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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of estimating sparse precision matrices from data with missing values. Direct statistical inference with likelihoods of observed values is a minimization program that uses the Schur complement. The objective function is not convex but rather a "difference of convex" program (DC program) which can be solved with the Concave-Convex procedure (CCP). The technique presented uses a concave-convex decomposition that is more natural than the one used in Expectation-Maximization (EM) algorithms, and simulation studies also show that the CCP compares favorably to EM.

matrices and their inverses, precision matrices. 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), 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. Promoting sparsity has compelling advantages, such as producing more robust models that generalize well to unseen data [1] or uncovering the interactions between variables [2]. It 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 [3, 4, 5, 6] 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 $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ the empirical covariance of the dataset $\mathcal{D} = \{x_i\}_i$, these algorithms solve the problem:

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

where $|S|$ is the determinant of matrix S, Tr the trace operator, and $\|S\|_1 = \sum_{ij}|S_{ij}|$.

In practice, datasets often suffer from missing values 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* for short. Inference for μ and S can be based on the observed log-likelihood if we assume that the underlying missing data mechanism is ignorable. With an arbitrary pattern of missing values, no explicit maximization of the likelihood is possible even for the mean values and covariance matrices [7]. Concretely, let x_i be an observation of the *full-dimensional* random vector $X \sim \mathcal{N}(\mu, \Sigma)$, and $x_{i,\text{obs}}$ a *marginal* (or subset) of the full distribution: $x_{i,\text{obs}}$ is also drawn from the marginal Gaussian distribution with mean $\mu_{i,\text{obs}}$ and covariance $\Sigma_{i,\text{obs}}$. The marginal mean $\mu_{i,\text{obs}}$ and covariance $\Sigma_{i,\text{obs}}$ are obtained by dropping the irrelevant variables from μ and Σ. Calling $\mathcal{S}_i(S) := (\Sigma_{i,\text{obs}})^{-1}$ the precision matrix of the marginal distribution, and $\hat{\Sigma}_{i,\text{obs}} = (x_{i,\text{obs}} - \mu_{i,\text{obs}})(x_{i,\text{obs}} - \mu_{i,\text{obs}})^T$ the empirical covariance, the log-likelihood of the observations for this subset of samples is $-\log|\mathcal{S}_i(S)| + \mathrm{Tr}(\hat{\Sigma}_{i,\text{obs}}\mathcal{S}_i(S))$,

and the original problem (1) becomes the new problem of minimizing the objective function:

$$\frac{1}{n}\sum_{i=1}^{n}\Big(-\log|\mathcal{S}_i(S)| + \mathrm{Tr}(\hat{\Sigma}_{i,\mathrm{obs}}\mathcal{S}_i(S))\Big) + \lambda\|S\|_1 \tag{2}$$

Due to the structure of $\mathcal{S}_i(S)$, the observed log-likelihood is a non-convex function of S for a general missing data pattern, with possible existence of multiple stationary points [8], [9]. Thus, optimization of (2) is a non-trivial problem.

Our central observation is that the problem of minimizing (2), although nonconvex, has an objective function which can be decomposed into the sum of a convex and a concave function. Hence we have a "difference of convex" (DC) program for which the ConCave-Convex procedure (CCCP) is a natural method for performing the optimization [10].

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 [11, 12]; majorization-minimization algorithms for problems in statistics such as least-squares multidimensional scaling [13]; regularized regression with nonconvex penalties [14]. Finally, the CCCP algorithm has been used in various machine learning applications and studied theoretically [10, 15].

We present several contributions. We propose a novel approach called m-CCCP to solve the problem of minimizing (2) which differs from previous works [16, 17]. The emphasis is placed on the DC decomposition of the log-likelihood rather than the statistical analysis used in the Expectation-Maximization method (EM) typically used in this setting. Nonetheless, an interesting connection is drawn between the two methods: EM is also a CCCP method using a different DC program. This provides a powerful analytical framework for comparing the two algorithms, which 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.

The article is organized as follows. We first show that the *determinant of the Schur complement* is a *log-concave function*, which is of interest by itself when optimizing marginal likelihoods of Gaussian distributions, and in convex optimization in general. As a consequence, the objective function (2) is a *difference of convex functions* and can be optimized using the CCCP algorithm. We then present the difference with EM and we compare our new algorithm against EM on standard synthetic datasets as well as gene expression datasets. In both cases, CCCP compares favorably in terms of convergence speed and quality of output.

1 Missing values as a DC program

We begin by noting that the problem of minimizing (2) is a DC program. We will first see that the marginal precision $\mathcal{S}_i(S) = (\Sigma_{i,\mathrm{obs}})^{-1}$ is a Schur complement from S (after permuting the variables). The DC decomposition will ensue by proving some new properties of the Schur complements.

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 $x_{i,\mathrm{obs}}$, then the inverse of the observed block in Σ is a Schur complement. In other words, with $x_i := P_i^T\begin{pmatrix} y_i \\ z_i\end{pmatrix}$, where $y_i := x_{i,\mathrm{obs}}$ and $z_i = x_{i,\mathrm{mis}}$, it follows that $P_i\Sigma P_i^T = \begin{pmatrix}\Sigma_{y_iy_i} & \Sigma_{y_iz_i}\\ \Sigma_{z_iy_i} & \Sigma_{z_iz_i}\end{pmatrix}$ has inverse $P_iSP_i^T = \begin{pmatrix}S_{y_iy_i} & S_{y_iz_i}\\ S_{z_iy_i} & S_{z_iz_i}\end{pmatrix}$. where the inverse $\mathcal{S}_i(S) = (\Sigma_{y_iy_i})^{-1}$ is the Schur complement of the block $S_{y_iy_i}$ of the matrix $P_iSP_i^T$: $\mathcal{S}_i(S) = S_{y_iy_i} - S_{y_iz_i}S_{z_iz_i}^{-1}S_{z_iy_i}$. In the p-dimensional Hilbert space $\mathbb{R}^p$ with inner product x^Ty, 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 [18] (Theorem 1.3.3, Corollary 1.5.3) that the Schur complement (**??**) is concave. We present the following result (the proof is omitted for space constraints):

Theorem 1. *The function $S \mapsto log|\mathcal{S}(S)|$ is concave on S_{++}^p.*

Given the concavity of $\log|\mathcal{S}(S)|$ and the concavity of $x^T\mathcal{S}(S)x = \text{Tr}(xx^T\mathcal{S}(S))$, the problem of maximizing the log-likelihood can be restated as the following minimization problem:

$$\begin{aligned} &\min_{S\succ 0} f_0(S) - g_0(S) + \lambda\|S\|_1 \\ &\quad f_0(S) = -\tfrac{1}{n}\textstyle\sum_{i=1}^n \log|\mathcal{S}_i(S)| \\ &\quad g_0(S) = -\tfrac{1}{n}\textstyle\sum_{i=1}^n \text{Tr}(\hat{\Sigma}_{i,\text{obs}}\mathcal{S}_i(S)) \end{aligned} \tag{3}$$

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. The following proposition shows that the problem above has a solution (the proof is omitted):

Proposition 2. *Assuming that $\hat{\Sigma}_{i,obs} \succ 0$ for all i, the minimization problem (3) is bounded below.*

To promote sparsity when learning the precision matrix, we can also add the ℓ_1-penalty $\lambda\|S\|_1$, which does not alter the DC structure of the objective. So long as $\lambda \geq 0$ the regularization term $\lambda\|S\|_1$ is also a convex function, so it can be added to the convex part f_0 of the DC program.

2 A CCCP algorithm for solving the DC program

A local minimum for DC programs can be found with the Concave-Convex Procedure (CCCP). Here, we briefly present the general framework of the CCCP along with some convergence guarantees of the algorithm. CCCP attempts to solve the optimization problem by solving a sequence of convex programs [15], and it solves problems of the form

$$\min f_0(x) - g_0(x) \quad \text{s.t. } x \in \mathcal{C} \tag{4}$$

where f_0 and g_0 are convex and $\mathcal{C}$ is some convex set. Starting from a first estimate $x^{(0)}$, 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

$$\min f_0(x) - g_0(x^{(t)}) - \nabla g_0(x^{(t)})^T(x - x^{(t)}) \quad \text{s.t. } x \in \mathcal{C} \tag{5}$$

Proposition 3. *Let $h_0^{(t)}(x)$ be the objective function in (5). Assuming that the minimization problem (4) is bounded, the convex program (5) is bounded for all t. Moreover, if we can solve each convex program (5), the objective function in (4) is non-increasing at each iteration of CCCP and convergent.*

When CCCP is used to solve Problem (2), the variable x becomes a symmetric matrix S, and the feasible sets of the convex problems are also the same as the feasible set of the original problem, that is, the set of positive definite matrices $\mathcal{C} = \mathcal{S}^n_{++}$. Since the concave part of our objective $g_0(S) = -\sum_{i=1}^n \text{Tr}(\hat{\Sigma}_{i,\text{obs}}\mathcal{S}_i(S))$ is smooth, we can take the first order Taylor expansion of g_0 at $S^{(t)}$, that is $g_0(S) \approx g_0(S^{(t)}) + \text{Tr}\left((\nabla_S g_0)_{S^{(t)}}(S - S^{(t)})\right)$. The sequence of convex programs about the current best estimate $S^{(t)}$ is:

$$S^{(t+1)} = \operatorname*{argmin}_{S\succ 0} \{f_0(S) - \text{Tr}(D_{(t)}^T S) + \lambda\|S\|_1\} \tag{6}$$

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

Comparison with EM Expectation-Maximization is also a choice of decomposition CCCP for Gaussians with Missing Values. In fact, EM solves the following CCCP problem:

$$\begin{aligned} &\min_{S\succ 0} f_1(S) - g_1(S) \\ &\quad f_1(S) = -\log|S| \\ &\quad g_1(S) = -\tfrac{1}{n}\textstyle\sum_{i=1}^n \left(\log|S_{i,\text{mis}}| + \text{Tr}(\hat{\Sigma}_{i,\text{obs}}\mathcal{S}_i(S))\right) \end{aligned} \tag{7}$$

This other decomposition can also be recovered using the decomposability of the determinant: $\log|\mathcal{S}(S)| = \log|S| - \log|S_{\text{mis}}|$. Since EM linearizes more terms in the objective than m-CCCP, the objective in EM is a weaker approximation of the true objective. This intuition can be formalized in terms of convergence.

3 Evaluation

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 [19]. We found that in general: (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.

For synthetic experiments, we considered the datasets of [6] with $p = 10,\ 50,\ 100$, $x_1, \cdots, x_n \sim \mathcal{N}(0, \Sigma)$. These experiments present different sparsity patterns and condition numbers on the covariance matrix. For all 12 settings (4 models with $p = 10,\ 50,\ 100$) we perform 20 simulations. 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.

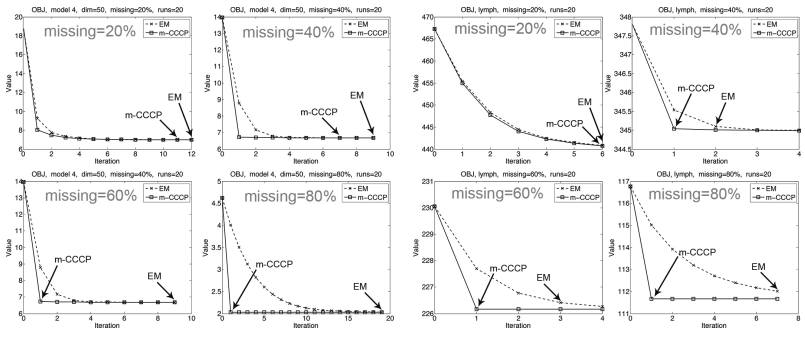

Figure 1: On the left, objective values of m-CCCP and EM averaged over 20 runs for model 4 and p=50. On the right, objective values of m-CCCP and EM on lymph dataset. The arrows indicate convergence (fixed at a threshold of 10^{-3}).

The gap is large for 60% and 80% of missing values because EM provides a weaker approximation of the objective. We note that m-CCCP and EM consistently give very close estimates of the inverse 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).

Following experiments done by [19] and [16], we use the biology datasets preprocessed by [20] to compare the performance of our algorithm with EM on real datasets, 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. 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 good local minimum, while EM requires many more iterations.

4 Conclusion

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. 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 [19].

References

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

[2] 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.

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

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

[5] 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.

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

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

[8] 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.

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

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

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

[12] 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.

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

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

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

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

[17] 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.

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

[19] 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.

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