Paper Review

Minzhe Zhang mxz 163730

Sparse inverse covariance estimation with the graphical lasso

The graphical lasso (glasso) is an algorithm for learning the structure in an undirected Gaussian graphical model, using ℓ_1 regularization to control the number of zeros in the precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$. Given a data matrix $\boldsymbol{X}_{n\times p}$, a sample of n observations from a p-dimensional Gaussian distribution with zero mean and positive definite covariance matrix $\mathbf{\Sigma}$. The task is to estimate the unknown $\mathbf{\Sigma}$ based on the n samples. This is a challenging problem especially when n $\ll$ p which leads to the nonexistence of or poorly behaved ordinary maximum likelihood estimate, therefore regularization is called for. Friedman et al. proposed the Graphical Lasso regularization framework for estimating the covariance matrix $\mathbf{\Sigma}$, under the assumption that its inverse $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ is sparse. $\mathbf{\Theta}$ is the precision matrix; element $\theta_{jk} = 0$ implies the corresponding variables X_j and X_k are conditionally independent, given the rest. The goal of graphical lasso is to maximize a ℓ_1-regularized negative log-likelihood:

$$l(\mathbf{\Theta}) = \log|\mathbf{\Theta}| - \mathrm{tr}(\boldsymbol{S}\mathbf{\Theta}) - \lambda\|\mathbf{\Theta}\|_1 \tag{1}$$

where $\boldsymbol{S}$ is the sample covariance matrix, tr($\mathbf{A}$) is the trace of $\mathbf{A}$ and $\|\mathbf{A}\|_1$ is the ℓ_1 norm of $\mathbf{A}$ for $\mathbf{A}\in\mathbb{R}_{p\times p}$.

To be specific, let W be the estimate of the covariance matrix Σ and consider partitioning W and S

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}_{11} & \boldsymbol{w}_{12} \\ \boldsymbol{w}_{12}^T & \boldsymbol{w}_{22} \end{pmatrix},\ \boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{s}_{12} \\ \boldsymbol{s}_{12}^T & \boldsymbol{s}_{22} \end{pmatrix},\ \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{12}^T & \boldsymbol{\omega}_{22} \end{pmatrix}$$

Motivated by Banerjee et al., Friedman et al. show that the solution $\widehat{\boldsymbol{\Omega}}$ of (1) is equivalent to the inverse of $\boldsymbol{W}$ whose partitioned entity $\boldsymbol{w}_{12}$ satisfies $\boldsymbol{w}_{12} = \boldsymbol{W}_{11}\beta^*$, where β^* is the solution of the lasso problem

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \| \boldsymbol{W}_{11}^{1/2}\boldsymbol{\beta} - \boldsymbol{W}_{11}^{-1/2}\boldsymbol{s}_{12} \|_2^2 + \boldsymbol{\lambda} \| \boldsymbol{\beta} \|_1 \tag{2}$$

Based on the above property, the graphical lasso sets the diagonal elements $\boldsymbol{w}_{ii} = \boldsymbol{s}_{ii} + \rho$ and obtains the off-diagonal elements of W by repeatedly applying the following two steps:

1. Permuting the columns and rows to locate the target elements at the position of $\boldsymbol{w}_{12}$.
2. Finding the solution $\boldsymbol{w}_{12} = \boldsymbol{W}_{11}\beta^*$ by solving the lasso problem (2).

until convergence occurs. After finding $\boldsymbol{W}$, the estimate $\widehat{\boldsymbol{\Omega}}$ is obtained from the relationships $\omega_{12} = -\hat{\beta}\widehat{\omega}_{22}$ and $\widehat{\omega}_{22} = 1/(w_{22} - w_{12}^T\hat{\beta})$, where $\hat{\beta} = W_{11}^{-1}w_{12}$.

However, as noted earlier, glasso operates on $\boldsymbol{W}$, it does not explicitly compute the inverse $\boldsymbol{W}^{-1}$. It does however keep track of the estimates for θ12 after every row/column update. The copy of $\boldsymbol{\Theta}$ retained by glasso along the row/column updates is not the exact inverse of the optimization variable $\boldsymbol{W}$. This can have important consequences. Since the glasso is a block coordinate procedure on the covariance matrix, it maintains a positive definite covariance matrix at every row/column update. However, since the estimated precision matrix is not the exact inverse of $\boldsymbol{W}$, it need not be positive definite. Although it is relatively straightforward to maintain an exact inverse of $\boldsymbol{W}$ along the row/column updates (via simple rank-one updates as before), this inverse $\boldsymbol{W}^{-1}$ need not be sparse. Arbitrary thresholding rules may be used to set some of the entries to zero, but that might destroy the positive-definiteness of the matrix. Since a principal motivation of solving (1) is to obtain a sparse precision matrix (which is also positive definite), returning a dense $\boldsymbol{W}^{-1}$ to (1) is not desirable.

Reference

1. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso." *Biostatistics* 9.3 (2008): 432-441.