Inverse Covariance Estimation from Data with Missing Values using the Concave-Convex Procedure

Jérôme Thai Timothy Hunter Anayo K. Akametalu Claire J. Tomlin Alexandre M. Bayen

***Abstract*— We study the problem of estimating sparse precision matrices from data with missing values. We show that the corresponding maximum likelihood problem is a Difference of Convex (DC) program by proving some new concavity results on the Schur complements. We propose a new algorithm to solve this problem based on the ConCave-Convex Procedure (CCCP), and we show that the standard EM procedure is a weaker CCCP for this problem. Numerical experiments show that our new algorithm, called m-CCCP, converges much faster than EM in number of iterations on both synthetic and biology datasets.**

I. Introduction

Many applications in statistics rely on estimating accurately covariance matrices and their inverses, precision matrices. This includes analysis of multivariate data such as principal component analysis and discriminant analysis [25], and such models have a wide range of practical applications, from array processing to functional genomics [10]. The most common probability model for studying correlations in continuous data is the multivariate Gaussian distribution with mean μ and covariance matrix Σ. In the context of Gaussian distributions defined over undirected graphs, also known as Gaussian Markov Random Fields (GMRFs), it is well known that the non-zero entries S_{ij} of the precision matrix $S = \Sigma^{-1}$ of the GMRF correspond precisely to the conditional dependencies between the variables [21]. Recent work in the field of learning the structure of graphical models and especially on sparse GMRFs [12], has demonstrated that promoting sparsity has compelling advantages, such as producing more robust models that generalize well to unseen data [8], cost-effective belief propagation algorithms [18], and uncovering the interactions between variables (interactions between genes [9] for example). Promoting sparsity in GMRFs is typically done with an additional ℓ_1 penalty term on the objective function that increases the sparsity of the solution S. Researchers have proposed algorithms for the exact optimization of the ℓ_1-penalized log-likelihood [27], [11], [2], [20] specifically in high-dimensional settings where the number of variables p is much larger than the sample size n. Most of these algorithms assume *full-dimensional* observations. With $\hat{\mu}$ the empirical mean and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$ the empirical covariance of the dataset $\mathscr{D} = \{x_i\}_i$, these algorithms solve:

$$\min_{S \succ 0} -\log|S| + \mathrm{Tr}(\hat{\Sigma}S) + \lambda\|S\|_1 \tag{1}$$

The authors are with the department of Electrical Engineering and Computer Science, University of California at Berkeley {jerome.thai, timothy.hunter, kakametalu, bayen}@berkeley.edu, tomlin@eecs.berkeley.edu

where $|S|$ is the determinant of matrix S, Tr the trace operator, and $\|S\|_1 = \sum_{ij}|S_{ij}|$. These results can also be extended to allow different nonnegative weights assigned to different entries, i.e. $\|\lambda \circ S\|_1 = \sum_{ij}\lambda_{ij}|S_{ij}|$.

In practice, datasets often suffer from missing values [17] due to mistakes in data collection, dropouts, or limitations from experimental design. Instead of using the full likelihood of the samples, we need to consider the marginal likelihood of the observed values, or *observed log-likelihood*. Inference for μ and S can be based on the observed log-likelihood if we assume that the underlying missing data mechanism is *ignorable*, i.e. the probability that an observation is missing may depend on the observed values but not on the missing values (Missing at Random) and the parameters of the data model and the parameters of the missing values mechanism are distinct [17]. Unfortunately, with an arbitrary pattern of missing values, no explicit maximization of the likelihood is possible even for the mean values and covariance matrices [17]. Concretely, we denote x_i a *full-dimensional* sample from $X \sim \mathscr{N}(\mu, \Sigma)$, and $x_{i,\mathrm{obs}}$ (resp. $x_{i,\mathrm{mis}}$) the observed values (resp. missing data) in x_i. Then, $x_{i,\mathrm{obs}}$ is drawn from the marginal Gaussian distribution with marginal mean $\mu_{i,\mathrm{obs}}$ and marginal covariance $\Sigma_{i,\mathrm{obs}}$, obtained by dropping the irrelevant variables in μ and Σ.

Following the work of Chandrasekaran [6], we first note that the marginal precision matrix $\mathscr{S}_i(S) := (\Sigma_{i,\mathrm{obs}})^{-1}$ is the Schur complement with respect to the block of missing or latent variables. However, we suppose in our paper that the number of latent variables and their indices are known, and that they can change for each observation. This additional information enables us to formulate a well-posed optimization problem. The observed log-likelihood is:

$$-\log|\mathscr{S}_i(S)| + (x_{i,\mathrm{obs}} - \mu_{i,\mathrm{obs}})^T\mathscr{S}_i(S)(x_{i,\mathrm{obs}} - \mu_{i,\mathrm{obs}}) \tag{2}$$

Let $\Sigma_i(\mu), i = 1, \cdots, n$ be the functions such that:

$$\Sigma_i(\mu) := (x_{i,\mathrm{obs}} - \mu_{i,\mathrm{obs}})(x_{i,\mathrm{obs}} - \mu_{i,\mathrm{obs}})^T \tag{3}$$

then the inference of μ and $S = \Sigma^{-1}$ is given by:

$$\min_{\mu, S \succ 0} \frac{1}{n}\sum_i \{-\log|\mathscr{S}_i(S)| + \mathrm{Tr}(\Sigma_i(\mu)\mathscr{S}_i(S)) + \lambda\|S\|_1 \tag{4}$$

Due to the structure of $\mathscr{S}_i(S)$, the observed log-likelihood is a non-convex function of S for a general missing data pattern, with possible existence of multiple stationary points [19], [22]. Thus, optimization of (4) is a non-trivial problem.

For a given vector of values $x_{i,\mathrm{obs}}$ observed from a full vector x_i, denote by $x_{i,\mathrm{mis}}$ the values of the rest of the variables. Hypothetically, if we have access to the *missing*

values $x_{i,\text{mis}}$, we can complete the observed values to obtain a full observation $x_i = (x_{i,\text{obs}}, x_{i,\text{mis}})$ (modulo some permutation). Naturally, this is not possible because the missing values are unobserved. However, they can be imputed using the Expectation Maximization (EM) algorithm to obtain a completed dataset on which standard methods can be applied. Under mild regularity conditions including differentiability and continuity, EM converges to a stationary point of the observed log-likelihood [26], [24]. In the E-step, the imputation is done by conditional means of the sufficient statistics $\sum_i x_i$ and $\sum_i x_i x_i^T$ given the current estimates of the parameters $\mu^{(t)}$ and $S^{(t)}$. In the M-step, the optimization of the resulting *complete log-likelihood* is solved using standard tools for estimating sparse precision matrices. In [15], an imputation based on plug-in estimator of the covariance matrix has also been proposed.

Our central observation is that the problem of minimizing (4), although nonconvex, has an objective function which can be decomposed into the sum of a convex and a concave function. This leads us to apply the concave-convex procedure (CCCP), which is a majorization-minimization (MM) algorithm that solves difference of convex (DC) programs as a sequence of convex programs [28], [23]. This approach has been applied in the past for estimating sparse covariance matrices [4]. However, that work considers the optimization problem $\min_{\Sigma \succ 0}\{\log|\Sigma| + \text{Tr}(\hat{\Sigma}\Sigma^{-1}) + \lambda\|\Sigma\|_1\}$, which differs from the objective (1) because the sparsity is attained in the covariance matrix itself rather than in the precision matrix, and because the data set is complete.

There is a significant body of prior work in machine learning and statistics that takes advantage of this structure to develop specialized algorithms: DC programs focusing on general techniques to find exact and approximate solutions of such problems [13], [1]; majorization-minimization algorithms for problems in statistics such as least-squares multidimensional scaling [7]; regularized regression with nonconvex penalties [29]. The CCCP algorithm has also been used in various machine learning applications [28], [23].

The article presents several important contributions. We propose a novel approach using CCCP with Gauss-Seidel update rule (or block coordinate CCCP), called m-CCCP to solve the problem of minimizing (4) which differs from previous works [24], [15]. The emphasis is placed on the DC decomposition of the log-likelihood rather than the statistical analysis used in the EM method. Moreover, we show that EM is also a block coordinate CCCP, using a different DC program, which provides a powerful analytical framework for comparing the two algorithms. This enables us to show that our algorithm compares favorably to EM in theoretical speed of convergence. Our results are also supported by numerical experiments. We hope our analysis will be the starting point for the design of new algorithms that outperform the well-studied EM-based methods by developing optimal DC programs.

II. Difference of Convex problem

We solve problem (4) with a block coordinate descent where we alternate between blocks μ and S. CCCP is applied to the optimization with respect to S with fixed $\Sigma_i := \Sigma_i(\mu)$:

$$\min_{\mu, S\succ 0} \frac{1}{n}\sum_i \{-\log|\mathscr{S}_i(S)| + \text{Tr}(\Sigma_i \mathscr{S}_i(S)) + \lambda\|S\|_1 \quad (5)$$

When we re-order the lines and columns of Σ and S using a permutation P_i such that the first block corresponds to the observed values $x_{i,\text{obs}}$, then the inverse of the observed block in Σ is a Schur complement. In other words, with $x_i := P_i^T \begin{bmatrix} y_i \\ z_i \end{bmatrix}$, where $y_i := x_{i,\text{obs}}$ and $z_i = x_{i,\text{mis}}$, it follows that $P_i \Sigma P_i^T = \begin{bmatrix} \Sigma_{y_i y_i} & \Sigma_{y_i z_i} \\ \Sigma_{z_i y_i} & \Sigma_{z_i z_i} \end{bmatrix}$ has inverse

$$P_i S P_i^T = \begin{bmatrix} S_{y_i y_i} & S_{y_i z_i} \\ S_{z_i y_i} & S_{z_i z_i} \end{bmatrix} = \begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix} \quad (6)$$

where the inverse $\mathscr{S}_i(S) = (\Sigma_{y_i y_i})^{-1}$ is the Schur complement of the block A_i of the matrix $P_i S P_i^T$ (with $C_i = B_i^T$ because $S = S^T$): $\mathscr{S}_i(S) = A_i - B_i D_i^{-1} C_i$. We follow this notation for the rest of the article. Because the permutation is a linear operator, we can directly consider the block matrix $S = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}$ and prove our results on the Schur complement

$$\mathscr{S}(S) = A - BD^{-1}B^T = S_{yy} - S_{yz}S_{zz}^{-1}S_{zy} \quad (7)$$

In the p-dimensional Hilbert space $\mathbb{R}^p$ with inner product $x^T y$, we denote the set of symmetric matrices S^p, the set of positive semidefinite matrices S_+^p, and the set of positive definite matrices S_{++}^p, respectively. It has been seen in [3] (Th. 1.3.3, Corollary 1.5.3) that the Schur complement (7) is concave. We now prove that the determinant of the Schur complement is also log-concave, which is a new result to the best of our knowledge. We first restate Lemmas 1 and 2 from [3], which lead to the main result of this section. The proofs are provided and will be useful for Proposition 1.

Lemma 1: Let D be positive definite. Then the block matrix $M = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}$ is positive definite if and only if $A \succ BD^{-1}B^T$.

Proof: We have $\begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A & B \\ B^T & D \end{bmatrix} \begin{bmatrix} I & 0 \\ -D^{-1}B^T & I \end{bmatrix} = \begin{bmatrix} A - BD^{-1}B^T & 0 \\ 0 & D \end{bmatrix}$ ■

We note that M defined in Lemma 1 has determinant $|A - BD^{-1}B^T| \cdot |D|$, hence $\log|A - BD^{-1}B^T| = \log|M| - \log|D|$.

Lemma 2: The map $(A,B,D) \mapsto A - BD^{-1}B^T$ is jointly concave on $S_+^p \times \mathbb{R}^{p\times q} \times S_{++}^q$.

Proof: We need to prove that $(B,D) \mapsto BD^{-1}B^T$ is jointly convex. Applying Lemma 1 with $A = BD^{-1}B^T$, we get $\begin{bmatrix} B_i D_i^{-1} B_i^T & B_i \\ B_i^T & D_i \end{bmatrix} \succeq 0,\ i = 1,2$, hence summing the two matrices gives $\begin{bmatrix} \frac{B_1 D_1^{-1} B_1^T + B_2 D_2^{-1} B_2^T}{2} & \frac{B_1+B_2}{2} \\ \left(\frac{B_1+B_2}{2}\right)^T & \frac{D_1+D_2}{2} \end{bmatrix} \succeq 0$. Using Lemma 1 again $\left(\frac{B_1+B_2}{2}\right)\left(\frac{D_1+D_2}{2}\right)^{-1}\left(\frac{B_1+B_2}{2}\right)^T \preceq \frac{B_1 D_1^{-1} B_1^T + B_2 D_2^{-1} B_2^T}{2}$ ■

If ensures that $\mathscr{S}$ is a concave function on S^p_{++} and that $\mathscr{S}(S)$ is positive definite. It is also known that the determinant is log-concave on S^p_{++} as seen in [5, Section 3.1.5]. This leads to one of our main results:

Theorem 1: The function $S \mapsto \log|\mathscr{S}(S)|$ is concave on S^p_{++}.

Proof: We have $\mathscr{S}\left(\frac{S_1+S_2}{2}\right) \succeq \frac{\mathscr{S}(S_1)+\mathscr{S}(S_2)}{2}$ from Lemma 2. Then $\left|\mathscr{S}\left(\frac{S_1+S_2}{2}\right)\right| \geq \left|\frac{\mathscr{S}(S_1)+\mathscr{S}(S_2)}{2}\right| \geq |\mathscr{S}(S_1)|^{\frac{1}{2}}|\mathscr{S}(S_2)|^{\frac{1}{2}}$ where the first inequality comes from $A \succeq B \Rightarrow |A| \geq |B|$ for positive definite matrices, and the second inequality is from the log-concavity of the determinant. Taking the log terminates the proof. ∎

Given the concavity of $\log|\mathscr{S}(S)|$ from Theorem 1 and the concavity of $x^T\mathscr{S}(S)x = \mathrm{Tr}(xx^T\mathscr{S}(S))$ from Lemma 2, the problem of maximizing the log-likelihood with respect to block S has the following DC structure:

$$\begin{array}{l} \min\limits_{S\succ 0} f_0(S) - g_0(S) + \lambda\|S\|_1 \\ \quad f_0(S) = -\frac{1}{n}\sum_{i=1}^n \log|\mathscr{S}_i(S)| \\ \quad g_0(S) = -\frac{1}{n}\sum_{i=1}^n \mathrm{Tr}(\Sigma_i\mathscr{S}_i(S)) \end{array} \tag{8}$$

where f_0 and g_0 are both convex functions. We now have a DC program. The following result gives a guarantee of convergence to a stationary point if the objective is non-increasing at each iteration of our algorithm.

Lemma 3: There exists $\alpha > 0$ such that $\forall A, B \in \mathbb{R}^{p\times p}$ similar, $\|A\|_1 \geq \alpha\|B\|_1$ with $\|M\|_1 = \sum_{ij}|M_{ij}|$.

Proof: All norms on $\mathbb{R}^{p\times p}$ are equivalent, so $\exists r, s > 0$ such that $r\|\cdot\|_1 \leq \|\cdot\|_2 \leq s\|\cdot\|_1$, with $\|\cdot\|_2$ the spectral norm. And $\|A\|_1 \geq \frac{1}{s}\|A\|_2 = \frac{1}{s}\|B\|_2 \geq \frac{r}{s}\|B\|_1$ since two similar matrices have the same spectral norm. ∎

Lemma 4: If A, B are positive semidefinite, then $\mathrm{Tr}(AB)$ is nonnegative.

Proof: We have $A = A^{\frac{1}{2}}A^{\frac{1}{2}}$, hence $\mathrm{Tr}(AB) = \mathrm{Tr}(A^{\frac{1}{2}}A^{\frac{1}{2}}B)$ which is equal to $\mathrm{Tr}(A^{\frac{1}{2}}BA^{\frac{1}{2}})$. Since $A^{\frac{1}{2}}BA^{\frac{1}{2}}$ is positive semidefinite, its trace is nonnegative. ∎

Proposition 1: Providing that $\lambda > 0$, the general minimization problem (4) is bounded below.

Proof: Since $\mathscr{S}_i(S), \Sigma_i(\mu), i = 1, \cdots, n$ are positive semidefinite matrices for all S and μ, then $\mathrm{Tr}(\Sigma_i(\mu)\mathscr{S}_i(S))$ is always nonnegative. Hence problem (4) is bounded below by $\min \frac{1}{n}\sum_i\{-\log|\mathscr{S}_i(S)| + \lambda\|S\|_1\}$. Introducing the auxiliary variables $Z_i = \mathscr{S}_i(S)$, the bound can be rewritten as $\min \frac{1}{n}\sum_{i=1}^n(-\log|Z_i| + \lambda\|S\|_1)$ s.t. $S \succ 0, Z_i = \mathscr{S}_i(S), i = 1, \cdots, n$. From the proof of Lemma 1, the matrices S and $\begin{bmatrix} \mathscr{S}_i(S) & 0 \\ 0 & S_{z_iz_i} \end{bmatrix}$ are similar, hence $\|S\|_1 \geq \alpha(\|\mathscr{S}_i(S)\|_1 + \|S_{z_iz_i}\|_1) \geq \alpha\|\mathscr{S}_i(S)\|_1$ from Lemma 3. Hence problem (8) is bounded by $\min \frac{1}{n}\sum_{i=1}^n\{-\log|Z_i| + \lambda\alpha\|Z_i\|_1\}$ s.t. $Z_i = \mathscr{S}_i(S)$ which can be relaxed in a sum of n programs $\frac{1}{n}\sum_{i=1}^n \min_{Z_i\succ 0}\{-\log|Z_i| + \lambda\alpha\|Z_i\|_1\}$. Since $\lambda\alpha > 0$, each program has a unique solution $\hat{Z}_i$ from [2, Th. 1]: $\operatorname{argmin}_{Z_i\succ 0}\{-\log|Z_i| + \lambda\alpha\|Z_i\|_1\} = \hat{Z}_i$. Hence (8) is bounded below by $\frac{1}{n}\sum_{i=1}^n \hat{Z}_i$. ∎

The ℓ_1-penalty $\lambda\|S\|_1$ promotes sparsity when learning the precision matrix, without altering the DC structure of the objective. With $\lambda \geq 0$, $\lambda\|S\|_1$ is also a convex function, so it can be added to the convex part f_0 of the DC program. This generalizes to any regularization of the form $\lambda\|S\|$ where $\|\cdot\|$ is a norm (e.g. the Euclidian norm $\|\cdot\|_2$ for ridge regression) since all norms are convex.

To conclude the section, the note below Lemma 1 gives a simplification of the objective in (4):

$$-\log|S| + \frac{1}{n}\sum_{i=1}^n(\log|S_{i,\mathrm{mis}}| + \mathrm{Tr}(\Sigma_i\mathscr{S}_i(S))) + \lambda\|S\|_1 \tag{9}$$

This reformulation will be useful in Section IV.

III. Concave-Convex procedure

The Concave Convex Procedure (CCCP) computes a stationary point of DC programs by solving a sequence of convex programs (see [23] for further details). Here, we present the general framework of the CCCP along with some convergence guarantees of the algorithm. CCCP solves problems of the form

$$\min\ f_0(x) - g_0(x) \quad \text{s.t. } x \in \mathscr{C} \tag{10}$$

where f_0 and g_0 are convex and $\mathscr{C}$ is some convex set. Assuming g is differentiable at every iteration, CCCP solves a sequence of convex programs by linearizing g_0 about the current best estimate $x^{(t)}$ in order to obtain the next point $x^{(t+1)}$, which is solution of

$$\min\ f_0(x) - g_0(x^{(t)}) - \nabla g_0(x^{(t)})^T(x - x^{(t)}) \quad \text{s.t. } x \in \mathscr{C} \tag{11}$$

Proposition 2: Let $h_0^{(t)}(x)$ be the objective function in (11). Assuming that the minimization problem (10) is bounded, the convex program (11) is bounded. Moreover, solving the convex program (11) decreases the objective function in (10).

Proof: Using the first order condition for convex functions $h_0^{(t)}(x) \geq f_0(x) - g_0(x)$, problem (11) is also bounded. Now let us assume that we can solve (11) at each iteration of CCCP. Let $x^{(t+1)}$ be a solution to it at iteration t. Then we have $(f_0 - g_0)(x^{(t+1)}) \leq h_0^{(t)}(x^{(t+1)}) \leq h_0^{(t)}(x^{(t)}) = (f_0 - g_0)(x^{(t)})$. Hence the objective function is non-increasing. ∎

Therefore, CCCP is guaranteed to converge. More importantly, we apply block-coordinate CCCP to solve problem (4) by alternating between blocks μ and S and applying one step of CCCP when minimizing with respect to S (the problem is convex in μ).

When applying one step of CCCP to (5), the variable x becomes a symmetric matrix S, and the feasible set is $\mathscr{C} = \mathscr{S}^n_{++}$. Since the concave part of our objective $g_0(S) = -\frac{1}{n}\sum_{i=1}^n \mathrm{Tr}(\Sigma_i\mathscr{S}_i(S))$ is smooth, we can take the first order Taylor expansion of g_0 at $S^{(t)}$, that is $g_0(S) \approx g_0(S^{(t)}) + \mathrm{Tr}\left((\nabla_S g_0)_{S^{(t)}}(S - S^{(t)})\right)$. The sequence of convex programs about the current best estimate $S^{(t)}$ is:

$$S^{(t+1)} = \operatorname*{argmin}_{S\succ 0}\{f_0(S) - \mathrm{Tr}(D_{(t)}^T S) + \lambda\|S\|_1\} \tag{12}$$

where $D_{(t)}$ is the gradient of the concave part of the objective:

$$D_{(t)} := (\nabla_S g_0)_{S^{(t)}} = -\frac{1}{n}\sum_{i=1}^{n}(\nabla_S \mathrm{Tr}(\Sigma_i \mathscr{S}_i(S))_{S^{(t)}} \qquad (13)$$

The difference between the approaches in [24], [15] and our work stems from the sequence of convex programs solved. While EM imputes missing values to approximate the complete log-likelihood, we place the emphasis on the DC decomposition. We now derive a closed-form expression for $D_{(t)}$ in (12):

Proposition 3: With permutation matrix P_i in (6), and denoting $E_i := B_i D_i^{-1} = S_{y_i z_i}(S_{z_i z_i})^{-1}$, we have

$$\nabla_S \mathrm{Tr}(\Sigma_i \mathscr{S}_i(S)) = P_i^T \begin{pmatrix} \Sigma_i & -\Sigma_i E_i \\ -E_i^T \Sigma_i & E_i^T \Sigma_i E_i \end{pmatrix} P_i \qquad (14)$$

Proof: Consider the block matrix $S = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ and the Schur complement $\mathscr{S}(S) = A - BD^{-1}C$. Using $\mathrm{Tr}(M\mathscr{S}(S)) = \mathrm{Tr}(MA) - \mathrm{Tr}(MBD^{-1}C)$ we have:
$\nabla_S \mathrm{Tr}(M\mathscr{S}(S)) = \begin{pmatrix} \nabla_A \mathrm{Tr}(M\mathscr{S}(S)) & \nabla_B \mathrm{Tr}(M\mathscr{S}(S)) \\ \nabla_C \mathrm{Tr}(M\mathscr{S}(S)) & \nabla_D \mathrm{Tr}(M\mathscr{S}(S)) \end{pmatrix} = \begin{pmatrix} \nabla_A \mathrm{Tr}(MA) & -\nabla_B \mathrm{Tr}(MBD^{-1}C) \\ -\nabla_C \mathrm{Tr}(MBD^{-1}C) & -\nabla_D \mathrm{Tr}(MBD^{-1}C) \end{pmatrix}$
$= \begin{pmatrix} M^T & -M^T C^T D^{-T} \\ -D^{-T} B^T M^T & D^{-T} B^T M^T C^T D^{-T} \end{pmatrix}$, where the last equality is obtained using these facts:
$\nabla_X \mathrm{Tr}(M_1 X M_2) = M_1^T M_2^T$,
$\nabla_X \mathrm{Tr}(M_1 X^{-1} M_2) = -X^{-T} M_1^T M_2^T X^{-T}$.
Finally, we substitute $M, A, D, C = B^T$ and order the lines and columns back to the initial order. ∎

Eq. (14) gives a closed-form expression for the objective in (12). We conclude the section with the following result:

Proposition 4: As long as $\lambda > 0$, the minimization problem (12) is convex and bounded.

Proof: Since $\lambda > 0$, the convex program (12) is bounded by problem (8) from Proposition 2. ∎

IV. Comparison with EM

In this section, we compare the different approaches taken by EM and our CCCP algorithm, which we call m-CCCP for the rest of the article. We show that EM is also a choice of decomposition CCCP for Gaussians with Missing Values in which a term of the form of a log barrier function on $S_{i,\mathrm{mis}}$ is included in the concave part.

We recall that when we re-order the lines and columns of the covariance Σ and its inverse S using a permutation P_i such that the first block corresponds to the observed values $x_{i,\mathrm{obs}}$, we obtain $x_i = P_i^T \begin{pmatrix} y_i \\ z_i \end{pmatrix}$, where $y_i = x_{i,\mathrm{obs}}$ are the observed entries of x_i, and $z_i = x_{i,\mathrm{mis}}$ are the missing values.

In EM, a probability distribution of the missing values z_i is inferred, conditioned on the observed values x_i and based on some current estimate of the precision matrix S. It is then used to optimize the score function over the entire data set. The probabilities computed for z_i are the posterior probabilities computed in the E step, and the new parameter values are computed in the M step.

$$\begin{aligned} \text{E:}\quad & \rho(\{z_i\}_i) := f(\{z_i\}_i \,|\, \{y_i\}_i, \mu', S') \\ & Q(\mu, S|\mu', S') := \mathbb{E}_{z\sim\rho}[-\ell(\{y_i\}_i, \{z_i\}_i; \mu', S')] \\ \text{M:}\quad & \mu^+ := \arg\min_{\mu} Q(\mu, S|\mu', S') \\ & S^+ := \arg\min_{S\succ 0} Q(\mu, S|\mu^+, S') \end{aligned} \qquad (15)$$

where the likelihood ℓ is defined by $-\ell(\{y_i\}_i, \{z_i\}_i; \mu, S) = -\log|S| + \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^T S (x_i-\mu)$. Since the data points z_i are i.i.d., the distribution ρ in step E decomposes as $\Pi_{i=1}^n \rho_i$ where $\rho_i(z_i) := f(z_i|y_i)$ and $z_i|y_i$ is a Gaussian with conditional mean $\mu_{i,\mathrm{mis}} - E_i^T(y_i - \mu_{i,\mathrm{obs}})$ (where $E_i = S_{y_i z_i}(S_{z_i z_i})^{-1}$ as in (14)) and covariance $(S_{z_i z_i})^{-1}$. Hence the objective in step M of EM decomposes as $-\log|S| + \frac{1}{n}\sum_{i=1}^n \mathrm{Tr}(\mathbb{E}_{z_i\sim\rho_i}[x_i x_i^T] S)$. Hence, the E step is:

$$\begin{aligned} & \rho(z_i) \sim \mathscr{N}(\mu'_{i,\mathrm{mis}} - E_i^T(y_i - \mu'_{i,\mathrm{obs}}), S'^{-1}_{z_i z_i}) \\ & Q(\mu, S|\mu', S') := -\log|S'| + \tfrac{1}{n}\textstyle\sum_{i=1}^n \mathrm{Tr}(\mathbb{E}_{z_i\sim\rho_i}[(x_i-\mu')(x_i-\mu')^T] S') \end{aligned} \qquad (16)$$

Beyond the statistical interpretation of EM presented above, we now show that this algorithm is also a CCCP similar to the one presented in Section II. To the best of our knowledge, this is a new result on the problem of inverse covariance estimation from data with missing values. This stems from the following key observation on the average of the sufficient statistics $x_i x_i^T$ over the posterior probabilities:

Lemma 5: $\mathbb{E}_{z_i\sim\rho_i}[(x_i-\mu)(x_i-\mu)^T] = \nabla_S(\mathrm{Tr}(\Sigma_i \mathscr{S}_i(S)) + \log|S_{i,\mathrm{mis}}|)$, $i = 1, \cdots, n$.

Proof: We have $\mathbb{E}_{z_i\sim\rho_i}[(x_i-\mu)(x_i-\mu)^T] = P_i^T \begin{pmatrix} (y_i-\mu_{i,\mathrm{obs}})(y_i-\mu_{i,\mathrm{obs}})^T & (y_i-\mu_{i,\mathrm{obs}})\mathbb{E}[z_i-\mu_{i,\mathrm{mis}}]^T \\ \mathbb{E}[z_i-\mu_{i,\mathrm{mis}}](y_i-\mu_{i,\mathrm{obs}})^T & \mathbb{E}[(z_i-\mu_{i,\mathrm{mis}})(z_i-\mu_{i,\mathrm{mis}})^T] \end{pmatrix} P_i$.
Plugging in
(i) $(y_i-\mu_{i,\mathrm{obs}})(y_i-\mu_{i,\mathrm{obs}})^T = \Sigma_i$
(ii) $\mathbb{E}[z_i-\mu_{i,\mathrm{mis}}] = -E_i^T(y_i-\mu_{i,\mathrm{obs}})$
(iii) $\mathbb{E}[(z_i-\mu_{i,\mathrm{mis}})(z_i-\mu_{i,\mathrm{mis}})^T] = \mathbb{E}[z_i z_i^T] + \mu_{i,\mathrm{mis}}\mu_{i,\mathrm{mis}}^T - \mu_{i,\mathrm{mis}}\mathbb{E}[z_i]^T - \mathbb{E}[z_i]\mu_{i,\mathrm{mis}}^T = (S_{z_i z_i})^{-1} + (\mathbb{E}[z_i]-\mu_{i,\mathrm{mis}})(\mathbb{E}[z_i]-\mu_{i,\mathrm{mis}})^T = (S_{z_i z_i})^{-1} + E_i^T \Sigma_i E_i$
where we used $\mathbb{E}[z_i z_i^T] = (S_{z_i z_i})^{-1} + \mathbb{E}[z_i]\mathbb{E}[z_i]^T$ in (iii), gives $\mathbb{E}[x_i x_i^T] = P_i^T \begin{pmatrix} \Sigma_i & -\Sigma_i E_i \\ -E_i^T\Sigma_i & E_i^T\Sigma_i E_i + (S_{z_i z_i})^{-1} \end{pmatrix} P_i$.
We recognize in (14) the term $\nabla_S \mathrm{Tr}(\Sigma_i \mathscr{S}_i(S))$ plus an additional term $(S_{z_i z_i})^{-1}$. Observing that $\nabla_S \log|S_{z_i z_i}| = P_i^T \begin{pmatrix} 0 & 0 \\ 0 & (S_{z_i z_i})^{-1} \end{pmatrix} P_i$ finishes the proof. ∎

In words, the conditional expectation of $(x_i-\mu)(x_i-\mu)^T$ is the gradient of a concave part of the objective rewritten in (9), that is $\mathrm{Tr}(\Sigma_i \mathscr{S}_i(S)) + \log|S_{i,\mathrm{mis}}|$.

Proposition 5: For Gaussians, the optimization with respect to S in step M of EM is a decomposition CCCP:

$$\begin{aligned} & \min_{S\succ 0} f_1(S) - g_1(S) \\ & \quad f_1(S) = -\log|S| \\ & \quad g_1(S) = -\tfrac{1}{n}\textstyle\sum_{i=1}^n (\log|S_{i,\mathrm{mis}}| + \mathrm{Tr}(\Sigma_i \mathscr{S}_i(S))) \end{aligned} \qquad (17)$$

Proof: From the results in Section II, f_1 and g_1 are both convex. Plugging in the result from Lemma 5 in step M of EM for Gaussians in (16), the objective becomes $f_1(S) - \mathrm{Tr}(\nabla_S g_1(S) S)$. ∎

Note that for simplicity the ℓ_1-penalty term is not included in the analysis because it does not affect the CCCP decompositions discussed in this section and the next one. We have shown that EM is a CCCP. Since EM linearizes more terms in the objective than m-CCCP, the objective in EM is a weaker approximation of the true objective. In fact, we show in the next section that m-CCCP gives a tighter upper bound than EM on the objective of the minimization problem with respect to block S.

V. Convergence analysis

From the results presented in the previous section, EM can be seen as a block descent in which the minimization with respect to μ is done by imputation of the missing values by their conditional means, (or simply the minimization with respect to μ is one step of an EM with μ as the only unknown parameters), and the minimization with respect to S uses CCCP with the DC structure (17).

In m-CCCP, we do not restrict the minimization with respect to μ to follow a particular method. However, the minimization with respect to S is assumed to follow the DC structure (8). We apply the same analysis as the one made in section III, and show that m-CCCP provides a tighter upper bound to the true objective than EM.

Defining the convex function $g_2(S) := g_1(S) - g_0(S) = -\frac{1}{n}\sum_{i=1}^{n} \log|S_{i,\text{mis}}|$ the objective becomes

$$\arg\min_{S\succ 0} \{f_1(S) - g_2(S) - g_0(S)\} \tag{18}$$

where f_1, g_2, and g_0 are convex. Denoting the linear map $\mathscr{L}$ such that $\mathscr{L}f$ is the 1st-order Taylor expansion of f at x_0, the objectives in m-CCCP and EM at x_0 are:

$$\begin{array}{ll} \text{m-CCCP:} & f_0 - \mathscr{L}g_0 = f_1 - g_2 - \mathscr{L}g_0 \\ \text{EM:} & f_1 - \mathscr{L}g_1 = f_1 - \mathscr{L}g_2 - \mathscr{L}g_0 \end{array} \tag{19}$$

Hence, EM linearizes the concave part $-g_2$ of the objective in m-CCCP. In other words, m-CCCP applies CCCP to the true objective, and EM applies CCCP to the objective in m-CCCP:

$$\begin{array}{l} (f_0 - \mathscr{L}g_0) - (f_0 - g_0) = g_0 - \mathscr{L}g_0 \geq 0 \\ (f_1 - \mathscr{L}g_1) - (f_0 - \mathscr{L}g_0) = g_2 - \mathscr{L}g_2 \geq 0 \end{array} \tag{20}$$

where the inequalities are obtained from the first-order condition on the convex functions g_0 and g_2. As a consequence:

Proposition 6: The convex program solved by EM at x_0 is an upper bound on the convex program solved by m-CCCP at x_0:

$$f_1 - g_1 = f_0 - g_0 \leq f_0 - \mathscr{L}g_0 \leq f_1 - \mathscr{L}g_1 \tag{21}$$

Thus, m-CCCP provides a tighter upper bound on the true objective (4) than EM, and $\min_{S\succ 0}(f_0 - g_0)(S) \leq \min_{S\succ 0}(f_0 - \mathscr{L}g_0)(S) \leq \min_{S\succ 0}(f_1 - \mathscr{L}g_1)(S)$.

VI. Experiments on synthetic datasets

We present several evaluations of our algorithm against current state of the art for missing data. Our first set of experiments evaluate m-CCCP against EM on standard synthetic datasets. In the case of EM, the maximization step is implemented using the QUIC algorithm [14] run until convergence.

In a first synthetic experiment, we study the rate of convergence of m-CCCP compared with EM: given an inverse covariance matrix, we generate small subsets of the covariance matrix as synthetic observations. This corresponds to the ideal case in which we observe empirical covariances without sampling noise. Given these covariance subsets, we run m-CCCP and EM to attempt to recover the original precision matrix (assuming the sparsity pattern is known). Figure 1 presents an example in which $n = 40$, $m = 50$. As one can see (this is confirmed in subsequent experiments), (1) m-CCCP reaches a good local minimum in the first iteration, while EM takes a substantial number of iterations to reach the same level, and (2) after having found a local minimum, m-CCCP performs more progress per iteration.

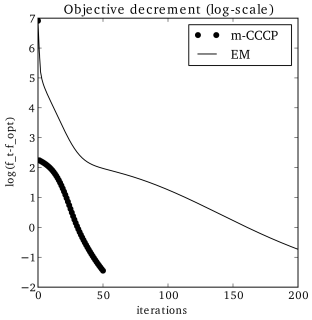

Fig. 1. Comparing m-CCCP and EM.

In a second experiment, we consider model 1, model 2, model 3, and model 4 of [20] with $p = 10, 50, 100$, $x_1, \cdots, x_n \sim \mathscr{N}(0, \Sigma)$ with

Model 1: $n = 100$, $\Sigma_{ij} = 0,7^{|j-i|}$, so that the entries of the covariance matrix decay exponentially.

Model 2: $n = 150$,, $\Sigma_{ij} = \mathbb{I}_{\{i=j\}} + 0.4\mathbb{I}_{\{|i-j|=1\}}$ $+0.2\mathbb{I}_{\{|i-j|=2\}} + 0.2\mathbb{I}_{\{|i-j|=3\}} + 0.1\mathbb{I}_{\{|i-j|=4\}}$ where $\mathbb{I}_C$ is the indicator function which is 1 if condition C is true and 0 otherwise.

Model 3: $n = 200$, $\Sigma = B + \delta I$, where each off-diagonal entry of B is generated independently and equals 0.5 with probability $\alpha = 0.1$ or 0 with probability $1 - \alpha$. Diagonal entries of B are zero, and δ is chosen so that the condition number of Σ is p.

Model 4: $n = 250$, same as model 3 except $\alpha = 0.5$.

Note that in all models Σ^{-1} is sparse. In models 1 and 2 the number of non-zeros in Σ^{-1} is linear in p, whereas in models 3 and 4 it is proportional to p^2.

For all 12 settings (4 models with $p = 10, 50, 100$) we perform 20 simulation runs. In each run we proceed as follows:

- We generate n training observations from the model.
- In the training set we delete uniformly at random 20%, 40%, 60% and 80% of the data. Per setting, hence we get four training sets with different degree of missing data, for a total of 48 training sets.

- The m-CCCP estimator is fitted on each of the three mutilated training sets, with the tuning parameter λ selected by minimizing the BIC criterion.

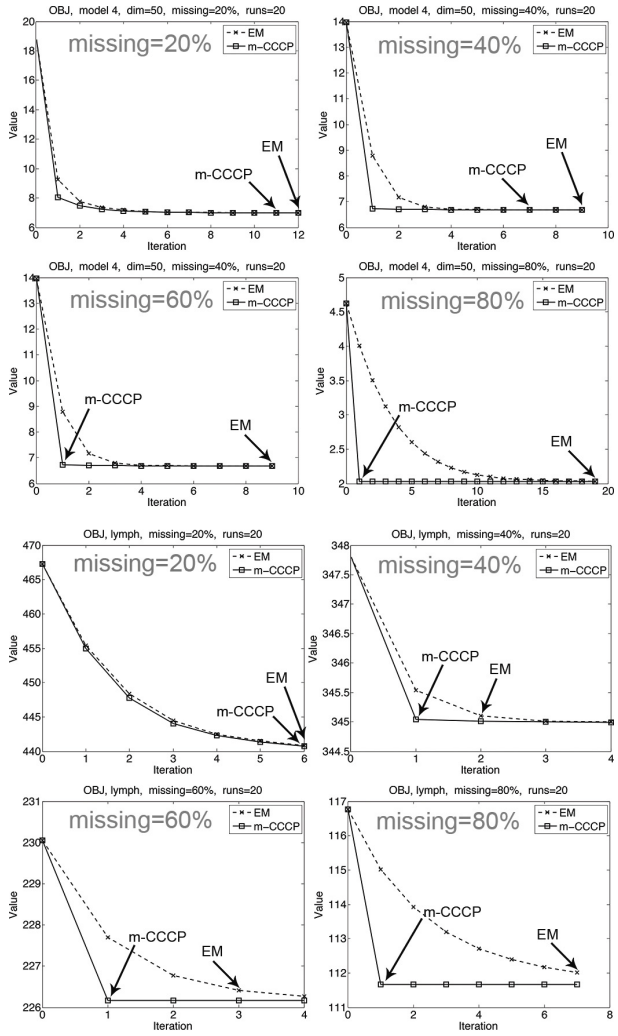

Fig. 2. Top four: Objectives of m-CCCP and EM averaged over 20 runs for model 4 and p=50. Bottom four: objectives of m-CCCP and EM on lymph dataset. The arrows mark the convergence of both algorithms.

We note that since we sample from a Gaussian distribution with zero-mean, there is no minimization with respect to block μ. Hence, both m-CCCP and EM are exactly CCCP algorithms.

Tuning parameter: Recall that the tuning parameter λ described in the previous sections is required to control the sparsity of the solution to fit the data better. A common approach is to form a grid of exponentially increasing values for λ and choose to minimize a modified BIC criterion:
$\mathrm{BIC}(\lambda) = -2\ell(\{x_{i,\mathrm{obs}}\}_i; \hat{\mu}, \hat{S}) + \log(n) \sum_{i \le j} \mathbb{I}_{\{\hat{S}_{ij} \neq 0\}}$
where $(\hat{\mu}, \hat{S})$ are estimated obtained using the tuning parameter λ, $-2\ell(\{x_{i,\mathrm{obs}}\}_i; \hat{\mu}, \hat{S}) = n(f_0(\hat{S}) - g_0(\hat{S}))$ is the observed log-likelihood (modulo some constants), and $\sum_{i \le j} \mathbb{I}_{\{\hat{S}_{ij} \neq 0\}}$ measures the degrees of freedom.

We observe that for percentages of missing values $20\%, 40\%, 60\%$, the tuning parameters that minimize the BIC criterion results in good estimates. However, for 80% of missing values, minimizing the BIC criterion often gives excessively large tuning parameters λ which results in poor estimates. Hence we minimize the BIC criterion under the constraint $\lambda \le \lambda_{\max}$ where $\lambda_{\max}$ is the maximum tolerated value for the tuning parameters λ.

Initialization: Following the work of Städler and Bühlmann [24], we initialize m-CCCP and EM with the full samples $x_i, i = 1, \cdots, n$ by imputing the missing values by their row means. Since this imputation scheme can be quite good, this can result in a fast convergence of both m-CCCP and EM in number of iterations.

We assume m-CCCP and EM converge when the relative change of objective value is less than 10^{-3} and we report in Table I the number of iterations required for convergence for both algorithms. m-CCCP converges faster in number of iterations for all 48 training sets. The gap is large for 60% and 80% of missing values because the weight of $g_2 = -\frac{1}{n}\sum_{i=1}^{n} \log |S_{i,\mathrm{mis}}|$ in (19) is more important, hence by linearizing out g_2, EM provides a weaker approximation of the objective.

We note that m-CCCP and EM consistently give *very close* estimates of the inserve covariance matrix $\hat{S}$, and same values of $\|\hat{S} - S\|_F$ where S is the true precision matrix and $\|\cdot\|_F$ is the Frobenius norm (results not reported here).

VII. Experiments on real datasets

	miss	model 1		model 2		model 3		model 4	
	(%)	**mC**	EM	**mC**	EM	**mC**	EM	**mC**	EM
p=10	20	**5**	8	**6**	7	**5**	6	**1**	2
	40	**11**	16	**11**	13	**8**	10	**1**	3
	60	**1**	5	**1**	4	**13**	18	**1**	5
	80	**1**	10	**1**	8	**2**	10	**1**	8
p=50	20	**8**	10	**9**	10	**8**	9	**11**	12
	40	**16**	20	**3**	4	**16**	18	**7**	9
	60	**1**	7	**13**	17	**16**	21	**1**	10
	80	**1**	15	**4**	13	**3**	15	**1**	19
p=100	20	**8**	12	**11**	12	**11**	12	**18**	19
	40	**22**	26	**25**	29	**14**	16	**6**	8
	60	**4**	8	**1**	7	**37**	49	**1**	13
	80	**1**	16	**1**	14	**1**	20	**1**	29

	miss	arabidop.		leukemia		lymph		-	
	(%)	**mC**	EM	**mC**	EM	**mC**	EM	**mC**	EM
	20	**7**	7	**9**	9	**6**	6	**-**	-
	40	**1**	2	**1**	2	**1**	2	**-**	-
	60	**1**	3	**1**	4	**1**	3	**-**	-
	80	**1**	8	**1**	9	**1**	7	**-**	-

Table I: Number of iterations for convergence of the algorithms: m-CCCP outperforms EM for all 51 datasets (synthetic and real).

Following experiments done by [14] and [24], we use the biology datasets preprocessed by [16] to compare the performance of our algorithm with EM, in the hypothetical case that values were missing from data. This is an interesting case in practice, as collecting hundreds of biological

parameters for each experiment may become expensive. We first decimate the data at random, and then perform centering and variance scaling using the observed data points.

For all three biology datasets (arabidopsis, leukemia, and lymph) we set $\lambda = 0.5$ as in [14], and we use a convergence threshold of 10^{-3} on the change of objective value for both m-CCCP and EM. The number of iterations required for convergence are reported in Table I. In all cases, the number of CCCP iterations is (much) lower than the number of iterations required by EM. In particular, when most of the data is unobserved ($> 80\%$ missing values), m-CCCP converges in one iteration to a local minimum, while EM requires many more iterations.

VIII. Conclusion and future work

The problem of learning sparse inverse covariance from incomplete data is proven to be a difference of convex problem. When the data is sparingly observed, the Expectation-Maximization algorithm leads to slow convergence in number of iterations. Based on the observation that the determinant of a Schur complement is a log-concave function, we propose a new Concave-Convex procedure that shows superior convergence results on standard synthetic datasets. We are currently working on extending these results to larger problems by exploiting the structure of the Schur complement in the case of small observations combined with a quadratic approximation similar to the state-of-the-art QUIC algorithm [14], the aim being to develop an algorithm that is faster than EM in runtime.

IX. Acknowledgments

The authors gratefully thank Prof. Suvrit Sra from Carnegie Mellon University and Prof. Laurent El Ghaoui from Univesity of California at Berkeley for their insightful discussions

References

[1] L. An and P. Tao. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Oper. Res.*, 133:23–46, 2005.

[2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or Binary data. *J. Mach. Learn. Res.*, 9:485–516, 2008.

[3] Rajendra Bhatia. *Positive Definite Matrices*. Princeton in Applied Mathematics, December 18 2006.

[4] J. Bien and R. J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98, 4:807–820, 2011.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, March 8 2004.

[6] V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40:1935–1967, 2012.

[7] J. DeLeeuw and P. Mair. Multidimensional scaling using majorization: SMACOF in R. *J. Statist. Software*, 31:1–30, 2009.

[8] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[9] A. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196–212, 2004.

[10] E. Dougherty, A. Datta, and C. Sima. Research issues in genomic signal processing. *IEEE Signal Processing Magazine*, 22:46–68, 2005.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9:432–441, 2007.

[12] L. Gu, E. P. Xing, and T. Kanade. Learning GMRF structures for spatial priors. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, 2007.

[13] R. Horst and N. V. Thoai. DC programming: Overview. *J. Optimiz.*, 103:1–43, 1999.

[14] C-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar. Sparse Inverse Covariance Matrix Estimation Using Quadratic Approximation. *Advances in Neural Information Processing Systems (NIPS)*, page 24, 2011.

[15] M. Kolar and E. P. Xing. Estimating Sparse Precision Matrices from Data with Missing Values. *Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK*, 2012.

[16] L. Li and K.-C. Toh. An inexact interior point method for l1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2:291–315, 2010.

[17] R. Little and D. Rubin. *Statistical analysis with missing data*. Wiley, New York, 1987.

[18] D. M. Malioutov, J. K. Johnson, M. J. Choi, and A. S. Willsky. Low-rank variance approximation in gmrf models: Single and multiscale approaches. *EEE Transactions on Signal Processing*, 56(10):4621–4634, 2008.

[19] G. D. Murray. Comments on "Maximum likelihood from incomplete data via the EM algorithm" by Dempster, Laird, and Rubin. J. R. *Stat. Soc., Ser. B 39*, pages 27–28, 1977.

[20] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu. Sparse Permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[21] H. Rue and L. Held. Gaussian Markov random fields: theory and applications. *CRC Press*, 2005.

[22] J. L. Schafer. Analysis of Incomplete Multivariate Data. *Monographs on Statistics and Applied Probability, Chapman and Hall, London*, 72, 1997.

[23] B. K. Sriperumbudur and G. R. G. Lanckriet. On the Convergence of the Concave-Convex Procedure. *NIPS*, 2009.

[24] N. Stadler and P. Buhlmann. Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, pages 1–17, 2009.

[25] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. SpringeSpringer., 2002.

[26] C. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11:95–103, 1983.

[27] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika 94*, pages 19–35, 2007.

[28] A. L. Yuille and A. Rangarajan. The Concave-Convex Procedure. *Neural Computation*, 15:915–36, 2003.

[29] T. Zhang. Analysis of multi-stage convex relaxation for sparse regularization. 11:1081–107, 2010.