Estimating Sparse Precision Matrices from Data with Missing Values

Mladen Kolar MLADENK@CS.CMU.EDU
Eric P. Xing EPXING@CS.CMU.EDU
Machine Learning Department, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA

Abstract

We study a simple two step procedure for estimating sparse precision matrices from data with missing values, which is tractable in high-dimensions and does not require imputation of the missing values. We provide rates of convergence for this estimator in the spectral norm, Frobenius norm and element-wise ℓ_∞ norm. Simulation studies show that this estimator compares favorably with the EM algorithm. Our results have important practical consequences as they show that standard tools for estimating sparse precision matrices can be used when data contains missing values, without resorting to the iterative EM algorithm that can be slow to converge in practice for large problems.

1. Introduction

Covariance matrices and their inverses, precision matrices, arise in a number of applications including principal component analysis, classification by linear and quadratic discriminant analysis, and the identification of conditional independence assumptions in the context of Gaussian graphical models. As a result, obtaining good estimators of covariance and precision matrices under various contexts is of essential importance in statistics and machine learning research. In the context of Gaussian Markov Random Fields (MRFs), the graph structure encodes certain conditional independence assumptions; if variables corresponding to nodes a and b are conditionally independent given the remaining variables, then there is no edge between nodes a and b. As a precision matrix parametrizes a Gaussian MRF and a zero element in the precision matrix implies that two variables are conditionally independent, the problem of estimating precision matrices commonly arises in the context of learning the structure and parameters of Gaussian MRFs. The availability of high-dimensional data, where the sample size n can be small relative to the dimension p, has pushed the focus of research towards methods for estimating sparse precision matrices under proper regularizations. See, for example, Meinshausen & Bühlmann (2006), Peng et al. (2009), Cai et al. (2010), Ravikumar et al. (2008), Rothman et al. (2008), and Yuan & Lin (2007). Many theoretical results have been obtained for the high-dimensional problems, including consistency and rate of convergence results under a variety of assumptions, as well as efficient algorithms to numerically find estimates. However, all of the above approaches have been devised to deal with the case where all data are fully observed.

In practice, we often have to analyze data that contains missing values (Little & Rubin, 1987). Missing values may occur due to a number of reasons, for example, faulty machinery that collects data, subjects not being available in subsequent experiments (longitudinal studies), limits from experimental design, etc. When missing values are present, they are usually imputed to obtain a complete data set on which standard methods can be applied. However, methods that directly perform statistical inference, without imputing missing values, are preferred. A systematic approach to missing values problem is based on likelihoods of observed values. However, with an arbitrary pattern of missing values, no explicit maximization of the likelihood is possible even for the mean values and covariance matrices (Little & Rubin, 1987). Expectation maximization algorithms, which are iterative methods, are commonly used in cases where explicit maximization of the likelihood is not possible; however, providing theoretical guarantees for such procedures is difficult. This approach was employed in Städler & Bühlmann (2009) to estimate sparse inverse covariance matrices, which we will review in the following section. In recent work, Lounici (2012) deals with the estimation of covariance matrices from data with missing values under the assumption that the true covariance matrix

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

is approximately low rank. Loh & Wainwright (2011) recently studied high-dimensional regression problems when data contains missing values. Casting the estimation of a precision matrix as a sequence of regression problems, they obtain an estimator of the precision matrix without maximizing partially observed likelihood function using an EM algorithm.

In this work, we present a simple, principled method that directly estimates a large dimensional precision matrix from data with missing values. We form an unbiased estimator of the covariance matrix from available data, which is then plugged into the penalized maximum likelihood objective for a multivariate Normal distribution to obtain a sparse estimator of the precision matrix. Even though the initial estimator of the covariance matrix is not necessarily positive-definite, we can show that the final estimator of the precision matrix is positive definite. Furthermore, unlike the EM algorithm, which is only guaranteed to converge to a local maximum, we prove consistency and convergence rate for our estimator in the Frobenius norm, spectral norm and ℓ_∞ norm. Our results have important practical consequences as they allow practitioners to use existing tools for penalized covariance selection (see, e.g., Friedman et al., 2008), which are very efficient in high-dimensions for data sets with missing values without changing the algorithm or resorting to the iterative EM algorithm.

Throughout the paper we assume that the missing values are missing at random in the sense of Rubin (1976). Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times p}$ be a matrix of observations with samples organized into rows, and let $\mathbf{R} = (r_{ij}) \in \mathbb{R}^{n\times p}$ be a matrix of indicators of observed values, that is, $r_{ij} = 1$ if the value x_{ij} was observed and $r_{ij} = 0$ otherwise. We assume that the data is missing completely at random (MCAR), which means that $\mathbb{P}[\mathbf{R}|\mathbf{X}, \varphi] = \mathbb{P}[\mathbf{R}|\varphi]$ for all $\mathbf{X}$ and φ, where φ denotes unknown parameters. The MCAR assumption implies that the missingness does not depend on the observed values, e.g., in a distributed environment, each sensor may fail independently from other sensors. This assumption is relaxed in the experimental section where we test the robustness of our procedure when the missing data mechanism departs from the MCAR assumption. Another more realistic assumption is called missing at random (MAR), which assumes $\mathbb{P}[\mathbf{R}|\mathbf{X}, \varphi] = P[\mathbf{R}|\mathbf{X}_{\text{obs}}, \varphi]$ for all $\mathbf{X}_{\text{mis}}$ and φ, where $\mathbf{X}_{\text{obs}}$ denotes the observed components of $\mathbf{X}$ and $\mathbf{X}_{\text{mis}}$ denotes the missing components. The MAR assumes that the distribution of $\mathbf{R}$ depends on the observed values of $\mathbf{X}$, but not on the missing values, e.g., cholesterol level may be measured only if patient has high blood pressure. Finally, the missing data mechanism is called not-missing at random (NMAR) if the distribution of $\mathbf{R}$ depends on the non-observed values of $\mathbf{X}$. Estimation under NMAR is a hard problem, as one needs to make assumptions on the model for missing values. The method presented in this paper can, in theory, be extended to handle the MAR case.

2. Problem setup and the EM algorithm

Let $\{\mathbf{x}_i\}_{i=1}^n$ be an *i.i.d.* sample from the multivariate Normal distribution with the mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$. Let $\mathbf{R} \in \mathbb{R}^{n\times p}$ be a matrix of missing values indicators with $r_{ij} = 1$ if x_{ij} is observed and 0 otherwise. The goal is to estimate the sparse precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ from the data with missing values.

Estimating covariance matrices from data with missing values is quite an old problem. See, for example, Afifi & Elashoff (1966), Wilks (1932), Anderson (1957), Hocking & Smith (1968), and Hartley & Hocking (1971). However, literature on high-dimensional estimation of covariance matrices from incomplete data is missing. Recently Städler & Bühlmann (2009) proposed to use an EM algorithm, called MissGlasso, to estimate sparse precision matrices, which we review below.

Yuan & Lin (2007) proposed to estimate the sparse precision matrix by solving the following ℓ_1-norm penalized maximization problem

$$\hat{\boldsymbol{\Omega}} = \arg\max_{\boldsymbol{\Omega}\succeq 0} \{\log|\boldsymbol{\Omega}| - \operatorname{tr}\boldsymbol{\Omega}\hat{\mathbf{S}} - \lambda||\boldsymbol{\Omega}^-||_1\}, \qquad (1)$$

where $\hat{\mathbf{S}}$ is the empirical covariance matrix, $\boldsymbol{\Omega}^- := \boldsymbol{\Omega} - \operatorname{diag}(\boldsymbol{\Omega})$ and $||\mathbf{A}||_1 = \sum_{ij}|A_{ij}|$. The tuning parameter $\lambda > 0$ controls the sparsity of the solution and hence the complexity of the solution. The optimization problem in (1) can be solved efficiently using a number of procedures (e.g., Friedman et al., 2008; Hsieh et al., 2011).

When the data are fully observed, Yuan & Lin (2007) arrived at the optimization procedure in (1) from the penalized maximum likelihood approach, with $\hat{\mathbf{S}} = n^{-1}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$. In the case when data contains missing values, the log-likelihood of observed data takes the following form

$$\ell(\boldsymbol{\mu}, \boldsymbol{\Omega}; \{\mathbf{x}_{i,\text{obs}}\}_i) = -\frac{1}{2}\sum_{i=1}^n \Big(\log|(\boldsymbol{\Omega}^{-1})_{i,\text{obs}}| + (\mathbf{x}_{i,\text{obs}} - \boldsymbol{\mu}_{i,\text{obs}})'((\boldsymbol{\Omega}^{-1})_{i,\text{obs}})^{-1}(\mathbf{x}_{i,\text{obs}} - \boldsymbol{\mu}_{i,\text{obs}})\Big),$$

where for a sample point $\mathbf{x}_i$ we write $\mathbf{x}_i$ =

$(\mathbf{x}_{i,\text{obs}}, \mathbf{x}_{i,\text{mis}})$ to denote observed and missing components, and $\boldsymbol{\mu}_{i,\text{obs}}$ and $\boldsymbol{\Omega}_{i,\text{obs}}$ are the mean and precision matrix of the observed components of $\mathbf{x}_i$. MissGlasso is an EM algorithm that finds a local maximum $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}})$ of the ℓ_1 penalized observed log-likelihood. In the E-step, MissGLasso imputes the missing values by conditional means of the distribution. That is, imputation is done by $\hat{\mathbf{x}}_{i,\text{mis}} = \hat{\boldsymbol{\mu}}_{\text{mis}} - (\hat{\boldsymbol{\Omega}}_{\text{mis,mis}})^{-1}\hat{\boldsymbol{\Omega}}_{\text{mis,obs}}(\mathbf{x}_{i,\text{obs}} - \mu_{\text{obs}})$, where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Omega}}$ are the current estimates of the parameters. In the M-step, the optimization problem (1) is solved using the GLasso on data with imputed missing values. The procedure iterates between the E-step and the M-step until convergence to a local optimum of the penalized observed log-likelihood. We will denote $\hat{\boldsymbol{\Omega}}^{\text{EM}}$, the final estimator of the precision matrix obtained by the EM algorithm. As the objective is non-convex, it is difficult to establish theoretical guarantees on the solution produced by the EM. Next, we present our estimator.

3. Plug-in estimator and related procedures

In this section, we propose a simple procedure based on the plug-in estimator of the covariance matrix from available data that can be used together with existing procedures for estimating precision matrices from fully observed data. Specifically, we will use the penalized likelihood approach, which was introduced in the previous section in (1). From (1) it is obvious that we only need a sample estimate of the covariance matrix, which is plugged into a convex program that produces an estimate of the precision matrix.

We form a sample covariance matrix from the available samples containing missing values as follows. Let $\hat{\mathbf{S}} = [\hat{\sigma}_{ab}]_{ab}$ be the sample covariance matrix with elements

$$\hat{\sigma}_{ab} = \frac{\sum_{i=1}^{n} r_{ia} r_{ib} (x_{ia} - \hat{\mu}_a)(x_{ib} - \hat{\mu}_b)}{\sum_{i=1}^{n} r_{ia} r_{ib}} \qquad (2)$$

where $\hat{\mu} = (\hat{\mu}_a)$ is the sample mean defined as $\hat{\mu}_a = (\sum_{i=1}^{n} r_{ia})^{-1} \sum_{i=1}^{n} r_{ia} x_{ia}$. Observe that the missing values in $\mathbf{X}$ are taken into account naturally and that the mean and covariance elements are estimated only from the observed sample. Under the MCAR assumption, it is simple to show that $\hat{\mathbf{S}}$ is an unbiased estimator of $\boldsymbol{\Sigma}$, that is, $\mathbb{E}[\hat{\mathbf{S}}] = \boldsymbol{\Sigma}$.

Our estimator is formed by plugging $\hat{\mathbf{S}}$ into the objective in (1), which we will denote as $\hat{\boldsymbol{\Omega}}^{\text{mGLasso}}$. Note that $\hat{\mathbf{S}}$ is not necessarily a positive definite matrix, however, the minimization problem in (1) is still convex and the resulting estimator $\hat{\boldsymbol{\Omega}}^{\text{mGLasso}}$ will be positive definite and unique. In the next section, we leverage the analysis of Ravikumar et al. (2008) to establish a number of good statistical properties of the estimator $\hat{\boldsymbol{\Omega}}^{\text{mGLasso}}$.

3.1. Selecting tuning parameters

The procedure described in the previous section requires selection of the tuning parameters λ, which controls the sparsity of the solution and balances it to the fit to data. A common approach is to form a grid of candidate values for the tuning parameter λ and choose one that minimizes a modified BIC criterion

$$\text{BIC}(\lambda) = -2\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}}; \{\mathbf{x}_{i,\text{obs}}\}_i) + \log(n) \sum_{a \le b} \mathbb{1}\{\hat{\omega}_{ab} \neq 0\}.$$

Here $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}})$ are estimates obtained using the tuning parameter λ and $\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}}; \{\mathbf{x}_{i,\text{obs}}\}_i)$ is the observed log-likelihood. Yuan & Lin (2007) proposed to use $\sum_{a \le b} \mathbb{1}\{\hat{\omega}_{ab} \neq 0\}$ to measure the degrees of freedom.

Performing cross-validation is another possibility for finding the optimal parameter λ. In the V-fold cross-validation, samples are divided into V disjoint folds, say $\mathcal{D}_v$ for $v = 1, \ldots, V$, and the score is computed as

$$\begin{aligned}\text{CV}(\lambda) = \sum_{v=1}^{V} \sum_{i \in \mathcal{D}_v} &\log |(\hat{\boldsymbol{\Omega}}_v^{-1})_{i,\text{obs}}| + (\mathbf{x}_{i,\text{obs}} - (\hat{\boldsymbol{\mu}}_v)_{i,\text{obs}})' \\ &\times ((\hat{\boldsymbol{\Omega}}_v^{-1})_{i,\text{obs}})^{-1} (\mathbf{x}_{i,\text{obs}} - (\hat{\boldsymbol{\mu}}_v)_{i,\text{obs}}),\end{aligned}$$

where $(\hat{\boldsymbol{\mu}}_v, \hat{\boldsymbol{\Omega}}_v)$ denote estimates obtained from the sample $\{\mathbf{x}_i\}_{i=1}^n \backslash \mathcal{D}_v$. The optimal tuning parameter $\hat{\lambda}$ is the one that minimizes $\text{CV}(\lambda)$. The final estimates $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}})$ are constructed using the optimization procedure with the tuning parameter $\hat{\lambda}$ on all the data.

3.2. Related procedures

Lounici (2012) and Loh & Wainwright (2011) have recently proposed procedures for estimating approximately low-rank covariance matrices and sparse precision matrices, respectively, from high-dimensional data with missing values. In both works, a sample covariance estimator is formed, which is then plugged into an optimization procedure. The sample covariance estimator they consider, assuming $(r_{ia})_{ia} \overset{iid}{\sim} \text{Bern}(\gamma)$ with $\gamma \in (0, 1]$ known, is defined as

$$\tilde{\boldsymbol{\Sigma}} = (\gamma^{-1} - \gamma^{-2})\text{diag}(\check{\boldsymbol{\Sigma}}) + \gamma^{-2}\check{\boldsymbol{\Sigma}}$$

where $\check{\boldsymbol{\Sigma}} = [\check{\sigma}_{ab}]_{ab}$ and $\check{\sigma}_{ab} = n^{-1} \sum_{i=1}^{n} r_{ia} r_{ib} x_{ia} x_{ib}$. The estimator $\tilde{\boldsymbol{\Sigma}}$ is an unbiased estimator of the covariance matrix, however, it requires knowledge of the parameter γ.

Procedure of Lounici (2012) is focused on estimating a covariance matrix under the assumption that the true covariance matrix is approximately low rank and hence is not comparable to our procedure. Loh & Wainwright (2011) used a projected gradient descent method to obtain a solution to a high-dimensional regression problem when data contains missing values. A sparse precision matrix can be obtained by maximizing an ℓ_1 penalized pseudo-likelihood, which reduces to a sequence of regression problems. We note that the estimator $\hat{\mathbf{\Omega}}^{\text{mGLasso}}$ can be obtained using any convex program solver that can solve (1), while the results of Loh & Wainwright (2011) rely on the usage of projected gradient descent.

4. Theoretical results

In this section, we provide theoretical analysis of the estimates $\hat{\mathbf{\Omega}}^{\text{mGLasso}}$, which we denote $\hat{\mathbf{\Omega}}$ throughout the section for notational simplicity, under the MCAR assumption. We start by analyzing the sample covariance matrix $\hat{\mathbf{S}}$ in (2). We will assume that each element of the missing values indicator matrix $\mathbf{R}$ is independently distributed as $r_{ia} \sim \text{Bern}(\gamma)$, $i = 1, \ldots, n,\ a = 1, \ldots, p$. Furthermore, we assume that a distribution of $\mathbf{X}$ has sub-Gaussian tails, that is, there exists a constant $\sigma \in (0, \infty)$ such that

$$\mathbb{E}[\exp(t(X_{ia} - \mu_a))] \leq \exp(\sigma^2 t^2), \text{ for all } t \in \mathbb{R}.$$

A multivariate Gaussian distribution satisfies this condition. We define the function $f(n, \gamma, \delta)$, which will be useful for characterizing probabilistic deviation of different quantities, as

$$f(n, \gamma, \delta) = (n\gamma^2 - \sqrt{2n\gamma^2 \log(2/\delta)})^{-1} \log(8/\delta).$$

Our first result characterizes the deviation of the sample covariance matrix from the true covariance matrix.

Lemma 1. *Assume that $X_a/\sqrt{\Sigma_{aa}}$ is sub-Gaussian with parameter σ^2. Fix $\delta > 0$ and assume that n is big enough so that $f(n, \gamma, \delta) \leq 1/2$. Then for any fixed $(a, b) \in \{1, \ldots, p\}^2$, $a \neq b$, with probability at least $1 - \delta$, we have that $|\hat{\sigma}_{ab} - \sigma_{ab}| \leq C_\sigma \sqrt{f(n, \gamma, \delta)}$ where $C_\sigma = 16\sqrt{2}(1 + 4\sigma^2) \max_a \sigma_{aa}$.*

Similarly, we can obtain that for any diagonal elements of $\hat{\mathbf{S}}$ the statement $|\hat{\sigma}_{aa} - \sigma_{aa}| \leq C_\sigma \sqrt{f(n, \sqrt{\gamma}, \delta)}$ holds with probability $1 - \delta$.

We use Lemma 1 to prove properties of the estimate $\hat{\mathbf{\Omega}}^{\text{mGLasso}}$. We start by introducing some additional notation and assumptions. Following Ravikumar et al. (2008), we introduce the *irrepresentable condition*:

$$\|\mathbf{\Gamma}_{S^C S}(\mathbf{\Gamma}_{SS})^{-1}\|_\infty \leq 1 - \alpha, \quad \alpha \in (0, 1] \tag{3}$$

where $\mathbf{\Gamma} = \mathbf{\Omega} \otimes \mathbf{\Omega}$, $S := \{(a, b) \ : \ \omega_{ab} \neq 0\}$ is support of $\mathbf{\Omega}$ and $S^C := \{(a, b) \ : \ \omega_{ab} = 0\}$, and $\|\cdot\|_\infty$ is the ℓ_∞/ℓ_∞-operator norm. Furthermore, we define $K_{\mathbf{\Sigma}} := \|\Sigma\|_\infty$ and $K_{\mathbf{\Gamma}} := \|(\mathbf{\Gamma}_{SS})^{-1}\|_\infty$. The maximum number of non-zero elements in a row of $\mathbf{\Omega}$ is denoted $d := \max_{a=1,\ldots,p} |\{b \ : \ \omega_{ab} \neq 0\}|$. The rate of convergence will depend on these quantities.

Theorem 2. *Suppose that the distribution of $\mathbf{X}$ satisfies the irrepresentable condition in (3) with parameter $\alpha \in (0, 1]$ and that the missing values indicator matrix $\mathbf{R}$ has i.i.d. Bern(γ) elements, that is, the data is missing completely at random with probability $1 - \gamma$. Furthermore, assume that the conditions of Lemma 1 hold. Let $\hat{\mathbf{\Omega}}$ be the unique solution for the regularization parameter $\lambda = \frac{8}{\alpha} C_\sigma \sqrt{f(n, \gamma, p^{-\tau})}$ with some $\tau > 2$ and $C_\sigma = 16\sqrt{2}(1 + 4\sigma^2) \max_a \sigma_{aa}$. If the sample size satisfies*

$$n > \frac{2(C_1^2(1 + 8\alpha^{-1})^2 d^2 + C_1(1 + 8\alpha^{-1})d) \log 8p^\tau}{\gamma^2}$$

where $C_1 = 6C_\sigma \max\{K_{\mathbf{\Sigma}} K_{\mathbf{\Gamma}}, K_{\mathbf{\Sigma}}^3 K_{\mathbf{\Gamma}}^2\}$ then with probability at least $1 - p^{2-\tau}$

$$\max_{a,b} |\hat{\omega}_{ab} - \omega_{ab}| \leq 2(1 + 8\alpha^{-1}) K_{\mathbf{\Gamma}} C_\sigma \sqrt{f(n, \gamma, p^{-\tau})},$$

where $\hat{\mathbf{\Omega}} = [\hat{\omega}_{ab}]_{ab}$ and $\mathbf{\Omega} = [\omega_{ab}]_{ab}$.

The result follows from application of Theorem 1 in Ravikumar et al. (2008) to the tail bound in Lemma 1 and some algebra. A simple consequence of Theorem 2 is that the support $\hat{S}$ of $\hat{\mathbf{\Omega}}$ consistently estimates the support S of $\mathbf{\Omega}$ if all the elements of $\mathbf{\Omega}$ are large enough in absolute values.

Corollary 3. *Under the same assumptions as in Theorem 2, we have that $\mathbb{P}[\hat{S} = S] \geq 1 - p^{2-\tau}$ if $\min_{ab} |\omega_{ab}| \geq 2(1 + 8\alpha^{-1}) K_{\mathbf{\Gamma}} C_\sigma \sqrt{f(n, \gamma, p^{-\tau})}$.*

Proof follows by straightforward algebra from Theorem 2. Using the element-wise ℓ_∞ bound on deviation of $\hat{\mathbf{\Omega}}$ from $\mathbf{\Omega}$ established in Theorem 2, we can simply establish the bounds on the convergence in the Frobenius and spectral norms.

Corollary 4. *Under the same assumptions as in Theorem 2, we have that with probability at least $1 - p^{2-\tau}$,*

$$||\hat{\mathbf{\Omega}} - \mathbf{\Omega}||_F \leq K \sqrt{|S| f(n, \gamma, p^{-\tau})}, \text{ and}$$

$$\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2 \leq K \min\{\sqrt{|S|}, d\} \sqrt{f(n, \gamma, p^{-\tau})}$$

where $K = 2(1 + 8\alpha^{-1}) K_{\mathbf{\Gamma}} C_\sigma$.

Proof follows by straightforward algebra from Theorem 2. We can compare the established results for $\hat{\mathbf{\Omega}}$ under the MCAR assumption to results of Ravikumar

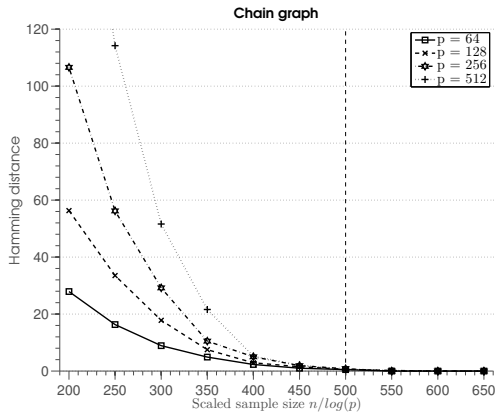

Figure 1. Hamming distance between the support of $\hat{\mathbf{\Omega}}$ and $\mathbf{\Omega}$ averaged over 100 runs. Vertical line marks a threshold at which the graph structure is consistently estimated.

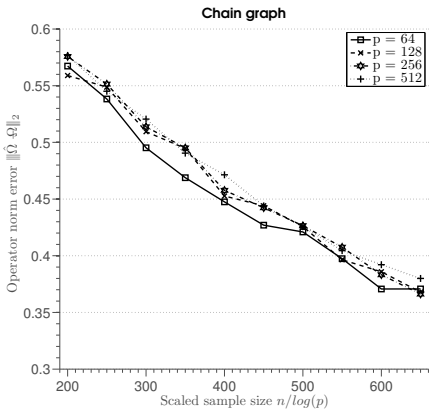

Figure 2. Operator norm error averaged over 100 runs. We observe that the error curve align when plotted against the rescaled sample size.

et al. (2008) for the fully observed case. We observe that the sample size increases by a factor of $\mathcal{O}(\gamma^{-2})$, while the rate of convergence in the element-wise ℓ_∞ norm is slower by a factor of $\mathcal{O}(\gamma^{-1})$. The parameter γ which controls the rate of missing data is commonly considered a constant, however, it is clear from Theorem 2 that we could let $\gamma \to 0$ slowly as a function of n and p, while maintaining the convergence properties of the procedure.

5. Simulation Analysis

In this section, we perform a set of simulation studies to illustrate finite sample performance of our procedure. First, we show that the scalings predicted by the theory are sharp. Next, we compare our procedure to the EM algorithm, MissGLasso (Städler & Bühlmann, 2009) and the projected gradient method (Loh & Wainwright, 2011), PGLasso. Furthermore, we can explore robustness of our method when the data generating process departs from the one assumed in Section 4.

5.1. Verifying theoretical scalings

Theoretical results given in Section 4 predict behavior of the error when estimating the precision matrix. In particular, Corollary 3 suggests that we need $\mathcal{O}(d^2 \log(p))$ samples to estimate the graph structure consistently and Corollary 4 states that the error in the operator norm decreases as $\mathcal{O}(d\sqrt{\log(p)/n})$. Therefore, if we plot the error curves against appropriately rescaled sample size, we expect them to align for different problem sizes. To verify this, we create a chain-structured Gaussian graphical model (following Loh & Wainwright (2011)), so that $d = 2$ and the precision matrix $\mathbf{\Omega}$ is created as follows. Each diagonal element is set to 1, and all the entries corresponding to the chain are set equal to 0.1. The precision matrix is rescaled so that $|||\mathbf{\Omega}|||_2 = 1$ and $\gamma = 0.8$.

Figure 1 shows the hamming distance between the support of $\hat{\mathbf{\Omega}}$ and $\mathbf{\Omega}$ plotted against the rescaled sample size. Vertical line marks a threshold in scaled sample size after which the pattern of non-zero element of the precision matrix is consistently recovered. Figure 2 shows that the error curves align when the sample size is rescaled, as predicted by the theory.

5.2. Data missing completely at random

Our first simulation explores the MCAR assumption. We use models from Städler & Bühlmann (2009):

Model 1: $\sigma_{ab} = 0.7^{|a-b|}$, so that the elements of the covariance matrix decay exponentially.

Model 2:

$$\sigma_{ab} = \mathbb{1}_{\{a=b\}} + 0.4\,\mathbb{1}_{\{|a-b|=1\}} + 0.2\,\mathbb{1}_{\{|a-b|=2\}} + 0.2\,\mathbb{1}_{\{|a-b|=3\}} + 0.1\,\mathbb{1}_{\{|a-b|=4\}},$$

where the symbol $\mathbb{1}$ represents the indicator function which is 1 if $a = b$ and 0 otherwise.

Model 3: $\mathbf{\Omega} = \mathbf{B} + \delta\mathbf{I}$, where each off-diagonal entry of $\mathbf{B}$ is generated independently and equals 0.5 with probability $\alpha = 0.1$ or 0 with probability $1 - \alpha$. Diagonal entries of $\mathbf{B}$ are zero, and δ is chosen so that the condition number of $\mathbf{\Omega}$ is p.

We report convergence results in the operator norm. We also report precision and recall for the performance on recovering the sparsity structure of $\mathbf{\Omega}$, where precision $= \frac{|\hat{S} \cap S|}{|\hat{S}|}$ and recall $= \frac{|\hat{S} \cap S|}{|S|}$. As described in Section 3.1, the tuning parameter λ is selected by minimizing the BIC criterion. We observed that using the

			Recall			Precision		
			MissGLasso	mGLasso	PGLasso	MissGLasso	mGLasso	PGLasso
Model 1	p=100	0%	NA	1.000(0.000)	1.000(0.000)	NA	0.973(0.045)	0.991(0.015)
		10%	1.000(0.000)	1.000(0.000)	0.998(0.008)	0.608(0.068)	0.915(0.059)	0.998(0.010)
		20%	0.999(0.004)	1.000(0.003)	0.967(0.006)	0.636(0.081)	0.897(0.073)	0.999(0.003)
		30%	0.977(0.062)	0.989(0.003)	0.759(0.140)	0.642(0.064)	0.836(0.057)	0.998(0.009)
	p=200	0%	NA	1.000(0.000)	0.891(0.005)	NA	0.950(0.046)	0.999(0.004)
		10%	0.860(0.022)	0.950(0.006)	0.782(0.024)	0.858(0.043)	0.803(0.046)	0.984(0.027)
		20%	0.833(0.053)	0.930(0.001)	0.556(0.006)	0.763(0.048)	0.734(0.062)	0.952(0.091)
		30%	0.794(0.138)	0.923(0.003)	0.553(0.009)	0.729(0.059)	0.731(0.060)	0.941(0.052)
	p=500	0%	NA	1.000(0.001)	0.889(0.015)	NA	0.912(0.022)	0.995(0.003)
		10%	0.931(0.011)	0.933(0.031)	0.855(0.023)	0.834(0.029)	0.862(0.044)	0.966(0.010)
		20%	0.852(0.064)	0.920(0.024)	0.767(0.026)	0.811(0.037)	0.841(0.037)	0.965(0.025)
		30%	0.808(0.045)	0.887(0.028)	0.526(0.031)	0.739(0.043)	0.781(0.030)	0.963(0.033)
Model 2	p=100	0%	NA	0.330(0.008)	0.403(0.006)	NA	0.420(0.012)	0.297(0.012)
		10%	0.278(0.019)	0.280(0.011)	0.380(0.007)	0.342(0.012)	0.375(0.010)	0.319(0.008)
		20%	0.240(0.022)	0.253(0.018)	0.259(0.012)	0.339(0.028)	0.372(0.027)	0.320(0.026)
		30%	0.231(0.031)	0.241(0.027)	0.174(0.030)	0.267(0.033)	0.281(0.037)	0.331(0.042)
	p=200	0%	NA	0.281(0.011)	0.410(0.013)	NA	0.570(0.012)	0.270(0.021)
		10%	0.331(0.011)	0.261(0.010)	0.361(0.011)	0.354(0.013)	0.471(0.015)	0.257(0.018)
		20%	0.261(0.012)	0.243(0.015)	0.283(0.013)	0.274(0.018)	0.354(0.021)	0.313(0.021)
		30%	0.218(0.017)	0.232(0.017)	0.208(0.017)	0.281(0.019)	0.267(0.031)	0.453(0.059)
	p=500	0%	NA	0.309(0.006)	0.302(0.012)	NA	0.510(0.007)	0.540(0.018)
		10%	0.305(0.007)	0.307(0.005)	0.357(0.009)	0.461(0.008)	0.462(0.010)	0.224(0.012)
		20%	0.297(0.010)	0.315(0.027)	0.243(0.015)	0.272(0.026)	0.223(0.048)	0.383(0.019)
		30%	0.238(0.025)	0.242(0.023)	0.203(0.028)	0.267(0.031)	0.259(0.033)	0.396(0.021)
Model 3	p=100	0%	NA	0.943(0.002)	0.971(0.015)	NA	0.532(0.017)	0.251(0.051)
		10%	0.857(0.010)	0.857(0.003)	0.994(0.005)	0.857(0.009)	0.882(0.004)	0.200(0.006)
		20%	0.829(0.017)	0.857(0.012)	0.886(0.035)	0.691(0.022)	0.588(0.015)	0.307(0.059)
		30%	0.771(0.053)	0.829(0.033)	0.595(0.096)	0.780(0.050)	0.671(0.050)	0.797(0.053)
	p=200	0%	NA	0.783(0.005)	1.000(0.003)	NA	0.921(0.002)	0.245(0.023)
		10%	0.747(0.005)	0.733(0.006)	0.998(0.007)	0.887(0.009)	0.921(0.004)	0.233(0.030)
		20%	0.667(0.009)	0.747(0.030)	0.931(0.014)	0.909(0.015)	0.737(0.031)	0.311(0.023)
		30%	0.480(0.037)	0.600(0.052)	0.801(0.045)	0.837(0.059)	0.804(0.033)	0.412(0.035)
	p=500	0%	NA	0.744(0.005)	0.998(0.002)	NA	0.844(0.003)	0.191(0.019)
		10%	0.627(0.006)	0.718(0.006)	0.994(0.003)	0.893(0.003)	0.835(0.005)	0.180(0.020)
		20%	0.601(0.010)	0.699(0.031)	0.923(0.029)	0.887(0.034)	0.789(0.037)	0.259(0.054)
		30%	0.511(0.039)	0.614(0.038)	0.851(0.041)	0.800(0.043)	0.755(0.027)	0.355(0.047)

Table 1. Average (standard deviation) recall and precision under the MCAR assumption.

tuning parameters that minimize the cross-validation loss result in complex estimates with many falsely selected edges (results not reported).

We set the sample size and number of dimensions $(n, p) = (100, 100), (150, 200), (200, 500)$ for each model and report results averaged over 50 independent runs for each setting. For each generated data set, we remove completely at random 10%, 20% and 30% entries. Results on recall and precision for different degrees of missingness are reported in Table 1, while the operator norm convergence results are reported in Table 2. From the simulations, we observe that mGLasso performs better than the EM algorithm on the task of recovering the sparsity pattern of the precision matrix. PGLasso does well on Model 1, but does not perform so well under Model 2 and 3. Model 2 is a difficult one for recovering non-zero patterns, as the true precision matrix contains many small non-zero elements. The EM algorithm performs better than mGLasso and PGLasso measured by $\|\hat{\mathbf{\Omega}} - \mathbf{\Omega}\|_2$, with mGLasso doing better than PGLasso. However, on average the EM algorithm requires 20 iterations for convergence, which makes mGLasso about 20 times faster on average.

5.3. Data missing at random

In the previous section, we have simulated data with missing values completely at random, under which consistency of the estimator $\hat{\mathbf{\Omega}}^{\text{mGLasso}}$ given in Section 3 can be proven. When the missing values are produced at random (MAR), the EM algorithm described is still valid, however, the estimator $\hat{\mathbf{\Omega}}^{\text{mGLasso}}$ is not. Little (1988) provided a statistical test for checking whether missing values are missing completely at random, however, no such tests exist for high-dimensional data. In this section, we will observe how robust our estimator is when the data generating mechanism departs from the MCAR assumption. When the missing data mechanism is NMAR, then neither the EM algorithm, nor the procedures described Section 3 are valid.

We will use the model considered in Städler & Bühlmann (2009) in Section 4.1.2. The model is a Gaussian with $p = 30$, $n = 100$ and the covariance

		MissGLasso	mGLasso	PGLasso
			Model 1	
p=100	0%	NA	2.10(0.01)	4.35(0.01)
	10%	2.25(0.01)	2.31(0.01)	4.69(0.01)
	20%	2.35(0.04)	2.42(0.03)	4.78(0.04)
	30%	2.69(0.05)	2.85(0.04)	4.82(0.06)
p=200	0%	NA	2.26(0.01)	4.49(0.01)
	10%	2.32(0.01)	2.73(0.01)	4.76(0.02)
	20%	2.51(0.01)	2.88(0.01)	4.86(0.02)
	30%	2.96(0.02)	3.04(0.01)	4.98(0.05)
p=500	0%	NA	3.59(0.03)	4.94(0.03)
	10%	3.71(0.02)	3.85(0.02)	5.25(0.04)
	20%	3.99(0.03)	3.99(0.02)	5.32(0.04)
	30%	4.11(0.05)	4.77(0.04)	5.76(0.05)
			Model 2	
p=100	0%	NA	1.25(0.01)	1.63(0.01)
	10%	1.32(0.01)	1.66(0.01)	1.75(0.01)
	20%	1.59(0.01)	1.75(0.01)	1.88(0.02)
	30%	1.66(0.02)	1.86(0.01)	1.99(0.02)
p=200	0%	NA	1.31(0.01)	1.69(0.01)
	10%	1.41(0.01)	1.71(0.01)	1.71(0.01)
	20%	1.61(0.01)	1.79(0.02)	1.99(0.01)
	30%	1.69(0.01)	1.87(0.01)	2.08(0.01)
p=500	0%	NA	1.44(0.01)	1.73(0.01)
	10%	1.49(0.01)	1.74(0.01)	1.84(0.02)
	20%	1.66(0.01)	1.81(0.02)	2.05(0.03)
	30%	1.72(0.02)	1.95(0.02)	2.22(0.04)
			Model 3	
p=100	0%	NA	1.12(0.01)	1.35(0.01)
	10%	1.16(0.01)	1.32(0.01)	1.42(0.02)
	20%	1.20(0.01)	1.64(0.02)	1.70(0.03)
	30%	1.49(0.05)	1.67(0.03)	1.83(0.03)
p=200	0%	NA	1.35(0.01)	1.59(0.01)
	10%	1.43(0.01)	1.62(0.01)	1.83(0.01)
	20%	1.46(0.03)	1.71(0.02)	1.87(0.01)
	30%	1.52(0.03)	1.82(0.01)	1.93(0.03)
p=500	0%	NA	1.42(0.01)	1.64(0.02)
	10%	1.47(0.01)	1.69(0.02)	1.86(0.01)
	20%	1.55(0.02)	1.73(0.04)	1.92(0.03)
	30%	1.59(0.02)	1.87(0.03)	2.01(0.03)

Table 2. Average (standard deviation) distance in the operator norm $\|\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}\|\|_2$ under the MCAR assumption.

matrix is block-diagonal, $\boldsymbol{\Sigma} = \text{diag}(\mathbf{B}, \mathbf{B}, \ldots, \mathbf{B})$ with $\mathbf{B} \in \mathbb{R}^{3\times 3}$, $b_{ab} = 0.7^{|a-b|}$. Missing values are created using the following three mechanisms:

1. For all $j = 1, \ldots, \lfloor p/3 \rfloor$ and $i = 1, \ldots, n$: $x_{i,3*j}$ is missing if $r_{i,j} = 0$ where $r_{i,j} \overset{iid}{\sim} \text{Bern}(\pi)$.
2. For all $j = 1, \ldots, \lfloor p/3 \rfloor$ and $i = 1, \ldots, n$: $x_{i,3*j}$ is missing if $x_{i,3*j-2} < T$.
3. For all $j = 1, \ldots, \lfloor p/3 \rfloor$ and $i = 1, \ldots, n$: $x_{i,3*j}$ is missing if $x_{i,3*j} < T$.

The threshold value T determines the percentage of missing values. We consider three settings: 1) $\pi = 0.25$ and $T = \Phi^{-1}(0.25)$, 2) $\pi = 0.5$ and $T = \Phi^{-1}(0.5)$. and 3) $\pi = 0.75$ and $T = \Phi^{-1}(0.75)$ where $\Phi(\cdot)$ is the standard Normal cumulative distribution function. The first missing data mechanism is MCAR as the missing values do not depend on the observed values. The second missing data mechanism is MAR as the missing value indicators depend on the observed values of other variables. Finally, the third missing data mechanism is NMAR as the missing data indicators depend on the unobserved values.

		MissGLasso	mGLasso	PGLasso
$\pi = 0.25$	MCAR	2.88(0.02)	3.16(0.01)	3.72(0.01)
	MAR	3.24(0.01)	3.92(0.03)	4.15(0.05)
	NMAR	5.78(0.05)	6.57(0.08)	7.64(0.10)
$\pi = 0.5$	MCAR	2.97(0.03)	3.28(0.02)	3.77(0.02)
	MAR	3.41(0.05)	4.16(0.06)	4.58(0.04)
	NMAR	6.15(0.07)	6.61(0.10)	8.12(0.12)
$\pi = 0.75$	MCAR	3.17(0.02)	3.31(0.03)	3.99(0.03)
	MAR	3.59(0.05)	4.47(0.04)	4.87(0.05)
	NMAR	6.87(0.11)	7.04(0.13)	8.76(0.15)

Table 3. Average (standard deviation) distance in the operator norm $\|\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}\|\|_2$ when missing values mechanism is MCAR, MAR and NMAR. The fraction of the observed data is controlled by π.

Results of the simulation, averaged over 50 independent runs, are summarized in Table 3 and Table 4. We first observe that when the missing values are not missing at random, performance of all procedures degrades. Furthermore, the EM algorithm performs better than the other two methods when the data is generated under MAR. This is expected, since our proposed procedure is not valid under this assumption. Note, however, that mGLasso performs better than PGLasso under this simulation scenario.

6. Discussion and extensions

We have proposed a simple estimator for the precision matrix in high-dimensions from data with missing values. The estimator is based on a convex program that can be solved efficiently. In particular, from our simulation studies, we observed that the algorithm runs on average 20 times faster than the EM algorithm. Furthermore, the estimator does not require imputation of the missing values and can be found using existing numerical procedures. As such, we believe that it represents a viable alternative to the iterative EM algorithm.

From the analysis in Section 4, it is clear that other procedures for estimating precision matrices from fully observed data, such as the Clime estimator (Cai et al., 2011), could be easily extended to handle data with missing values. Theoretical properties of those procedures would be established using the tail bounds on the sample covariance matrix given in Lemma 1.

There are two directions in which this work should be extended. First, the MCAR assumption is very strong and it is hard to check whether it holds in practice. However, we have observed in our simulation studies

		Recall			Precision		
		MissGLasso	mGLasso	PGLasso	MissGLasso	mGLasso	PGLasso
$\pi = 0.25$	MCAR	0.900(0.003)	0.950(0.005)	1.000(0.000)	0.900(0.002)	0.861(0.006)	0.333(0.030)
	MAR	0.512(0.026)	0.815(0.070)	0.501(0.067)	0.995(0.006)	0.471(0.052)	0.634(0.025)
	NMAR	0.500(0.015)	0.443(0.052)	0.465(0.112)	0.698(0.086)	0.188(0.021)	0.213(0.091)
$\pi = 0.5$	MCAR	0.800(0.005)	0.900(0.003)	1.000(0.000)	0.889(0.008)	0.774(0.068)	0.263(0.050)
	MAR	0.650(0.034)	0.900(0.005)	0.551(0.061)	0.921(0.021)	0.393(0.089)	0.453(0.072)
	NMAR	0.531(0.042)	0.613(0.477)	0.463(0.073)	0.684(0.092)	0.370(0.285)	0.315(0.109)
$\pi = 0.75$	MCAR	0.626(0.062)	0.635(0.220)	0.775(0.081)	0.924(0.053)	0.891(0.063)	0.221(0.039)
	MAR	0.619(0.014)	0.611(0.132)	0.431(0.075)	0.879(0.061)	0.555(0.074)	0.399(0.044)
	NMAR	0.491(0.046)	0.557(0.115)	0.411(0.076)	0.688(0.059)	0.464(0.067)	0.368(0.071)

Table 4. Average (standard deviation) recall and precision when missing values mechanism is MCAR, MAR and NMAR.

that under the MAR assumption, which is a more realistic assumption than MCAR, performance of the estimators does not degrade dramatically when estimating the support of the precision matrix. However, estimated parameters are quite far from the true parameters. This could be improved by using a weighted estimator for the sample covariance matrix (see, e.g., Robins et al., 1994). Second, it is important to establish sharp lower bounds for the estimation problem from data with missing values, which should reflect dependence on the proportion of observed entries γ (see Lounici, 2012).

Acknowledgments: This work is supported in part through the grants NIH R01GM087694 and AFOSR FA9550010247.

References

Afifi, A. A. and Elashoff, R. M. Missing observations in multivariate statistics: I. review of the literature. *Journal of the American Statistical Association*, 61(315):pp. 595–604, 1966.

Anderson, T. W. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):pp. 200–203, 1957.

Cai, T., Liu, W., and Luo, X. A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association, to appear*, 2011.

Cai, T.T., Zhang, C.H., and Zhou, H.H. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38(4):2118–2144, 2010.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008.

Hartley, H. O. and Hocking, R. R. The analysis of incomplete data. *Biometrics*, 27(4):pp. 783–823, 1971.

Hocking, R. R. and Smith, Wm. B. Estimation of parameters in the multivariate normal distribution with missing observations. *Journal of the American Statistical Association*, 63(321):pp. 159–173, 1968.

Hsieh, C.-J., Sustik, M., Dhillon, I., and Ravikumar, P. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems 24*, pp. 2330–2338. 2011.

Little, R. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404):1198–1202, 1988.

Little, R.J.A. and Rubin, D.B. *Statistical analysis with missing data*. Wiley, New York, 1987.

Loh, P.-L. and Wainwright, M. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems 24*, pp. 2726–2734. 2011.

Lounici, K. High-dimensional covariance matrix estimation with missing observations. *ArXiv*, January 2012.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

Peng, J., Wang, P., Zhou, N., and Zhu, J. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486): 735–746, 2009.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Nov 2008.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):pp. 846–866, 1994.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. Sparse permutation invariant covariance estimation. *Electronic Journal Of Statistics*, 2:494, 2008.

Rubin, D.B. Inference and missing data. *Biometrika*, 63 (3):581, 1976.

Städler, N. and Bühlmann, P. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, pp. 1–17, 2009.

Wilks, S. S. Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics*, 3(3):pp. 163–195, 1932.

Yuan, M. and Lin, Y. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.