Sparse Permutation Invariant Covariance Estimation: Motivation, Background and Key Results

David Prince

Biostat 572

dprince3@uw.edu

April 19, 2012

Electronic Journal of Statistics
Vol. 2 (2008) 494–515
ISSN: 1935-7524
DOI: 10.1214/08-EJS176

Sparse permutation invariant covariance estimation

Adam J. Rothman

University of Michigan
Ann Arbor, MI 48109-1107
e-mail: ajrothma@umich.edu

Peter J. Bickel

University of California
Berkeley, CA 94720-3860
e-mail: bickel@stat.berkeley.edu

Elizaveta Levina*

University of Michigan
Ann Arbor, MI 48109-1107
e-mail: elevina@umich.edu

Ji Zhu

University of Michigan
Ann Arbor, MI 48109-1107
e-mail: jizhu@umich.edu

Motivation

- Reductionism vs. Dynamism
- Gene-gene interaction networks
- High-dimensional setting, i.e. $n << p$
- Two key questions:
  1. Which gene products are directly dependent (yes/no for each pair of genes)?
  2. What is the strength and direction of this dependence (numeric for each pair of genes)?
- Roadmap for lab research

The Gaussian graphical models approach

- Multivariate normality assumption (with standardization)

$$\vec{X} \sim N_p(0, \Sigma)$$

- For the ith and jth elements of $\vec{X}$, we know:

$$\Sigma_{i,j} = cov(X_i, X_j)$$

By normality, a zero here implies indepedence, but what does a non-zero imply?

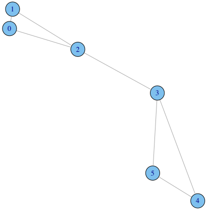
- Consider then $\Omega = \Sigma^{-1}$. A result from Lauritzen (1996):

$$\Omega_{i,j} = 0 \Leftrightarrow X_i \text{ and } X_j \text{ are conditionally independent}$$

A non-zero here implies dependence conditional on all other variables

The Gaussian graphical models approach

- The inverse covariance matrix (Question 2) can be thought of as implying a graphical model (Question 1)
- Hence the graph reflects conditional dependence



$$\Sigma^{-1} = \begin{pmatrix} 1 & .8 & .8 & 0 & 0 & 0 \\ .8 & 1 & .8 & 0 & 0 & 0 \\ .8 & .8 & 1 & .1 & 0 & 0 \\ 0 & 0 & .1 & 1 & .8 & .8 \\ 0 & 0 & 0 & .8 & 1 & .8 \\ 0 & 0 & 0 & .8 & .8 & 1 \end{pmatrix}$$

Previous attempts to solve similar problems

Answering question 1:

- Multiple testing procedures (Drton and Perlman, 2008)
- Multiple regressions (Meinshausen and Buehlmann, 2006)
- Graphical Lasso aka GLASSO (Friedman et al., 2007)
- Thresholding the sample covariance matrix

Answering question 2 (and therefore also 1):

- Sample covariance, inconsistent in high-dimensional setting (Johnstone, 2001)
- Shrink eigenvalues of covariance matrix (Ledoit and Wolf, 2003)
- Add structure to the covariance matrix, e.g. banding - think AR-1. (Bickel and Levina, 2008).
- ℓ_q penalized likelihood (d'Aspremont et al. and Yuan and Lin)

The sparse permutation invariant covariance estimator aka SPICE

$$\hat{\Omega}_\lambda = \arg\min_{\Omega \succ 0} \{ tr(\Omega\hat{\Sigma}) - \log det(\Omega) + \lambda|\Omega^-|_1 \}$$

where :

- $\Omega = \Sigma^{-1}$
- $\hat{\Sigma} = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$
- $\Omega^- = \Omega - diagonal(\Omega)$
- λ is the tuning parameter

Key results 1: Theoretical properties of SPICE

- Assumptions:
  1. the maximum and minimum of the eigenvalues of the true covariance matrix are bounded
  2. the number of non-zero inverse covariance elements are bounded (sparsity)
  3. $\frac{\lambda}{\sqrt{(\frac{\log p}{n})}} \rightarrow c$, where c is a constant
- Theorem 1:

$$\|\hat{\Omega}_\lambda - \Omega_0\|_F = O_P(\sqrt{\frac{(p+s)\log p}{n}})$$

- Theorem 2:

$$\|\Omega_\lambda - \Omega_0\| = O_P(\sqrt{\frac{(s+1)\log p}{n}})$$

Key results 2: The algorithm

- The algorithm uses the Cholesky decomposition to solve the minimization problem, i.e. that positive definite matrices can be written as:

  $$\hat{\Omega}_\lambda = T^T T$$

  where T is a lower triangular matrix
- The three terms from the SPICE estimator can then be rewritten as $f(T)$
- A quadratic approximation of the penalty term is used and each element of T is minimized one at a time (cyclical coordinate descent)
- Convergence because of cyclical coordinate descent and smooth functions (Bazaraa, 2006)

Key results 3: Applications

Simulation study

- Uses four different inverse covariance matrix generating mechanisms (all sparse)
- Compares sample inverse covariance matrix and Ledoit-Wolf estimator using Kullback-Leibler loss and true positive and true negative rates

Colon tumor classification example

- Compares classification error rates
- Uses Naive Bayes, Ledoit-Wolf and SPICE with three ways of deriving the tuning parameter

Next Steps

This quarter

- Prove key results 1, 2, 3 (potentially starting with 3)
- Linear algebra review (and new learning)
- Impact of standardization of covariance or inverse covariance matrix

In the future

- Nearly all approaches make the MVN assumption, if this does not hold what do the results mean if we use this or similar methods?
- Contrast the algorithm in this paper with the GLASSO algorithm
- Applied papers review
- Penalization requires training set to determine tuning parameter value