Sparse Estimation of Inverse Covariance and Partial Correlation Matrices via Joint Partial Regression

Samuel Erickson [1] **Tobias Rydén** [2]

Abstract

We present a new method for estimating high-dimensional sparse partial correlation and inverse covariance matrices, which exploits the connection between the inverse covariance matrix and linear regression. The method is a two-stage estimation method wherein each individual feature is regressed on all other features while positive semi-definiteness is enforced simultaneously. We provide statistical rates of convergence for the proposed method which match, and improve upon, previous methods for inverse covariance and partial correlation matrix estimation, respectively. We also propose an efficient proximal splitting algorithm for numerically computing the estimate. The effectiveness of the proposed method is demonstrated on both synthetic and real-world data.

1. Introduction

Two important and closely related problems in statistical learning are the problems of estimating a partial correlation network and the inverse covariance matrix, also known as the precision matrix, from data. Partial correlation networks, which generalize the Gaussian graphical model, are used to model the relationships between variables while conditioning on all other variables, and are useful for inferring causal relationships between variables. Partial correlation networks are used in a plethora of applications, such as in the analysis of gene expression data (de la Fuente et al., 2004), and psychological data (Epskamp & Fried, 2018). The precision matrix, from which we can obtain the partial correlation network, is also of interest in its own right, as it also appears in linear discriminant analysis (Hastie et al., 2009) and in Markowitz portfolio selection (Markowitz, 1952). However, due to the high-dimensionality of the problem, estimating a precision or partial correlation matrix is often challenging as the number of parameters are on the order of the squared number of features. For this reason, classical methods, such as using the inverse of the sample covariance matrix, are known to perform poorly whenever the number of observation is not extremely large. Additionally, they produce estimates which are almost surely dense. This makes regularization crucial, since in many applications we typically only have a moderate number of observations, and in particular, we are most often seeking a sparse estimate of which gives rise to a more parsimonious and interpretable network model.

[1]Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden [2]Lynx Asset Management AB, Stockholm, Sweden. Correspondence to: Samuel Erickson <samuelea@kth.se>.

2. Background

Suppose Z is a square-integrable mean-zero random vector taking values in $\mathbf{R}^p$, with non-singular covariance Σ, and inverse $\Omega = \Sigma^{-1}$. Then the best linear unbiased predictor $Z_{\setminus j}^\top \theta_j$ of a feature Z_j given all other features

$$Z_{\setminus j} = (Z_1, \ldots, Z_{j-1}, Z_{j+1}, \ldots, Z_p)$$

can be characterized via the precision matrix by

$$\theta_j = -\Omega_{\setminus j,j}/\Omega_{jj}.$$

Here $\Omega_{\setminus j,j}$ denotes the jth column of Ω with the jth elemented omitted. Moreover, the variance of the residual $\varepsilon_j = Z_j - Z_{\setminus j}^\top \theta_j$ is given by $\mathbf{Var}(\varepsilon_j) = 1/\Omega_{jj}$. For this reason, if Z follows a Gaussian distribution Z_j and Z_k are conditionally independent given the remaining features if and only if $\theta_{jk} = 0$, or equivalently, $\Omega_{jk} = 0$. This connection is the foundation of *Gaussian graphical models*.

For quantifying the correlation between two features Z_j and Z_k given the remaining features, we can define the *partial correlation* between them as the negative correlation between ε_j and ε_k,

$$\rho_{jk|\setminus\{j,k\}} = -\mathbf{Corr}(\varepsilon_j, \varepsilon_k) = -\frac{\mathbb{E}(\varepsilon_j \varepsilon_k)}{\sqrt{\mathbb{E}(\varepsilon_j^2)\mathbb{E}(\varepsilon_k^2)}},$$

which can be written in terms of the precision matrix and the linear regression as

$$\rho_{jk|\setminus\{j,k\}} = -\frac{\Omega_{jk}}{\sqrt{\Omega_{jj}\Omega_{kk}}} = \frac{\tau_k}{\tau_j}\theta_{jk},$$

respectively. Thus we define the *partial correlation matrix* Q as

$$Q = -T\Omega T,$$

where $T = \mathbf{diag}(\tau_1, \ldots, \tau_p)$. We call the weighted network that Q defines the *partial correlation network*.

3. Proposed method

Based on the connection described in §2, we propose the *joint partial regression method*, which is a two-step estimation method for simultaneous estimation of the precision and partial correlation matrices, and is described in Algorithm 1.

Algorithm 1 JOINT PARTIAL REGRESSION

Input: Data matrix $X \in \mathbf{R}^{n\times p}$, penalty parameter λ.

for $j = 1, \ldots, p$ **do**

lasso regression

$$\hat{\theta}_j = \underset{\theta\in\mathbf{R}^{p-1}}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|X_j - X_{\setminus j}\theta\|_2^2 + \lambda\|\theta\|_1 \right\},$$

and compute the estimate $\hat{\tau}_j^2$ of the residual variance.

end

Solve the convex program

$$\begin{array}{ll} \textit{minimize} & \displaystyle\sum_{j=1}^{p}\left(\frac{1}{2n}\|X_j - X_{\setminus j}\theta_j\|_2^2 + \lambda\|\theta_j\|_1\right) \\ \textit{subject to} & \Omega_{jj} = 1/\hat{\tau}_j^2, \quad \Omega_{\setminus j,j} = -\theta_j/\hat{\tau}_j^2, \\ & \Omega \succeq 0, \quad Q = -\widehat{T}\Omega\widehat{T} \end{array}$$

with the estimated residual variances $\hat{\tau}_j^2$ and regularization parameter λ to obtain the estimates $\widehat{\Omega}$ and $\widehat{Q}$ of the precision matrix and partial correlation matrix, respectively.

4. Theoretical results

We establish statistical estimation error rates under the following assumptions.

Assumption 4.1. The dimensionality is such that $p/n \leq 1-\delta$ for some $\delta \in (0,1)$, and the degree

$$d = \max_j \sum_{k\neq j} \mathbf{1}(\Omega^\star_{jk} \neq 0)$$

of the partial correlation network is such that $d\sqrt{\log(p)/n} \leq M$ for some constant $M > 0$.

Assumption 4.2. The rows of the design matrix $X \in \mathbf{R}^{n\times p}$ are n i.i.d. samples from a random vector with covariance matrix $\Sigma^\star$, and each X_{ij} is sub-Gaussian with associated norm $\|X_{ij}\|_{\psi_2} \leq K$ for some $K > 0$.

Assumption 4.3. There exists constants $\kappa \in (1,\infty)$ and $L \in (0,\infty)$ such that the precision matrix $\Omega^\star = (\Sigma^\star)^{-1}$ satisfies

$$1/\kappa \leq \lambda_{\min}(\Omega^\star) \leq \lambda_{\max}(\Omega^\star) \leq \kappa, \quad \text{and} \quad \|\Omega^\star\|_{\ell_1} \leq L.$$

Defining $s = \mathbf{card}\{(j,k)\colon \Omega^\star_{jk} \neq 0,\ j \neq k\}$ as the size of the partial correlation network, we can now state the main result of this section using the assumptions above.

Theorem 4.4. *Under Assumptions 4.1–4.3, there exist positive constants c, C_1 and C_2 such that Algorithm 1 with $\lambda = c\sqrt{\log(p)/n}$ outputs an estimate $\widehat{\Omega}$ of the precision matrix that satisfies*

$$\|\widehat{\Omega} - \Omega^\star\|_{\mathrm{F}} \leq C_1\sqrt{\frac{(s+p)\log p}{n}} \tag{1}$$

and an estimate $\widehat{Q}$ of the partial correlation matrix that satisfies

$$\|\widehat{Q} - Q^\star\|_{\mathrm{F}} \leq C_2\sqrt{\frac{s\log p}{n}} \tag{2}$$

with probability at least $1 - 6/p$.

The statistical rate of convergence (1) matches the state of the art for precision matrix estimation (Rothman et al., 2008), whereas (2) improves upon the state of the art for partial correlation estimation (Peng et al., 2009).

5. Numerical experiments

We evaluate the performance of the joint partial regression method on synthetic data for which the distribution and the true precision matrix are known. In Figure 1, the performance of the proposed method (blue) is compared with the graphical lasso (red), as well as the ideal oracle estimator (green). We compare the performances on different classes of precision matrices.

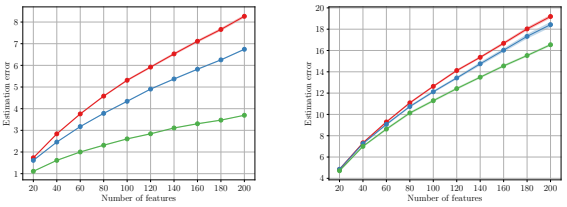

Figure 1. Average Frobenius error versus number of features with $\pm 2 \times$ SE bands for AR(1) model (left) and Hub network model (right).

References

de la Fuente, A., Bing, N., Pedro, I. H., and Mendes. Discovery of meaningful associations in genomic data

using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 07 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth445. URL https://doi.org/10.1093/bioinformatics/bth445.

Epskamp, S. and Fried, E. I. A tutorial on regularized partial correlation networks. *Psychological Methods*, 23 (4):617–634, 2018. doi: 10.1037/met0000167. URL https://doi.org/10.1037/met0000167.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2nd edition, 2009.

Markowitz, H. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 00221082, 15406261.

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. doi: 10.1198/jasa.2009.0126. URL https://doi.org/10.1198/jasa.2009.0126.

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