Sparse Maximum Likelihood Estimation for Gaussian and Binary Data

David Lee

Graphical Models Reading Group

August 14, 2015

Sparse graphical model

- The presentation is based on the article by Banerjee et al. (2008)[1].
- This was published at almost the same time as the article on graphical lasso by Friedman et al. (2008)[2]. Both articles are based on essentially the same optimization problem.
- Banerjee et al. (2008) proposed a different algorithm for optimization and showed how it can also be used for multivariate binary data.

[1] Banerjee, O., El Ghaoui, L., and d'Aspremont A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9, 485-516.

[2] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 432-441.

Problem specification

- We assume the p-variate random variables $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \Sigma)$ with sample covariance matrix S. The objective is to obtain a parsimonious (sparse) precision matrix $\hat{\Sigma}^{-1}$.
- In terms of graphical representation, a zero at (i, j) position in the precision matrix implies the absence of an edge between variables i and j in the corresponding concentration (precision) graph.
- To obtain the estimate, the penalized log-likelihood with respect to the parameter Σ^{-1} is optimized:

$$\hat{\Sigma}^{-1} = \underset{X \succ 0}{\arg\max} \left(\log |X| - \operatorname{tr}(SX) - \lambda ||X||_1 \right) \tag{1}$$

where λ is the regularization parameter and $||X||_1$ is the sum of the absolute values of the elements of X.

The dual problem

- Rather than solving (1), the authors developed an algorithm that solves its corresponding dual problem.
- We can write $||X||_1 = \max_{||U||_\infty \leq 1} \text{tr}(XU)$, where $||U||_\infty$ is the maximum absolute value of the elements of U. Hence

$$\begin{aligned} & \max_{X \succ 0} \left(\log |X| - \text{tr}\left(SX\right) - \lambda ||X||_1 \right) \\ = \; & \max_{X \succ 0} \min_{||U||_\infty \leq \lambda} \log |X| - \text{tr}\left[X\left(S + U\right)\right]. \end{aligned}$$

The interpretation is to find an estimate X with the worst-case log-likelihood over all perturbations U (not exceeding λ) on S.

The dual problem (cont.)

- The dual problem is obtained by exchanging the max and min operators. The resulting inner optimization in X has a closed-form solution, and the overall outer optimization becomes

  $$\min_{||U||_\infty \leq \lambda} -\log|S+U| - p$$

  with the primal variable $X = (S+U)^{-1}$. Hence the dual is

  $$\hat{\Sigma} = \arg\max_{W \succ 0} \left\{ \log|W| : ||W-S||_\infty \leq \lambda \right\}, \tag{2}$$

  i.e. maximize the determinant of W subject to the constraint that W is “close enough” to S.
- The dual problem estimates the covariance matrix while the primal problem deals with the precision matrix.

Choice of the penalty parameter

- Consider the penalty parameter as a function of α, such that

$$\lambda(\alpha) = \left(\max_{i>j} s_i^2 s_j^2\right) \frac{t_{n-2}\left(\alpha/2p^2\right)}{\sqrt{n-2+t_{n-2}^2\left(\alpha/2p^2\right)}},$$

  where s_k^2 is the sample variance of variable k and $t_{n-2}(x)$ is the $(100-x)\%$ quantile of the t distribution with $n-2$ degrees of freedom.

- With this choice of $\lambda(\alpha)$, the authors showed that

$$\mathbb{P}\left(\exists k \in \{1,\ldots,p\} : \hat{C}_k^\lambda \not\subseteq C_k\right) \leq \alpha,$$

  where C_k ($\hat{C}_k^\lambda$) is the true (fitted) set of nodes connected to variable k, directly or indirectly.

- In English: The probability that at least one node has a false connection in the fitted model is at most α.

Block coordinate descent algorithm

- For a symmetric matrix A, let A_{-kj} be the submatrix by removing column k and row j, and A_j be the column j without the diagonal element A_{jj}.
- The algorithm optimizes over one row and column of the matrix W at a time, sweeps through all columns until convergence.
- Steps:
  1. Initialize $W^{(0)} = S + \lambda I$, I being the identity matrix.
  2. For $j = 1, \ldots, p$:
     1. Let $W^{(j-1)}$ be the current iterate. Solve the quadratic program (with complexity $O(p^3)$)
        $$\hat{y} = \arg\min_y \left\{ y' \left(W_{-jj}^{(j-1)} \right)^{-1} y : ||y - S_j||_\infty \leq \lambda \right\}. \qquad (3)$$
     2. Update the jth column/row of $W^{(j-1)}$ by $\hat{y}$. This is our $W^{(j)}$.

Block coordinate descent algorithm (cont.)

3. When all columns are optimized, check the convergence condition. Convergence occurs when

$$\operatorname{tr}\left[\left(W^{(p)}\right)^{-1} S\right] - p + \lambda \left\| \left(W^{(p)}\right)^{-1} \right\|_1 \leq \epsilon$$

for some predetermined tolerance ϵ.

- One result for the solution is that, for a given $k \in \{1, \ldots, p\}$, if $\lambda \geq |S_{kj}|$ for all $j \neq k$, then column and row k of $\hat{\Sigma}$ in (2) are all zeroes except the diagonal element. Hence, if $\lambda \geq |S_{k,j}|$ for all $k > j$, the estimated covariance matrix is diagonal (i.e. all elements are independent).
- The computational cost of this algorithm is $O(Kp^4)$ where K is the number of sweeps.

Iterative penalized regression

- Let $Q = \left(W_{-jj}^{(j-1)} \right)^{1/2}$ and $b = Q^{-1} S_j / 2$, then the dual of the quadratic program (3) can be written as

$$\arg\min_{x} ||Qx - b||_2^2 + \lambda ||x||_1$$

  which is the lasso.
- Friedman et al. (2008) utilized this connection in developing the graphical lasso method.

Multivariate binary variables

- Consider a p-variate binary random variable. The objective is again to estimate the structure of the distribution. An Ising model (/ˈaɪsɪŋ/; German: [ˈiːzɪŋ]) admits the following density:

$$p(x;\theta) = \exp\left\{\sum_{i=1}^{p}\theta_i x_i + \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\theta_{ij}x_i x_j - A(\theta)\right\}$$

  where θ_i are the marginal parameters, θ_{ij} are the interaction parameters and $A(\theta)$ is a normalizing constant that often has too many terms to calculate.

Multivariate binary variables (cont.)

- The conditional distribution of node i given the rest can be calculated as

$$\mathbb{P}(X_i|X_{-i}) \propto \exp\left\{\theta_i x_i + \sum_{j\neq i} \theta_{ij} x_i x_j\right\},$$

  Where $\theta_{ij} := \theta_{ji}$ if $i > j$. Therefore, if $\theta_{ij} = 0$, nodes i and j are conditionally independent given the rest.

Multivariate binary variables (cont.)

- By rewriting the upper bound of $A(\theta)$ established in Wainwright and Jordan (2006)[3], the authors obtained the estimators of θ_i's and θ_{ij}'s for the penalized MLE:

$$\begin{aligned} \hat{\theta}_i &= \bar{x}_i; \\ \hat{\theta}_{ij} &= -\left(\hat{\Gamma}^{-1}\right)_{ij}, \end{aligned}$$

where $\bar{x}_i$ is the sample mean for component i, and

$$\hat{\Gamma} = \arg\max_{W} \left\{ \log|W| : W_{kk} = S_{kk} + \frac{1}{3}, |W_{kj} - S_{kj}| \leq \lambda \right\}$$

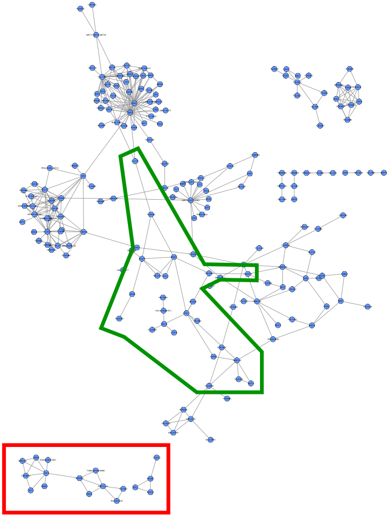
and S is the sample covariance matrix. This resembles the Gaussian case and can be optimized using the algorithms already developed.

[3] Wainwright, M. J., & Jordan, M. (2006). Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Transactions on Signal Processing*, 54, 2099-2109.

Example 1: Gene expression profile

- There are $n = 253$ observations with $p = 6136$ variables (genes), assumed to follow the Gaussian distribution. The penalty parameter used is $\lambda = 0.0313$.
- Dependence is only found among 270 genes.
- The fitted graph is shown on the next slide; the region highlighted in red contains genes associated with iron homeostasis, while those contained in the green polygon are genes associated with cellular membrane fusion.

Example 1: Gene expression profile (cont.)



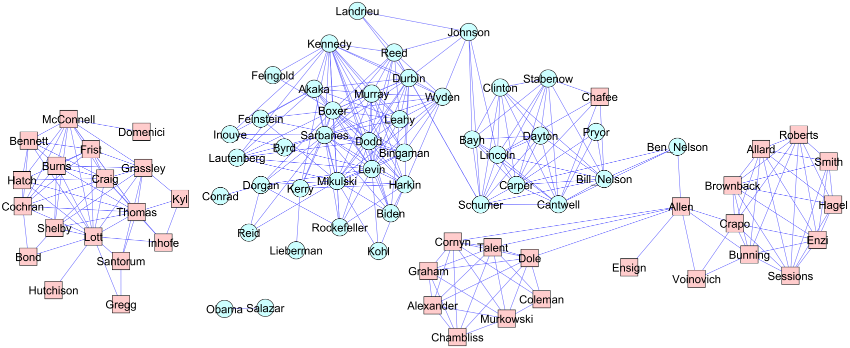
Example 2: Gene microarray data

- A subset of 500 out of 10,000 genes with highest variance is selected for analysis. There are $n = 160$ observations.
- The solution shows "all but 339" genes are independent from the rest.
- For the resulting graph, the first order neighbours of a node form the set of predictors for that variable. It was found that the LDL receptor had one of the largest number of first-order neighbours in the Gaussian graphical model.
- Some of those neighbours are directly involved in either lipid or steroid metabolism, while some are known to be global transcriptional regulators.

Example 3: Senate voting records data

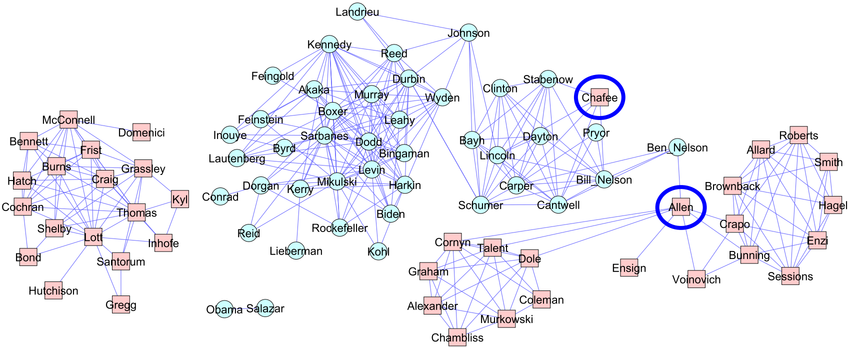
- A binary data set containing the $n = 542$ bills voted among the $p = 100$ senators in the US Congress between 2004 and 2006 is examined. Each observation is either a *no* (coded -1) or a *yes* (coded 1). Missing votes are counted as noes.
- The association among senators are shown in the graph on the next slide. Red and cyan nodes correspond to Republican and Democratic senators respectively.

Example 3: Senate voting records data (cont.)



- Observation 1: Most senators have only members from the same party as their neighbours.

Example 3: Senate voting records data (cont.)



- Observation 2: Chafee (R) has only Democrats as his neighbours, while Allen (R) unites two otherwise separate groups of Republicans and connects to the large cluster of Democrats through Ben Nelson (D). These are consistent with media statements.

The End

Thank you!