Midterm Presentation:

Risk Properties in Sparse Precision Matrix Estimation

Addison J. Hu
Statistics 490
01 March 2017

Outline

1. Graphical Models & Multivariate Gaussian
2. Pairwise Inference for Entrywise Recovery of Σ^{-1}
3. Risk Bounds for Entrywise Recovery in $\|\cdot\|_\infty$
4. Next Steps

Background

Graphical Models

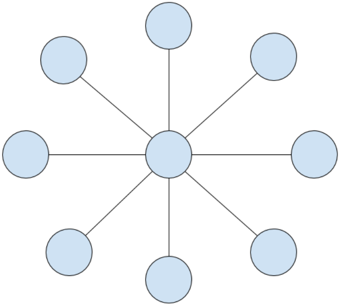
- Graphical models provide a framework within which to consider dependence structure within a group of variables.
- In doing so, we may relax the i.i.d. assumption and still perform inference feasibly.
- Examples:
  - Facebook users graph
  - Gene interaction networks

Markov Random Fields

- Consider a graph $G = (V, E)$, and a corresponding set of random variables $\{X_i\}_{i=1}^{|V|}$, where the random variables are indexed by $u \in V$.
- **Pairwise Markov property:** $X_u \perp\!\!\!\perp X_v | X_{V \setminus \{u,v\}}$ for any two non-adjacency nodes u, v.
- **Local Markov property:** $X_u \perp\!\!\!\perp X_{V \setminus \mathrm{cl}(u)} | X_{\mathrm{nb}(u)}$ for any node v.
- **Global Markov property:** $X_A \perp\!\!\!\perp X_B | X_S$ for disjoint $A, B \subset V$, and a separating subset S.
- Inference is easy when the edges are known; but is more interesting when they are unknown.

Example: Hub and Spoke Model

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$



Multivariate Gaussian

Suppose $X \sim \mathcal{N}(\mu, \Sigma)$. Its density function is given by:

$$p(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

- Closure properties:
  - Sum of independent Gaussian random variables is Gaussian.
  - Marginal of a joint Gaussian distribution is Gaussian.
  - Condition of a joint Gaussian distribution is Gaussian.
- The sparsity pattern of Σ^{-1} concides with the adjacency matrix of the associated MRF.

Multivariate Gaussian, cont.

- Closure under marginalization: Suppose $A \subset V$. Then

$$\Sigma_A = \left(\Sigma_{ij}\right)_{i \in A, j \in A}$$

- Closure under conditioning: Suppose $A, B \subset V$, $A \cup B = V$, $A \cap B = \varnothing$. Then:

$$\left(\Omega_A\right)^{-1} = \Sigma_{A|B}$$
$$\left(\Sigma_A\right)^{-1} = \Omega_{A|B}$$

Precision Matrix Estimation

Maximum Likelihood Estimation

Assume $\mu = 0$. Then the maximum likelihood estimation problem is:

$$\begin{array}{ll} \underset{\Sigma}{\text{maximize}} & -\log\det|\Sigma| - \langle\hat{\Sigma}, \Sigma^{-1}\rangle \\ \text{subject to} & \Sigma \succeq 0 \end{array}$$

- Maximum Likelihood Estimate given by $\hat{\Sigma} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$.
- Idea: $\hat{\Omega} = \hat{\Sigma}^{-1}$.
- Issues:
  - Invertibility & Conditioning
  - Noise & Sparsity

Graphical Lasso

To encourage sparsity, Tibshirani *et al* proposed imposing an entrywise ℓ_1 penalty on Ω.

$$\begin{array}{ll} \underset{\Omega}{\text{maximize}} & \log\det|\Omega| - \langle\hat{\Sigma},\Omega\rangle - \rho\,\|\Omega\|_1 \\ \text{subject to} & \Omega \succeq 0 \end{array}$$

Asymptotic Normal Thresholding (ANT)

- Goal: Obtain entrywise estimates $\hat{\omega}_{ij}$ of Ω that are asymptotically norm and minimax, and then threshold to enforce sparsity.
- Idea: For each pair $A = \{i, j\}$, regress the variables X_i, X_j on all other variables:

  $$\mathbf{X}_A = \mathbf{X}_{A^c}\beta + \epsilon_A$$

  where ϵ_A is a noise term, distributed normally with mean zero, and which are independent of A^c.
- Rationale: $\Omega_{A,A} = \Sigma_{A|A^c} = \text{var}(X_A|X_{A^c}) = \text{var}(\epsilon_A)$. Errors give entries of precision matrix.

Oracle MLE

- Suppose we could draw from the distribution of ϵ_A directly. How would we estimate $\Omega_{A,A}$?
- The maximum likelihood estimator in this case is:

$$\Theta_{A,A}^{ora} = (\theta_{kl}^{ora})_{k,l \in A} = \frac{\epsilon_A^\top \epsilon_A}{n}$$

  where we call Θ^{ora} the *oracle* MLE covariance estimates.
- The corresponding oracle MLE precision estimates are then given by:

$$\Omega_{A,A}^{ora} = (\omega_{kl}^{ora})_{k,l \in A} = \left(\Theta_{A,A}^{ora}\right)^{-1}$$

Residual Estimates

- In practice, we only observe $\mathbf{X}$, so we must estimate ϵ_A.
- Suppose we have an adequate estimates of the regression weights $\hat{\beta}$. Then:

$$\hat{\epsilon}_A = \mathbf{X}_A - \mathbf{X}_{A^c}\hat{\beta}$$

- Consequently:

$$\hat{\Theta}_{A,A} = \frac{\hat{\epsilon}_A^\top \hat{\epsilon}_A}{n}$$
$$\hat{\Omega}_{A,A} = \hat{\Theta}_{A,A}^{-1}$$

Scaled Lasso Estimator

For each $m \in A = \{i, j\}$, perform the optimization:

$$\left\{\hat{\beta}_m, \hat{\theta}_{mm}^{1/2}\right\} = \arg \min_{\substack{b \in \mathbf{R}^{p-2}, \\ \sigma \in \mathbf{R}^+}} \left\{ \frac{\|\mathbf{X}_m - \mathbf{X}_{A^c} b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_{k \in A^c} \frac{\|\mathbf{X}_k\|}{\sqrt{n}} |b_k| \right\}$$

Intuitively, the scaling factor on the ℓ_1 penalty implicitly standardizes the design vector to length $\sqrt{n}$ such that the ℓ_1 penalty is applied to the new coefficients $\frac{\|\mathbf{X}_k\|}{\sqrt{n}} b_k$.

Risk Bounds in $\|\cdot\|_\infty$

Minimaxity

Recall that we call an estimator δ^* *minimax* if it achieves the minimax risk:

$$\sup_{\theta \in \Theta} R(\delta^*, \theta) = \inf_{\delta} \sup_{\theta \in \Theta} R(\delta, \theta) \triangleq \text{minimax risk}$$

where $R(\delta, \theta)$ is a risk function:

$$R(\delta, \theta) \triangleq \mathbf{E}_{X|\theta}\, \ell(\delta(X), \theta)$$

Parameter Space Construction

We consider the following parameter space for $\lambda > 0$:

$$\mathcal{G}^* = \left\{\Omega : s_\lambda(\Omega) \leq s, M^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M\right\}$$

where

$$s_\lambda = s_\lambda(\Omega) = \max_j \sum_{i \neq j} \min\left\{1, \frac{|\omega_{ij}|}{\lambda}\right\}$$

for $\Omega = (\omega_{ij})_{1 \leq i,j \leq p}$.

The authors take λ on the order $\sqrt{\frac{\log p}{n}}$ in this paper.

Risk Upper Bound

- A risk upper bound on an estimator gives a guarantee on its worst case performance.
- The ANT estimator achieves, for some constants C_2, C_3, the following bounds in probability:

$$\max_{\Omega \in \mathcal{G}^*(M,s,\lambda)} \max_{1 \le i \le j \le p} \mathbf{P}\left\{ |\hat{\omega}_{ij} - \omega_{ij}| > C_2 \max\left\{ s\frac{\log p}{n}, \sqrt{\frac{1}{n}} \right\} \right\} \le \varepsilon_0$$

$$\max_{\Omega \in \mathcal{G}^*(M,s,\lambda)} \mathbf{P}\left\{ \left\| \hat{\Omega} - \Omega \right\|_\infty > C_3 \max\left\{ s\frac{\log p}{n}, \sqrt{\frac{\log p}{n}} \right\} \right\} = o(p^{-\delta+3})$$

Risk Upper Bound: Oracle Inequalities

First, we bound the distance from the estimator to the oracle MLEs. There exist constants C_1, C_1' such that:

$$\max_{A:A=\{i,j\}} \mathbf{P}\left\{\left\|\hat{\Theta}_{A,A} - \Theta_{A,A}^{ora}\right\|_\infty > C_1 s \frac{\log p}{n}\right\} \leq o(p^{-\delta+1})$$

and

$$\max_{A:A=\{i,j\}} \mathbf{P}\left\{\left\|\hat{\Omega}_{A,A} - \Omega_{A,A}^{ora}\right\|_\infty > C_1' s \frac{\log p}{n}\right\} \leq o(p^{-\delta+1})$$

Risk Upper Bound: Coupling Argument

- Denote $\kappa_{ij} = \sqrt{n}\frac{\omega_{ij}^{ora} - \omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj} + \omega_{ij}^2}}$.
- Under suitable conditions (KMT Inequality), κ_{ij} behaves roughly like a standard normal random variable.
- This gives a bound in probability on the deviation of ω_{ij}^{ora} from the true ω_{ij}:

$$\mathbf{P}\left\{|\omega_{ij}^{ora} - \omega_{ij}| > C_4\sqrt{\frac{1}{n}}\right\} \leq \frac{3}{4}\epsilon$$

- A similar argument gives:

$$\mathbf{P}\left\{\left\|\hat{\Omega}_{A,A} - \Omega_{A,A}\right\|_\infty > C_5\sqrt{\frac{\log p}{n}}\right\} = o(p^{-\delta})$$

Risk Lower Bound

Le Cam's Two-Point Argument

Next Steps