Midterm Presentation:

# Risk Properties in Bandable Precision Matrix Estimation

Addison Hu
Statistics 490
01 March 2017

# Outline

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

# Refresher

# Graphical Models

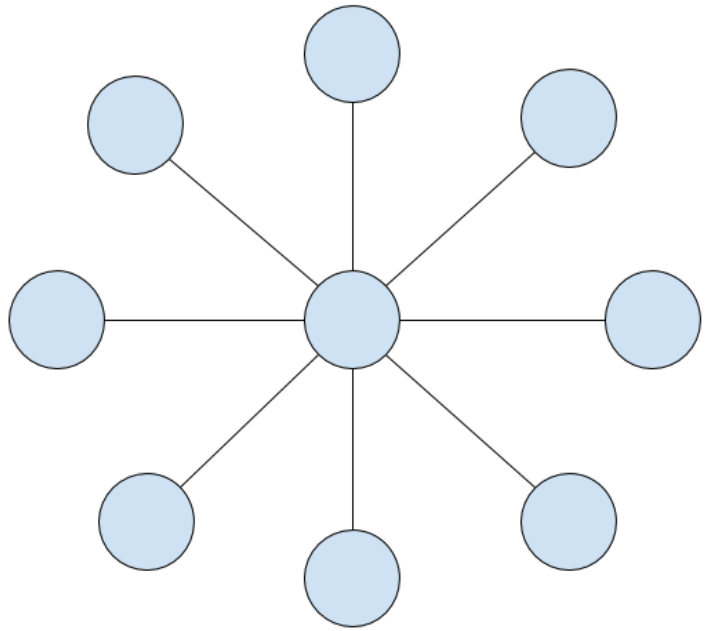
- Graphical models provide a framework within which to consider dependence structure within a group of variables.
- In doing so, we may relax the i.i.d. assumption and still perform inference feasibly.
- Examples:
  - Facebook users graph
  - Gene interaction networks

# 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 $\Sigma^{-1}$ concides with the adjacency matrix of the associated MRF.

# Multivariate Gaussian, cont.

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

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

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

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

# 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 $\ell_1$ penalty on $\Omega$.

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

# Asymptotic Normal Thresholding (ANT)

- Goal: Obtain entrywise estimates $\hat{\omega}_{ij}$ of $\Omega$ 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 $\epsilon_A$ 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 $\epsilon_A$ directly. How would we estimate $\Omega_{A,A}$?
- The maximum likelihood estimator in this case is:

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

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

$$\Omega^{ora}_{A,A} = (\omega^{ora}_{kl})_{k,l \in A} = \left(\Theta^{ora}_{A,A}\right)^{-1}$$

# Residual Estimates

- In practice, we only observe $\mathbf{X}$, so we must estimate $\epsilon_A$.
- 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 $\ell_1$ penalty implicitly standardizes the design vector to length $\sqrt{n}$ such that the $\ell_1$ penalty is applied to the new coefficients $\frac{\|\mathbf{X}_k\|}{\sqrt{n}} b_k$.

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

# Risk Upper Bound

- A risk upper bound on an estimator gives a guarantee on its worst case performance.

# Oracle Inequalities

# Coupling Argument

# Risk Lower Bound

# Le Cam's Two-Point Argument

# Next Steps