Sparse inverse covariance estimation using Graphical Lasso and BIGQUIC

STAT 557- Survey

Amal Agarwal

PENNSTATE

Department of Statistics
Pennsylvania State University

Graphical models

- $G = (V, E)$ formed by a collection of vertices $V = \{1, 2, ..., p\}$, and a collection of edges $E \subset V \times V$.
- Each edge consists of a pair of vertices $(s, t) \in E$.
- Undirected graph models: No distinction between edge (s, t) and edge (t, s).
- Directed graph models or Bayesian Networks: $(s \rightarrow t)$ to indicate the direction.

The problem of sparse concentration estimation

- GMRF: $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n \sim \mathcal{N}(\underset{\sim}{\mu}, \Sigma)$
- Goal: To estimate Σ or $C = \Sigma^{-1}$.
  - **Model Selection**: Identification of zero entries in C $\Leftrightarrow$ Finding the missing edges in the undirected graph $\Leftrightarrow$ Conditional independencies.
  - **Model estimation**: Estimation of the non-zero entries in C.
- Log likelihood[1]:

$$l(\underset{\sim}{\mu}, C) = \frac{n}{2} log|C| - \frac{1}{2}\sum_{i=1}^{n}(\underset{\sim}{X}_i - \underset{\sim}{\mu})^T C(\underset{\sim}{X}_i - \underset{\sim}{\mu})$$

- MLE of $(\underset{\sim}{\mu}, \Sigma)$ is $(\overline{X}, \overline{A})$, where

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{X}_i - \underset{\sim}{\overline{X}})(\underset{\sim}{X}_i - \underset{\sim}{\overline{X}})^T$$

[1] Ming Yuan & Yi Lin,"Model Selection and estimation in gaussian graphical model", Biometrika,94(1):19-35,2007

The problem of sparse concentration estimation

- L_1 Lasso penalty term in the log likelihood function by putting the following constraint.

$$\sum_{i \neq j} |c_{ij}| \leq t$$

- For centered data, 2nd term in the log likelihood function transforms as:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{X_i}^T C \underset{\sim}{X_i} &= \frac{1}{n}\sum_{i=1}^{n} tr(\underset{\sim}{X_i}^T C \underset{\sim}{X_i}) \\
&= \frac{1}{n}\sum_{i=1}^{n} tr(C \underset{\sim}{X_i} \underset{\sim}{X_i}^T) \\
&= tr\left(C\left[\frac{1}{n}\sum_{i=1}^{n} \underset{\sim}{X_i} \underset{\sim}{X_i}^T\right]\right) \\
&= tr(C\bar{A})
\end{aligned}$$

The problem of sparse concentration estimation

- Thus our optimization problem becomes: Minimize

$$l(C) = -\log|C| + tr(C\bar{A}) \quad \text{subject to}$$

$$\sum_{i \neq j} |c_{ij}| \leq t$$

- Objective function and the feasible region are convex; Minimize

$$-\log|C| + tr(C\bar{A}) + \lambda \sum_{i \neq j} |c_{ij}|$$

where λ is the tuning parameter.

Preliminaries for Graphical Lasso

- Let W be the estimate of Σ. Partitioning W and S as

$$\begin{bmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{bmatrix} \qquad \begin{bmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{bmatrix}$$

- It can be shown that the solution for w_{12} satisfies

$$w_{12} = \text{argmin}_y \left\{ y^T W_{11}^{-1} y : ||y - s_{12}||_\infty \leq \lambda \right\} \qquad (1)$$

- Using covex duality, it can be shown that this can be solved by solving the corresponding dual problem as

$$\min_\beta \left\{ \frac{1}{2} ||W_{11}^{1/2}\beta - b||^2 + \rho||\beta||_1 \right\} \qquad (2)$$

where $b = W_{11}^{-1/2} s_{12}$.

Graphical Lasso algorithm

- Start with $W = S + \lambda I$.[2] The diagonal of W remains unchanged in what follows.
- For each $j = 1, 2, ...p$, solve the lasso problem (1), which takes as input the inner products W_{11} and s_{12}. This gives a $p - 1$ vector solution $\hat{\beta}$.
- Fill in the corresponding row and column of W using $w_{12} = W_{11}\hat{\beta}$.
- Continue until convergence.

[2] J.Friedman, T. Hastie and R.Tibshirani, "Sparse inverse covariance estimation with the graphical lasso", Biostatistics, vol. 9, no. 3, pp. 432-441, 2008

BIGQUIC

- State of the art techniques do not scale to problems for more than $20,000$ variables.
- The BIGQUIC algorithm[3] can solve 1 million dimensional L_1 regularized Gaussian MLE problems (with 1000 billion parameters) using a single machine with bounded memory.
- Key features:
  - Block coordinate descent with blocks chosen via a clustering scheme to minimize repeated computations.
  - Allows inexact computations for specific components.
  - Super linear or quadratic convergence rates.

[3] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep K Ravikumar and Russell Poldrack, "Big & quic: Sparse inverse covariance estimation for a million variables", Advances in Neural Information Processing Systems, pp. 3165-3173, 2013