^{(t)}))\} \xrightarrow{p} G_n(\theta_n^{(t+1)} | \theta_n^{(t)}) - G_n\{M(\theta_n^{(t)}) | \theta_n^{(t)}\}.$$

From theorem 2, we have $\theta_n^{(t+1)} - M(\theta_n^{(t)}) \rightarrow_p 0$. Further, by the continuity of $G_n(\theta | \theta)$ with respect to θ , we have $G_n(\theta_n^{(t+1)} | \theta_n^{(t)}) - G_n\{M(\theta_n^{(t)}) | \theta_n^{(t)}\} \rightarrow_p 0$ and thus $(1/n)\log\{f(\tilde{x} | \theta_n^{(t+1)})\} - (1/n)\log\{f(\tilde{x} | M(\theta_n^{(t)}))\} \rightarrow_p 0$. Then, by the boundedness of $\log\{f(\tilde{x} | \theta)\}$ (condition 2) and the dominated convergence theorem,

$$E\left[\frac{1}{n}\log\{f(\tilde{x} | \theta_n^{(t+1)})\} - \frac{1}{n}\log\{f(\tilde{x} | M(\theta_n^{(t)}))\}\right] \rightarrow 0, \quad (26)$$

where the expectation is with respect to the joint density function of $\tilde{x} = (x^{\text{obs}}, \tilde{x}^{\text{mis}})$. Note that, for any $\theta \in \Theta_n$, we have

$$E_h \left[\frac{1}{n} \log \left\{ f(\tilde{\mathbf{x}} | \boldsymbol{\theta}) \right\} \right] = \frac{1}{n} \sum_{i=1}^n E_h [\log \{ f(\tilde{x}_i | \boldsymbol{\theta}) \}] \triangleq \frac{1}{n} \sum_{i=1}^n g(x_i^{\text{obs}}), \quad (27)$$

where the $g(x_i^{\text{obs}})$ are mutually independent, but not necessarily identically distributed because of the presence of missing data. Then, by result (26), condition 2 and Kolmogorov's strong law of large numbers, we have

$$E_h \left[\frac{1}{n} \log \left\{ f(\tilde{\mathbf{x}} | \boldsymbol{\theta}_n^{(t+1)}) \right\} - \frac{1}{n} \log [f(\tilde{\mathbf{x}} | M(\boldsymbol{\theta}_n^{(t)}))] \right] \rightarrow 0, \quad \text{almost surely,} \quad (28)$$

as $n \rightarrow \infty$. Therefore, there is a constant $c > 0$ and a large number N such that

$$-c < E_h \left[\frac{1}{n} \log \left\{ f(\tilde{\mathbf{x}} | \boldsymbol{\theta}_n^{(t+1)}) \right\} - \frac{1}{n} \log [f(\tilde{\mathbf{x}} | M(\boldsymbol{\theta}_n^{(t)}))] \right] < c, \quad \text{almost surely,} \quad (29)$$

for any $n > N$ and any $\epsilon > 0$. With a similar argument to that for equation (27), by invoking Kolmogorov's strong law of large numbers, it can be shown that there is a constant $\delta > 0$ such that

$$E_h \left[\frac{1}{n} \log [f(\tilde{\mathbf{x}} | M(\boldsymbol{\theta}_n^{(t)}))] - \frac{1}{n} \log \left\{ f(\tilde{\mathbf{x}} | \boldsymbol{\theta}_n^{(t)}) \right\} \right] \rightarrow \delta, \quad \text{almost surely,} \quad (30)$$

for any $\epsilon > 0$ as $n \rightarrow \infty$. Combining expressions (29) and (30), we have $-c - \delta < I < c$ holds almost surely for sufficiently large n .

Next, by Jensen's inequality, we have

$$\begin{aligned} \text{II} &= E_h \left[\frac{1}{n} \log \left\{ h(\tilde{\mathbf{x}}^{\text{mis}} | \mathbf{x}^{\text{obs}}, \boldsymbol{\theta}_n^{(t+1)}) \right\} - \frac{1}{n} \log \left\{ h(\tilde{\mathbf{x}}^{\text{mis}} | \mathbf{x}^{\text{obs}}, \boldsymbol{\theta}_n^{(t)}) \right\} \right] \leq \frac{1}{n} \log \left(E_h \left[\frac{h(\tilde{\mathbf{x}}^{\text{mis}} | \mathbf{x}^{\text{obs}}, \boldsymbol{\theta}_n^{(t+1)})}{h(\tilde{\mathbf{x}}^{\text{mis}} | \mathbf{x}^{\text{obs}}, \boldsymbol{\theta}_n^{(t)})} \right] \right) \\ &= \frac{1}{n} \log \left(\int h(\tilde{\mathbf{x}}^{\text{mis}} | \mathbf{x}^{\text{obs}}, \boldsymbol{\theta}_n^{(t+1)}) d\tilde{\mathbf{x}}^{\text{mis}} \right) = 0. \end{aligned}$$

Combining the results of I and II, we have that $v(\boldsymbol{\theta}) < c$ almost surely for all $\boldsymbol{\theta} \in \Theta_n$. Choose b as a positive number less than $c + \delta$ and D as a compact set including $\{\boldsymbol{\theta} \in \Theta_n : v(\boldsymbol{\theta}) \in [-b, c]\}$. In summary, we have

$$\Delta v(\boldsymbol{\theta}) \leq \begin{cases} c, & \boldsymbol{\theta} \in D \\ -b, & \boldsymbol{\theta} \in \Theta_n \setminus D, \end{cases}$$

almost surely. Hence, the strict drift condition V_2 (Meyn and Tweedie (2009), page 263) is almost surely satisfied.

Since $(\theta_n^{(t)})_{t \in \mathbb{N}_0}$ also has weak Feller property (see lemma 3), we can further conclude that an invariant probability measure π almost surely exists for this Markov chain (Meyn and Tweedie (2009), theorem 12.3.4). Since $(\theta_n^{(t)})_{t \in \mathbb{N}_0}$ is irreducible (shown in lemma 2), $\mathcal{D}$ is a compact set and thus a small set (shown in lemma 3), and the drift condition V_2 is stronger than the drift condition V_1 (Meyn and Tweedie (2009), page 189), we can show that $(\theta_n^{(t)})_{t \in \mathbb{N}_0}$ is Harris recurrent (Meyn and Tweedie (2009), theorem 9.1.8). Since $(\theta_n^{(t)})_{t \in \mathbb{N}_0}$ is irreducible and has an invariant probability measure π , it is also a positive chain (Meyn and Tweedie (2009), page 230). Therefore, it is a positive Harris recurrent chain (Meyn and Tweedie (2009), page 231). Finally, since $(\theta_n^{(t)})_{t \in \mathbb{N}_0}$ is aperiodic (shown in lemma 2) and positive Harris recurrent, we can conclude that it is almost surely ergodic (Meyn and Tweedie (2009), theorem 13.3.3).

A.3. Proof of consistency of the imputation–regularized optimization estimator

To prove the consistency of the IRO estimator, we consider the mapping that is defined in equation (25). For the RO step, we have $\theta_*^{(t)} = M(\theta_n^{(t)})$. Also, θ^* , the true value of θ_{π} is a fixed point of the mapping. Further, to show that the mean of the stationary distribution of the Markov chain forms a consistent estimate of θ^* , we make the following assumption.

Condition 11. The mapping $M(\theta)$ is differentiable. Let $\lambda_n(\theta)$ be the largest singular value of $M(\theta) - \theta$. There is a number $\lambda^* < 1$ such that $\lambda_n(\theta) \leq \lambda^*$ for all $\theta \in \Theta_n$ for sufficiently large n and an almost every $\mathbf{x}^{\text{obs}}$ -sequence.

Remark 4 (on contraction mapping). Condition 11 directly implies that

$$\|M(\theta_n^{(t)}) - \theta^*\| = \|M(\theta_n^{(t)}) - M(\theta^*)\| \leq \lambda^* \|\theta_n^{(t)} - \theta^*\|, \quad (31)$$

i.e. the mapping is a contraction. We note that a continuous application in the mapping, i.e. setting $\theta_n^{(t+1)} = \theta_*^{(t)} = M(\theta_n^{(t)})$ for all t , leads to a monotone increase in the expectation

$E_{\theta_n^{(t)}}[\log\{f_{\theta}(\tilde{\mathbf{x}})\}]$. Since $E_{\theta_n^{(t)}}[\log\{f_{\theta}(\tilde{\mathbf{x}})\}]$ attains its maximum at $E_{\theta^*}[\log\{f_{\theta^*}(\tilde{\mathbf{x}})\}]$, it is

reasonable to assume that $M(\theta_n^{(t)})$ is closer to θ^* than $\theta_n^{(t)}$. This condition should hold for sufficiently large n , at which the $\theta_*^{(t)}$ and θ^* are all unique as assumed in condition 4. We note that a similar contraction condition has been used in analysis of the SEM algorithm (proposition 3, Nielsen (2000)). Some other conditions can potentially be specified based on fixed point theory (see, for example, Khamsi and Kirk (2001)).

Theorem 4. Assume conditions 1–4 and 11, and $\sup_{n,t} E\|\theta_n^{(t)}\| < \infty$. Then, for sufficiently large n , sufficiently large t and an almost every $\mathbf{x}^{\text{obs}}$ -sequence, $\|\theta_n^{(t)} - \theta^*\| = o_p(1)$. Furthermore, the sample average of the Markov chain forms a consistent estimate of θ^* , i.e. $\|(1/T)\sum_{t=1}^T \theta_n^{(t)} - \theta^*\| = o_p(1)$, as $n \rightarrow \infty$ and $T \rightarrow \infty$.

Proof. By theorem 3, the Markov chain $\{\theta_n^{(t)}\}$ converges to a stationary distribution. For simplicity, we suppress the superscript t , let θ_n denote the current sample and let θ'_n denote the next iteration sample. Therefore,
 $\|\theta'_n - \theta^*\| \leq \|\theta'_n - M(\theta_n)\| + \|M(\theta_n) - \theta^*\| \leq \|\theta'_n - M(\theta_n)\| + \lambda^* \|\theta_n - \theta^*\|$, where the last inequality follows from result (31). Taking expectation on both sides leads to

$$E\|\theta'_n - \theta^*\| \leq E\|\theta'_n - M(\theta_n)\| + \lambda^* E\|\theta_n - \theta^*\| \leq \frac{1}{1-\lambda^*} E\|\theta'_n - M(\theta_n)\| = \frac{1}{1-\lambda^*} o(1) = o(1),$$

(32)

where the second in equality follows from the stationarity of the Markov chain, and the first equality follows from theorem 2 and the existence of $E\|\theta_n\|$. Finally, by Markov's inequality, we conclude the consistency of $\theta_n^{(t)}$ as an estimator of θ^* .

By result (32), we have $\|E(\theta_n) - \theta^*\| = E\|\theta_n - \theta^*\| = o(1)$, which implies that the mean of the stationary distribution of $\{\theta_n^{(t)}\}$ converges to θ^* for sufficiently large n . Further, by the ergodicity of the Markov chain $\{\theta_n^{(t)}\}$, we conclude the proof. $\square$

Corollary 3. Assume conditions 1–4 and 11, $\sup_{n,t} E\|\theta_n^{(t)}\| < \infty$, $h(\theta)$ is a Lipschitz function on Θ_n and $\sup_{n,t} E\|\theta_n^{(t)}\| < \infty$. Then for sufficiently large n , sufficiently large t and an almost every $\mathbf{x}^{\text{obs}}$ -sequence, $\|h(\theta_n^{(t)}) - h(\theta^*)\| = o_p(1)$. Furthermore,
 $\|(1/T)\sum_{t=1}^T (h(\theta_n^{(t)}) - h(\theta^*))\| = o_p(1)$, as $n \rightarrow \infty$ and $T \rightarrow \infty$.

Proof. The proof follows from the definition of a Lipschitz function and the proof of theorem 4. $\square$

A.4. Proof of ergodicity of the Markov chain for the imputation–conditional regularized optimization algorithm

Theorem 5. If conditions 1–3 hold, the Markov chain $\{(\theta_n^{(t,1)}, \dots, \theta_n^{(t,k)})\}$ is almost surely ergodic for sufficiently large n .

Theorem 5 can be proved in a similar way to theorem 3 with the detail given in the on-line supplementary material.

A.5. Proof of consistency of the imputation–conditional regularized optimization estimator

Condition 11'. Let M_i denote the mapping of the i th part of the CRO step, i.e. $\theta_n^{(t,i)} = M_i(\theta_n^{(t+1,1)}, \dots, \theta_n^{(t+1,i-1)}, \theta_n^{(t,i)}, \dots, \theta_n^{(t,k)})$. Let $M = M_k \circ M_{k-1} \circ \dots \circ M_1$ denote the joint mapping of $M_1, \dots, M_k$. Let $\lambda_n(\theta)$ denote the largest singular value of $M(\theta)/\theta$. There is a number $\lambda^* < 1$ such that $\lambda_n(\theta) \leq \lambda^*$ for all $\theta \in \Theta_n$, all sufficiently large n and an almost every $\mathbf{x}^{\text{obs}}$ -sequence.

This condition is reasonable: it is easy to see that a continuous application of the mapping M , i.e. applying M in a circular manner, leads to a monotone increase in the function $E_{\theta_n^{(t)}}[\log\{f_{\theta}(\tilde{\mathbf{X}})\}]$. Similarly to theorem 4, we can prove the following theorem with the detail given in the on-line supplementary material.

Theorem 6. Assume conditions 1–4 and 11', and $\sup_{n,t} E\|\theta_n^{(t)}\| < \infty$. Then for sufficiently large n , sufficiently large t and an almost every $\mathbf{x}^{\text{obs}}$ -sequence, $\|\theta_n^{(t)} - \theta^*\| = o_p(1)$. Furthermore, the sample average of the Markov chain also forms a consistent estimate of θ^* , i.e. $\|(1/T)\sum_{t=1}^T \theta_n^{(t)} - \theta^*\| = o_p(1)$, as $n \rightarrow \infty$ and $T \rightarrow \infty$.

Corollary 4. Assume conditions 1–4 and 11', $\sup_{n,t} E\|\theta_n^{(t)}\| < \infty$, $h(\theta)$ is a Lipschitz function on Θ_n and $\sup_{n,t} E\|\theta_n^{(t)}\| < \infty$. Then for sufficiently large n , sufficiently large t and an almost every $\mathbf{x}^{\text{obs}}$ -sequence, $\|h(\theta_n^{(t)}) - h(\theta^*)\| = o_p(1)$. Furthermore, $\|(1/T)\sum_{t=1}^T h(\theta_n^{(t)}) - h(\theta^*)\| = o_p(1)$, as $n \rightarrow \infty$ and $T \rightarrow \infty$.

Proof. The proof follows from the definition of a Lipschitz function and the proof of theorem 6.

References

- Besag J (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). J. R. Statist. Soc. B, 36, 192–236.
- Bo TH, Dysvik B and Jonassen I (2004) LSImpute: accurate estimation of missing values in microarray data with least square methods. Nucleic Acids Res, 32, article e34. [PubMed: 14978222]
- Burgette LF and Reiter JP (2010) Multiple imputation for missing data via sequential regression trees. Am. J. Epidemiol, 172, 1070–1076.
- van Buuren S and Groothuis-Oudshoorn K (2011) mice: multivariate imputation by chained equations in R. J. Statist. Softw, 45, no. 3.
- Cai J-F, Candes E and Shen Z (2010) A singular value thresholding algorithm for matrix completion. SIAM J. Optimizn, 20, 1956–1982.
- Castillo I, Schmidt-Hieber J and der Vaart AW (2015) Bayesian linear regression with sparse priors. Ann. Statist, 43, 1986–2018.

- Celeux G and Diebolt J (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computnl Statist. Q*, 2, 73–82.
- Dempster AP (1972) Covariance selection. *Biometrics*, 28, 157–175.
- Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39, 1–38.
- Dobra A, Hans C, Jones B, Nevins JR, Yao G and West M (2004) Sparse graphical models for exploring gene expression data. *J. Multiv. Anal*, 90, 196–212.
- Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Statist. Ass*, 99, 96–104.
- Fan J and Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass*, 96, 1348–1360.
- Fan Y and Li R (2012) Variable selection in linear mixed effects models. *Ann. Statist*, 40, 2043–2068.
- Fan J and Lv J (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B*, 70, 849–911.
- Fan J and Song R (2010) Sure independence screening in generalized linear model with NP-dimensionality. *Ann. Statist*, 38, 3567–3604.
- Firth D (1993) Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27–38.
- Friedman J, Hastie T and Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432–441. [PubMed: 18079126]
- Garcia RI, Ibrahim JG and Zhü H (2010) Variable selection for regression models with missing data. *Statist. Sin*, 20, 149–165.
- Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D and Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Molec. Biol. Cell*, 11, 4241–4257. [PubMed: 11102521]
- van de Geer S, Bühlmann P, Ritov Y and Dezeure R (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist*, 42, 1166–1202.
- Gelman A and Rubin DB (1992) Inference from iterative simulation using multiple sequences. *Statist. Sci*, 7, 457–472.
- Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning*, 2nd edn. New York: Springer.
- He S (2011) Extension of SPACE: R package ‘SpaceExt’. (Available from <http://cran.r-project.org/web/packages/spaceExt>.)
- He Y and Liu C (2012) The dynamic ‘expectation-conditional maximization either’ algorithm. *J. R. Statist. Soc. B*, 74, 313–336.
- Johnson VE and Rossell D (2012) Bayesian model selection in high-dimensional settings. *J. Am. Statist. Ass*, 107, 649–660.
- Khalili A and Chen J (2007) Variable selection in finite mixture of regression models. *J. Am. Statist. Ass*, 102, 1025–1038.
- Khamsi MA and Kirk WA (2000) *An Introduction to Metric Spaces and Fixed Point Theory*. Hoboken: Wiley.
- Liang F, Song Q and Qiü P (2015) An equivalent measure of partial correlation coefficients for high-dimensional Gaussian graphical models. *J. Am. Statist. Ass*, 110, 1248–1265.
- Liang F and Zhang J (2008) Estimating the false discovery rate using the stochastic approximation algorithm. *Biometrika*, 95, 961–977.
- Liu C and Rubin DB (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81, 633–648.
- Liu C, Rubin DB and Wu YN (1998) Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, 85, 755–770.
- Long Q and Johnson BA (2015) Variable selection in the presence of missing data: resampling and imputation. *Biostatistics*, 16, 596–610. [PubMed: 25694614]
- Mazumder R and Hastie T (2012) The graphical lasso: new insights and alternatives. *Electron. J. Statist*, 6, 2125–2149.

- Mazumder R, Hastie T and Tibshirani R (2010) Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 99, 2287–2322.
- McLachlan GJ and Krishnan T (2008) *The EM Algorithm and Extensions*, 2nd edn. Hoboken: Wiley.
- Meinshausen N and Bühlmann P (2006) High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34, 1436–1462.
- Meinshausen N and Bühlmann P (2010) Stability selection (with discussion). *J. R. Statist. Soc. B*, 72, 417–473.
- Meng X-L and Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80, 267–278.
- Meyn S and Tweedie RL (2009) *Markov Chains and Stochastic Stability*, 2nd edn. Cambridge: Cambridge University Press.
- Nielsen SF (2000) The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6, 457–489.
- Oba S, Sato M, Takemasa I, Monden M, Matsubara K-I and Ishii S (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19, 2088–2096. [PubMed: 14594714]
- Ouyang M, Welsh WJ and Georgopoulos P (2004) Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, 20, 917–923. [PubMed: 14751970]
- Raskutti G, Wainwright MJ and Yu B (2011) Minimax rates of estimation for high-dimensional linear regression over l_q -balls. *IEEE Trans. Inform. Theory*, 57, 6976–6994.
- Scheetz TE, Kim K-Y, Swiderski RE, Philp AR, Braun TA, Knudtson KL, Dorrance AM, Di Bona GF, Huang J, Casavant TL, Sheffield VC and Stone EM (2006) Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natn. Acad. Sci. USA*, 103, 1442914434.
- Song Q and Liang F (2015a) A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *J. R. Statist. Soc. B*, 77, 947–972.
- Song Q and Liang F (2015b) High dimensional variable selection with reciprocal L_1 -regularization. *J. Am. Statist. Ass.*, 110, 1607–1620.
- Stacklies W, Redestig H, Scholz M, Walther D and Selbig J (2007) *pcaMethods*—a Bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, 23, 1164–1167. [PubMed: 17344241]
- Städler N and Bühlmann P (2012) Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statist. Comput.*, 22, 219–235.
- Städler N, Stekhoven DJ and Bühlmann P (2014) Pattern alternating maximization algorithm for missing data in high-dimensional problems. *J. Mach. Learn. Res.*, 15, 1903–1928.
- Storey JD (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64, 479–498.
- Tanner MA and Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, 82, 528–540.
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58, 267–288.
- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D and Altman R (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17, 520–525. [PubMed: 11395428]
- Tseng P (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optimizn Theory Appl.*, 109, 475–494.
- Tseng P and Yun S (2009) A coordinate gradient descent method for nonsmooth separable minimization. *Math. Programmg.*, 117, 387–423.
- van der Vaart AW and Wellner JA (1996) *Weak Convergence and Empirical Processes*. New York: Springer.
- Vershynin R (2015) Estimation in high dimensions: a geometric perspective In *Sampling Theory, a Renaissance* (ed. Pfander G), pp. 3–66. Cham: Birkhäuser.
- Wei GCG and Tanner MA (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Ass.*, 85, 699–704.
- Wu CFJ (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, 11, 95–103.

- Yu G and Liu Y (2016) Sparse regression incorporating graphical structure among predictors. *J. Am. Statist. Ass.*, 111,707–720.
- Yuan M and Lin Y (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94, 19–35.
- Zhang C-H (2010) Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38,894–942.
- Zhang C-H and Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B*, 76, 217–242.
- Zhao Y and Long Q (2016) Multiple imputation in the presence of high-dimensional data. *Statist. Meth. Med. Res.*, 25, 2021–2035.

Fig. 1.

Precision–recall curves resulting from various imputation methods for one simulated data set with $p = 400$: —, curve obtained with complete data ('true'); —, curve produced by the misgLasso algorithm; - - -, curve obtained with the ψ -scores generated in the last IRO iteration; - - -, curve obtained with the ψ -scores averaged over the last 20 IRO iterations; - - -, curve obtained with missing values imputed by the median filling method; ..., curve obtained with missing data imputed by the BPCA method; ..., curve obtained with missing data imputed by the regression tree (multiple-imputation) method

Table 1.

Average areas (over 10 data sets) under the precision–recall curves resulting from various imputation methods[‡]

<i>p</i>	<i>Results for the following methods:</i>						
	<i>misgLasso</i>	<i>Median</i>	<i>BPCA</i>	<i>MIRegTree</i>	<i>IRO-Last</i>	<i>IRO-Ave</i>	<i>True</i>
100	0.678 (0.006)	0.882 (0.007)	0.874 (0.006)	0.877 (0.005)	0.877 (0.007)	0.904 (0.006)	0.949 (0.006)
200	0.633 (0.005)	0.856 (0.004)	0.855 (0.004)	0.844 (0.006)	0.887 (0.004)	0.902 (0.003)	0.941 (0.002)
300	0.599 (0.004)	0.830 (0.003)	0.833 (0.003)	0.807 (0.006)	0.869 (0.003)	0.901 (0.002)	0.936 (0.002)
400	0.580 (0.003)	0.824 (0.003)	0.824 (0.003)	0.799 (0.005)	0.868 (0.003)	0.900 (0.002)	0.932 (0.001)

[‡]Numbers in parentheses denote the standard deviation of the average.

Table 2.

Comparison of the ICRO algorithm with the median filling, BPCA and regression tree (with multiple imputation) methods for high dimensional variable selection with independent covariates[†]

<i>p</i>	<i>MR (%)</i>		<i>Results for the following methods:</i>				
			<i>BPCA</i>	<i>Median</i>	<i>MIRegTree</i>	<i>ICRO</i>	<i>True</i>
200	5	err_{β}^2	0.257 (0.267)	0.262 (0.261)	0.333 (0.08)	0.042(0.041)	0.046 (0.048)
		fsr	0.119(0.143)	0.082 (0.092)	0.075 (0.038)	0(0)	0(0)
		nsr	0(0)	0(0)	0(0)	0(0)	0(0)
	10	err_{β}^2	0.903 (0.396)	0.856 (0.421)	0.369(0.189)	0.065 (0.087)	0.046 (0.048)
		fsr	0.310(0.159)	0.308(0.178)	0.057 (0.031)	0(0)	0(0)
		nsr	0(0)	0(0)	0.016(0.016)	0(0)	0(0)
	500	err_{β}^2	0.339 (0.214)	0.350 (0.206)	0.254(0.121)	0.029 (0.034)	0.027 (0.023)
		fsr	0.249 (0.225)	0.266 (0.237)	0.142(0.117)	0(0)	0(0)
		nsr	0(0)	0(0)	0(0)	0(0)	0(0)
	10	err_{β}^2	1.532(1.071)	1.354(0.895)	1.075(0.149)	0.044 (0.022)	0.027 (0.023)
		fsr	0.470 (0.265)	0.420 (0.255)	0.083 (0.037)	0(0)	0(0)
		nsr	0.033 (0.070)	0.017 (0.053)	0.083 (0.037)	0(0)	0(0)

[†]MR denotes the missingness rate and true denotes the results obtained by the MCP method from the complete data. The values in the table are obtained by averaging over 10 independent data sets with the standard deviation reported in the parentheses.

Table 3.

Comparison of the ICRO algorithm with the median filling, BPCA and regression tree (with multiple imputation) methods for high dimensional variable selection with dependent covariates[†]

<i>p</i>	<i>MR (%)</i>		<i>Results for the following methods:</i>				
			<i>BPCA</i>	<i>Median</i>	<i>MIRegTree</i>	<i>ICRO</i>	<i>True</i>
200	5	err_{β}^2	0.580(0.413)	0.548(0.140)	0.537 (0.269)	0.118(0.097)	0.071 (0.050)
		fsr	0.262 (0.204)	0.263 (0.200)	0.076 (0.033)	0(0)	0(0)
		nsr	0.017 (0.052)	0.017(0.052)	0.083 (0.027)	0(0)	0(0)
	10	err_{β}^2	1.604(0.666)	1.575(0.974)	1.550(0.462)	0.424 (0.461)	0.071 (0.050)
		fsr	0.247 (0.229)	0.273 (0.238)	0.031 (0.021)	0(0)	0(0)
		nsr	0.100(0.086)	0.083 (0.088)	0.167(0.025)	0.033 (0.070)	0(0)
	500	err_{β}^2	0.669 (0.366)	0.717(0.358)	0.555(0.134)	0.172(0.195)	0.096 (0.083)
		fsr	0.262 (0.202)	0.289 (0.236)	0.083 (0.037)	0(0)	0(0)
		nsr	0.017(0.053)	0.017(0.053)	0.167(0.144)	0(0)	0(0)
	10	err_{β}^2	2.752 (2.306)	2.896(2.601)	1.433(1.151)	0.578 (0.587)	0.096 (0.083)
		fsr	0.297 (0.230)	0.327 (0.224)	0.083 (0.037)	0(0)	0(0)
		nsr	0.133(0.070)	0.133(0.070)	0.167(0.125)	0.050(0.081)	0(0)

[†]MR denotes the missingness rate, and true denotes the results obtained by the MCP method from the complete data.

Table 4.

Estimation errors of $\hat{\beta}$ (with respect to β_c) produced by the ICRO, median filling, BPCA and regression tree (with multiple imputation) methods for the Bardet–Biedl syndrome example[†]

<i>Method</i>	err_{β}^2	<i>sd</i>
BPCA	0.428	0.091
Median	0.397	0.086
MIRegTree	0.289	0.042
ICRO	0.187	0.040

[†] err_{β}^2 is calculated by averaging $\left\|\hat{\beta} - \beta_c\right\|^2$ over 10 incomplete data sets, and sd represents the standard deviation of err_{β}^2 .