An Introduction to Graphical Lasso

Bo Chang

Graphical Models Reading Group

May 15, 2015

Undirected Graphical Models

- An undirected graph, each vertex represents a random variable.
- The absence of an edge between two vertices means the corresponding random variables are conditionally independent, given other variables.
- The Gaussian distribution is widely used for such graphical models, because of its convenient analytical properties.
- Penalized regression methods for inducing sparsity in the precision matrix are central to the construction of Gaussian graphical models.

Precision Matrix

Denote the covariance matrix by $\mathbf{\Sigma}$, then the inverse covariance matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ is called precision matrix. Let θ_{ij} be the (i, j)th element of $\mathbf{\Theta}$.

$$\theta_{ij} = -\sigma_{ij;\mathrm{rest}} \det(\mathbf{\Sigma}^{(ij)}) \det(\mathbf{\Sigma})^{-1}.$$

- $\sigma_{ij;\mathrm{rest}}$: conditional/partial covariance of variables i and j, given the other variables.
- $\mathbf{\Sigma}^{(ij)}$: matrix $\mathbf{\Sigma}$ with ith row and jth column removed.
- If $\theta_{ij} = 0$, then variables i and j are conditionally independent, given other variables.

Precision Matrix

- Suppose we partition $X^T = (X_1^T, X_2)$, where X_1 consists of the first $d-1$ variables and X_2 is the last.
- We have the partition of $\mathbf{\Sigma}$ and $\mathbf{\Theta}$:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \sigma_{12} \\ \sigma_{12}^T & \sigma_{22} \end{pmatrix}, \qquad \mathbf{\Theta} = \begin{pmatrix} \mathbf{\Theta}_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix}.$$

- Let $\beta = \mathbf{\Sigma}_{11}^{-1}\sigma_{12}$ be the multiple linear regression coefficient of X_2 on X_1.
- Since $\mathbf{\Sigma\Theta} = \mathbf{I}$,

$$\mathbf{\Sigma}_{11}\theta_{12} + \sigma_{12}\theta_{22} = 0,$$

$$\beta = \mathbf{\Sigma}_{11}^{-1}\sigma_{12} = -\theta_{12}/\theta_{22}.$$

Precision Matrix

- Regression coefficient:

$$\beta = -\theta_{12}/\theta_{22}.$$

- We can learn about the dependence structure through multiple linear regression.
- Meinshausen and Bhlmann (2006) try to estimate which components θ_{ij} are zero, rather than fully estimate $\mathbf{\Theta}$. They fit a lasso regression using each variable as the response and the others as predictors.

Lasso

- Minimize

$$Q(\beta) = \frac{1}{2}\|Y - X\beta\|^2 + \lambda \sum_j |\beta_j|.$$

- When $n = p = 1$ and $X = 1$,

$$Q(\beta) = \frac{1}{2}(y - \beta)^2 + \lambda|\beta|.$$

$$Q'(\beta) = -y + \beta + \lambda \cdot \mathrm{sign}(\beta) = 0.$$

- Lasso solution

$$\hat{\beta}(\lambda) = \mathrm{sign}(y)(|y| - \lambda)_+ = S(y, \lambda),$$

where $S(y, \lambda)$ is called the soft-thresholding operator.

Graphical Lasso

A more systematic approach by Friedman, Hastie and Tibshirani (2008).

- Consider maximizing the penalized log-likelihood

$$\log(\det[\mathbf{\Theta}]) - \mathrm{trace}(\mathbf{S\Theta}) - \lambda\|\mathbf{\Theta}\|_1.$$

  $\mathbf{S}$: sample covariance matrix.
  $\|\mathbf{\Theta}\|_1$: element L_1 norm, the sum of the absolute values of the elements of $\mathbf{\Theta}$.

- The gradient equation

$$\mathbf{\Theta}^{-1} - \mathbf{S} - \lambda \cdot \mathrm{Sign}(\mathbf{\Theta}) = \mathbf{0}.$$

- The gradient equation

$$\mathbf{\Theta}^{-1} - \mathbf{S} - \lambda \cdot \mathrm{Sign}(\mathbf{\Theta}) = \mathbf{0}.$$

- Let $\mathbf{W} = \mathbf{\Theta}^{-1}$ and

$$\begin{pmatrix} \mathbf{W}_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} \begin{pmatrix} \mathbf{\Theta}_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0^T & 1 \end{pmatrix}.$$

$$w_{12} = -\mathbf{W}_{11}\theta_{12}/\theta_{22} = \mathbf{W}_{11}\beta,$$

  where $\beta = -\theta_{12}/\theta_{22}$.

- The upper right block of the gradient equation:

$$\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \mathrm{Sign}(\beta) = 0$$

  which is recognized as the estimation equation for the Lasso regression.

Algorithm 17.2 *Graphical Lasso.*

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains unchanged in what follows.

2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:

   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.

   (b) Solve the estimating equations $\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \text{Sign}(\beta) = 0$ using the cyclical coordinate-descent algorithm (17.26) for the modified lasso.

   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$

3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.

Graphical Lasso

- Coordinate descent: Let $\mathbf{V} = \mathbf{W}_{11}$,

$$\hat{\beta}_i \leftarrow S(s_{12i} - \sum_{k \neq j} V_{ki} \hat{\beta}_k, \lambda) / V_{ii},$$

where $S(y, \lambda)$ is the soft-thresholding operator.

Analysis of Protein-signalling Data

We analyze a flow cytometry dataset on $d = 11$ proteins and $n = 7466$ cells. Several methods are compared:

- Graphical Lasso
- Bayesian Network
- Truncated Vine (Sequential MST)
- Factor Analysis

Discrepancy Measure

- A common discrepancy measure in the psychometrics and structural equation modeling literatures is:

$$D = \log(\det[\mathbf{R}_{\mathrm{model}}(\hat{\delta})]) - \log(\det[\mathbf{R}_{\mathrm{data}}]) + \mathrm{tr}[\mathbf{R}^{-1}_{\mathrm{model}}(\hat{\delta})\mathbf{R}_{\mathrm{data}}] - d.$$

  d: number of variables.
  $\mathbf{R}_{\mathrm{data}}$: sample correlation matrix.
  $\mathbf{R}_{\mathrm{model}}(\hat{\delta})$: model-based correlation matrix based on the estimate of the parameter δ. If either model has some conditional independence relations, then the dimension of δ is less than $d(d-1)/2$.

Discrepancy Measure

- Other comparisons are the AIC/BIC based on a Gaussian log-likelihood.
- Also useful are the average and max absolute deviations of the model-based correlation matrix from the empirical correlation matrix:

$$\max_{j<k} |\mathbf{R}_{\mathrm{data},jk} - \mathbf{R}_{\mathrm{model},jk}(\hat{\delta})|.$$

Results

Model	Dfit	MaxAbsDiff	AIC($\times 10^5$)	BIC($\times 10^5$)	#Par
BN	0.013	0.019	1.969	1.972	36
glasso($\lambda = 0.13$)	1.232	0.200	2.060	2.062	33
glasso($\lambda = 0.10$)	0.930	0.159	2.038	2.040	37
glasso($\lambda = 0.08$)	0.700	0.126	2.020	2.023	41
1-truncated seq. MST	1.030	0.306	2.044	2.045	10
2-truncated seq. MST	0.568	0.242	2.010	2.012	19
3-truncated seq. MST	0.328	0.197	1.992	1.994	27
4-truncated seq. MST	0.224	0.229	1.985	1.987	34
5-truncated seq. MST	0.142	0.150	1.979	1.982	40
1-factor	2.682	0.571	2.168	2.169	11
2-factor	1.689	0.529	2.094	2.095	21
3-factor	0.832	0.456	2.030	2.032	30
4-factor	0.245	0.119	1.986	1.989	38

References

Hastie, T., Tibshirani, R., Friedman, J. (2009).
The elements of statistical learning.
New York: springer.

Pourahmadi, M. (2013).
High-Dimensional Covariance Estimation: With High-Dimensional Data.
John Wiley & Sons.

Meinshausen, N., & Bhlmann, P. (2006).
High-dimensional graphs and variable selection with the lasso.
The Annals of Statistics, 1436-1462.

Friedman, J., Hastie, T., & Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3), 432-441.Chicago

The End