STATISTICS 593C: Spring, 2007

Model Selection and Regularization

Jon A. Wellner

Lecture 16 (May 17): In this (half) lecture I will briefly discuss the paper by Meinshausen and Bühlmann (2006) on covariance selection in Gaussian graphical models via the lasso

Suppose that $\underline{X} = (X_1, \ldots, X_p) \sim N_p(\mu, \Sigma)$. Consider a graph $\mathcal{G} = (\Gamma, E)$ associated with the random vector X as follows:

$$\begin{aligned}
&\Gamma = \Gamma(n) = \{1, \ldots, p = p_n\} \text{ is the set of } \textit{nodes} \text{ in the graph,} \\
&E \subset \Gamma \times \Gamma \quad \text{is the set of } \textit{edges} \text{ in the graph, and} \\
&(a,b) \in E \text{ if and only if } X_a \not\perp X_b \Big| \{X_k : \ k \in \Gamma \setminus \{a,b\}\}.
\end{aligned}$$

Thus if $(a,b) \in E^c$ it follows that

$$X_a \perp X_b \Big| \{X_k : \ k \in \Gamma \setminus \{a,b\}\},$$

and this corresponds to a 0 in Σ^{-1}. (Here $X_a \perp X_b | W$ means "X_a and X_b are conditionally independent given the collection of random variables W.")

Now suppose that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. as $\underline{X}$. The goal of "covariance selection" as introduced by Dempster (1963) is to estimate (or "discover" or "learn") the graph $\mathcal{G}$ from the data. See e.g. Buhl (1993) for methods based on maximum likelihood and see Heckerman, Chickering, Meek, Rounthwaite, and Kadie (2000). The goal of Meinshausen and Buhlmann is to use the lasso to study covariance selection when $p = n^\alpha$ with (possibly) $\alpha > 1$.

Define

$$\begin{aligned}
\text{ne}_a = &\text{ smallest subset of } \Gamma \setminus \{a\} \text{ such that} \\
&X_a \text{ is conditionally independent of all other variables ;}
\end{aligned}$$

that is,

$$X_a \perp \{X_k : \ k \in \Gamma \setminus \text{cl}_a\} \Big| X_{ne_a}$$

where $\text{cl}_a \equiv \text{ne}_a \cup \{a\}$.

Assumptions:

A1 $p = O(n^\alpha)$ for some $0 < \alpha < \infty$.

A2 $Var(X_a) = 1$ for all $a \in \Gamma$ and there exists a constant $v^2 > 0$ such that $Var(X_a | X_{\Gamma \setminus \{a\}}) \geq v^2$ for all $a \in \Gamma$.

A3 $\max\{|\mathrm{ne}_a| : \ a \in \Gamma\} = O(n^\kappa)$ for some $0 \le \kappa < 1$. (This is a *sparsity assumption.*)

Before we introduce the rest of the assumptions needed, we define the lasso type procedures introduced by Meinshausen and Buhlmann (2006). For an arbitrary subset $\mathcal{A} \subset \Gamma$, let

$$\theta^{a,\mathcal{A}} \equiv \mathrm{argmin}_{\theta:\theta_k=0, k\notin\mathcal{A}} E\left(X_a - \sum_{k\in\Gamma}\theta_k X_k\right)^2,$$
$$\theta^{a,\Gamma\setminus\{a\}} \equiv \theta^a = (\theta^a_b, \ b \in \Gamma).$$

Then it turns out that

$$\theta^a_b = -\frac{\Sigma^{-1}_{a,b}}{\Sigma^{-1}_{aa}}.$$

Therefore $\mathrm{ne}_a = \{b \in \Gamma : \ \theta^a_b \neq 0\}$.

This suggests defining the lasso estimator of θ^a and hence the lasso estimator of ne_a as follows:

$$\widehat{\theta}^{a,\lambda} \equiv \mathrm{argmin}_{\theta:\theta_a=0}\left\{n^{-1}\|X_{n,a} - \underline{X}_n\theta\|_2^2 + \lambda\|\theta\|_1\right\},$$
$$\widehat{\mathrm{ne}}^\lambda_a \equiv \{b \in \Gamma : \ \widehat{\theta}^{a,\lambda}_b \neq 0\}.$$

where $\underline{X}_n \equiv \sum_1^n \underline{X}_i$.

Before stating the main result of Meinhausen and Buhlmann, we need three more assumptions:

Assumptions, continued:

A4 $\|\theta^{a,\mathrm{ne}_b\setminus\{a\}}\|_1 \le \eta$ for all $a, b \in \Gamma$ such that $(a, b) \in E$.

A5 If $\pi_{a,b}$ denotes the partial correlation between X_a and X_b (i.e. the correlation after correcting for linear effects from all the other variables), then

$$|\pi_{a,b}| \ge \delta n^{-(1-\xi)/2} \qquad \text{for all } \ (a, b) \in E$$

for some $\delta > 0$ and $\xi > \kappa$.

A6 (Neighborhood stability) There is a $\delta < 1$ such that for all $a, b \in \Gamma$ with $b \notin \mathrm{ne}_a$,

$$S_a(b) \equiv \sum_{k\in\mathrm{ne}_a} \mathrm{sign}(\theta^{a,\mathrm{ne}_a}_k)\theta^{k,\mathrm{ne}_a}_b$$

satisfies $|S_a(b)| < \delta$.

With this preparation we can state the main result of M & B (2006):

Theorem. Suppose that A1-A6 hold. Let $\lambda_n \sim dn^{-(1-\epsilon)/2}$ with some $\kappa < \epsilon < \xi$ and $d > 0$. Then there exists a $c > 0$ such that for all $a \in \Gamma$,

$$P(\widehat{\mathrm{ne}}^\lambda_a \subset \mathrm{ne}_a) = 1 - O(e^{-cn^\epsilon}) \qquad \text{as } \ n \to \infty, \ \text{ and}$$
$$P(\mathrm{ne}_a \subset \widehat{\mathrm{ne}}^\lambda_a) = 1 - O(e^{-cn^\epsilon}) \qquad \text{as } \ n \to \infty.$$