High dimensional graphs and variable selection with the Lasso
Nicolai Meinshausen and Peter Buhlmann
The annals of Statistics (2006)

presented by Jee Young Moon

Feb. 19. 2010
Inverse covariance matrix

0 in $\Sigma^{-1}$ = conditional independence ($X_a \perp X_b | \text{all the remaining variables}$) = no edge between these two variables (nodes)

Traditionally,

- Dempster (1972): introduced covariance selection. Discovering the conditional independence.
- Forward search: edges are added iteratively.
- MLE fit (Speed and Kiiveri 1986) for $O(p^2)$ different models. But, the existence of MLE is not guaranteed in general if the number of observation is smaller than the number of nodes (Buhl 1993)
- Neighborhood selection with the Lasso (here) : optimization of a convex function, applied consecutively to each node in the graph.
Neighborhood

Neighborhood $ne_a$ of a node $a \in \Gamma$

$= \text{smallest subset of } \Gamma \setminus \{a\} \text{ so that } X_a \perp \text{all the remaining} | X_{ne_a}$

$= \{ b \in \Gamma \setminus \{a\} : (a, b) \in E \}.$
Notations

$p(n) = |\Gamma(n)|$ = the number of nodes (the number of variables)

$n$: the number of observations

Optimal prediction of $X_a$ given all remaining variables

$$\theta^a = \operatorname{arg\ min}_{\theta: \theta_a = 0} E(X_a - \sum_{k \in \Gamma(n)} \theta_k X_k)^2$$

Optimal prediction $\theta^a, A$ where $A \subseteq \Gamma(n) \setminus \{a\}$

$$\theta^a, A = \operatorname{arg\ min}_{\theta: \theta_k = 0, \forall k \notin A} E(X_a - \sum_{k \in \Gamma(n)} \theta_k X_k)^2$$

$A$: active set.

Relation to conditional independence is

$$\theta^a_b = -\Sigma^{-1}_{ab} / \Sigma^{-1}_{aa}.$$

$\text{ne}_a = \{b \in \Gamma(n) : \theta^a_b \neq 0\}.$
Neighborhood selection with Lasso

Lasso estimate $\hat{\theta}^{a,\lambda}$ of $\theta^a$

$$\hat{\theta}^{a,\lambda} = \arg\min_{\theta: \theta_a = 0} (n^{-1} \|X_a - X\theta\|^2 + \lambda \|\theta\|_1)$$ (3)

Neighborhood estimate

$$\hat{ne}_a^\lambda = \{ b \in \Gamma(n) | \hat{\theta}_b^{a,\lambda} \neq 0 \}$$
prediction-oracle value

\[ \lambda_{\text{oracle}} = \arg \min_{\lambda} E(X_a - \sum_{k \in \Gamma(n)} \hat{\theta}_k^{a,\lambda} X_k)^2 \]

**Proposition 1.**

Let the number of variables grow to infinity, \( p(n) \to \infty \) for \( n \to \infty \) with \( p(n) = o(n^\gamma) \) for some \( \gamma > 0 \). Assume that the covariance matrices \( \Sigma(n) \) are identical to the identity matrix except for some pair \((a, b) \in \Gamma(n) \times \Gamma(n)\) for which \( \Sigma_{ab}(n) = \Sigma_{ba}(n) = s \) for some \( 0 < s < 1 \) and all \( n \in \mathbb{N} \). The probability of selecting the wrong neighborhood for node \( a \) converges to 1 under the prediction-oracle penalty

\[ P(\hat{\text{ne}}^\lambda_{a,\text{oracle}} \neq \text{ne}_a) \to 1 \text{ for } n \to \infty. \]
\[ \theta^a = (0, -K_{ab}/K_{aa}, 0, 0, \ldots) = (0, s, 0, 0, \ldots). \] To be \( \hat{n}_{e^a} = ne_a, \)
\[ \hat{\theta}^{a,\lambda} = (0, \tau, 0, 0, \ldots) \] is the oracle Lasso solution for some \( \tau \neq 0. \) Then, it is the same as
\[
P(\exists \lambda, \tau \geq s : \hat{\theta}^{a,\lambda} = (0, \tau, 0, 0, \ldots)) \to 0 \text{ as } n \to \infty.
\]
and 2. \((0, \tau, 0, 0, \ldots)\) cannot be the oracle Lasso solution as long as \( \tau < s. \)

1. If \( \hat{\theta} = (0, \tau, 0, \ldots) \) is a Lasso solution, from Lemma 1 and positivity of \( \tau, \)
\[
< X_1 - \tau X_2, X_2 > \geq | < X_1 - \tau X_2, X_k > | \ \forall k \in \Gamma(n), k > 2.
\]
Substituting \( X_1 = sX_s + W_1 \) yields
\[
< W_1, X_2 > - (\tau - s) < X_2, X_2 > \geq | < W_1, X_k > - (\tau - s) < X_2, X_k > | .
\]
Let \( U_k = < W_1, X_k > . \) \( U_k, k = 2, \ldots, p(n) \) are exchangeable. Let
\[
D = < X_2, X_2 > - \max_{k \in \Gamma(n), k > 2} | < X_2, X_k > | .
\]
It is sufficient to show
\[
P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k + (\tau - s)D) \to 0 \text{ for } n \to \infty.
\]
Since $\tau - s > 0,$

$$P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k + (\tau - s)D) \leq P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k) \text{ when } D \geq 0.$$  

$$P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k + (\tau - s)D) \leq 1 \text{ when } D < 0.$$  

$$P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k + (\tau - s)D) \leq P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k) + P(D < 0).$$

By Berstein inequality and $p(n) = o(n^{\gamma}),$

$$P(D < 0) \to 0 \text{ for } n \to \infty.$$  

Since $U_2, \ldots, U_{p(n)}$ are exchangeable, $P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k) = P(U_3 > \max_{k \in \Gamma(n), k=2,>3} U_k) = \cdots = P(U_p > \max_{2 \leq k < p-1} U_k)$ and sum of those should be 1. Therefore,

$$P(U_2 > \max_{k \in \Gamma(n), k > 2} U_k) = (p(n) - 1)^{-1} \to 0 \text{ for } n \to \infty.$$
2. \((0, \tau, 0, \ldots)\) with \(\tau < s\) cannot be the oracle Lasso solution.
Suppose \((0, \tau_{\text{max}}, 0, 0, \ldots)\) is the Lasso solution \(\hat{\theta}^{a,\lambda}\) for some \(\lambda = \tilde{\lambda} > 0\) with \(\tau_{\text{max}} < s\). Since \(\tau_{\text{max}}\) is the maximal value such that \((0, \tau, 0, \ldots)\) is a Lasso solution, there exists some \(k \in \Gamma(n) > 2\) such that

\[
|n^{-1} < X_1 - \tau_{\text{max}} X_2, X_2| = |n^{-1} < X_1 - \tau_{\text{max}} X_2, X_k| = \tilde{\lambda}.
\]

For sufficiently small \(\delta \lambda \geq 0\), a Lasso solution for the penalty \(\tilde{\lambda} - \delta \lambda\) is given by

\[
(0, \tau_{\text{max}} + \delta \theta_2, \delta \theta_3, 0, \ldots).
\]

From LARS, \(\delta \theta_2 = \delta \theta_3\). If we compare the squared error for these solution

\[
L_{\delta \theta} - L_0 = -2(s - \tau_{\text{max}})\delta \theta + 2\delta \theta^2 < 0 \text{ for any } 0 < \delta \theta < 1/2(s - \tau_{\text{max}})
\]
Lemma 1

Lasso estimate $\hat{\theta}^{a,A,\lambda}$ of $\theta^{a,A}$ is given by

$$\hat{\theta}^{a,A,\lambda} = \arg\min_{\theta : \theta_k = 0, \forall k \notin A} (n^{-1} \|X_a - X\theta\|^2 + \lambda \|\theta\|_1) \quad (10)$$

Lemma 1

Given $\theta \in \mathbb{R}^{p(n)}$, let $G(\theta)$ be a $p(n)$-dimensional vector with elements

$$G_b(\theta) = -2n^{-1} < X_a - X\theta, X_b > .$$

A vector $\hat{\theta}$ with $\hat{\theta}_k = 0, \forall k \in \Gamma(n) \setminus A$ is a solution to the above

$$\iff \text{ for all } b \in A, \ G_b(\hat{\theta}) = -\text{sign}(\hat{\theta}_b)\lambda \text{ in case } \hat{\theta}_b \neq 0 $$

and $|G_b(\hat{\theta})| \leq \lambda$ in case $\hat{\theta}_b = 0$. Moreover, if the solution is not unique and $|G_b(\hat{\theta})| < \lambda$ for some solution $\theta$, then $\hat{\theta}_b = 0$ for all solution of the above.
Proof.

\[ D(\theta) = \text{subdifferential of } (n^{-1}\|X_a - X\theta\|^2 + \lambda\|\theta\|_1) \text{ with respect to } \theta = \{ G(\theta) + \lambda e, e \in S \} \]

where \( S \subset \mathbb{R}^{p(n)} \) is given by

\[ S = \{ e \in \mathbb{R}^{p(n)} | e_b = \text{sign}(\theta_b) \text{ if } \theta_b \neq 0 \text{ and } e_b \in [-1, 1] \} \]

\( \hat{\theta} \) is a solution to the above iff \( \exists d \in D(\theta) \text{ so that } d_b = 0 \forall b \in A. \)
Assumptions

\( X \sim N(0, \Sigma) \)

**High-dimensionality** Assumption 1. There exists \( \gamma > 0 \) so that \( p(n) = O(n^{\gamma}) \) for \( n \to \infty \).

**Non-singularity** Assumption 2. (a) For all \( a \in \Gamma(n) \) and \( n \in \mathcal{N} \), \( \text{Var}(X_a) = 1 \). (b) There exists \( \nu^2 > 0 \) so that for all \( n \in \mathcal{N} \) and \( a \in \Gamma(n) \), \( \text{Var}(X_a | X_{\Gamma(n) \setminus \{a\}}) \geq \nu^2 \). [This excludes singular or nearly singular covariance matrices.]

**Sparsity** Assumption 3. There exists some \( 0 \leq \kappa < 1 \) so that \( \max_{a \in \Gamma(n)} |ne_a| = O(n^{\kappa}) \) for \( n \to \infty \). [restriction on the size of the neighborhood].

Assumption 4. There exists some \( \vartheta < \infty \) so that for all neighboring nodes \( a, b \in \Gamma(n) \) and all \( n \in \mathbb{N} \), \( \|\theta^{a,ne_b \setminus \{a\}}\|_1 \leq \vartheta \). [This is fulfilled if assumption 2 holds and the size of the overlap of neighborhoods is bounded by an arbitrarily large number from above. ]
Magnitude of partial correlations Assumption 5. There exists a constant $\delta > 0$ and some $\xi > \kappa$ so that for every $(a, b) \in E$, $|\pi_{a,b}| \geq \delta n^{-(1-\xi)/2}$.

Neighborhood stability $S_a(b) := \sum_{k \in ne_a} sign(\theta_{k}^{a,ne_a})\theta_{k}^{b,ne_a}$.

Assumption 6. There exists some $\delta < 1$ so that for all $a, b \in \Gamma(n)$ with $b \notin ne_a$, $|S_a(b)| < \delta$. 
Theorem 1: controlling type-I error

Let assumptions 1-6 be fulfilled. Let the penalty parameter satisfy $\lambda_n \sim dn^{-(1-\varepsilon)/2}$ with some $\kappa < \varepsilon < \xi$ and $d > 0$. There exists some $c > 0$ so that, for all $a \in \Gamma(n)$,

$$P(\hat{ne}_a^\lambda \subseteq ne_a) = 1 - O(\exp(-cn^\varepsilon)) \text{ for } n \to \infty.$$ 

It means that the probability of falsely including any of the non-neighboring variables is vanishing exponentially fast. Proposition 3 says that assumption 6 cannot be relaxed.

Proposition 3

If there exists some $a, b \in \Gamma(n)$ with $b \notin ne_a$ and $|S_a(b)| > 1$, then

$$P(\hat{ne}_a^\lambda \subseteq ne_a) \to 0 \text{ for } n \to \infty.$$
Proof of Thm 1

\[ P(\tilde{ne}_a^\lambda \subseteq ne_a) = 1 - P(\exists b \in \Gamma(n) \setminus cl_a : \hat{\theta}_b^{a,\lambda} \neq 0). \]

Consider the Lasso estimate $\hat{\theta}^{a,ne_a,\lambda}$ which is constrained to have non-zero components only in $ne_a$. Let $\mathcal{E}$ be the event

\[ \max_{k \in \Gamma(n) \setminus cl_a} |G_k(\hat{\theta}^{a,ne_a,\lambda})| < \lambda. \]

On this event, by Lemma 1, $\hat{\theta}^{a,ne_a,\lambda}$ is a solution of (3) with $\mathcal{A} = \Gamma(n) \setminus \{a\}$ as well as a solution of (10).

\[ P(\exists b \in \Gamma(n) \setminus cl_a : \hat{\theta}_b^{a,\lambda} \neq 0) \leq 1 - P(\mathcal{E}) = P(\max_{k \in \Gamma(n) \setminus cl_a} |G_k(\hat{\theta}^{a,ne_a,\lambda})| \geq \lambda). \]

It is sufficient to show there exists a constant $c > 0$ so that for all $b \in \Gamma(n) \setminus cl_a$,

\[ P(|G_b(\hat{\theta}^{a,ne_a,\lambda})| \geq \lambda) = O(\exp(-cn^\epsilon)). \]
cont. prof of Thm 1.

One can write for any \( b \in \Gamma(n) \setminus cl_a \),

\[
X_b = \sum_{m \in ne_a} \theta_{m}^{b, ne_a} X_m + V_b,
\]

where \( V_b \sim N(0, \sigma_b^2) \) for some \( \sigma_b^2 \leq 1 \) and \( V_b \) is independent of \( \{X_m | m \in cl_a\} \). Plugging this in gradient calculation,

\[
G_b(\hat{\theta}^a_{ne_a, \lambda}) = -2n^{-1} \sum_{m \in ne_a} \theta_{m}^{b, ne_a} < X_a - X\hat{\theta}^a_{ne_a, \lambda}, X_m > -2n^{-1} < X_a - X\hat{\theta}^a_{ne_a, \lambda}, V_b >.
\]

By lemma 2, there exists some \( c > 0 \) so that with probability

\[
1 - O(\exp(-cn^\epsilon)),
\]

\[
\text{sign}(\hat{\theta}^a_{ne_a, \lambda}) = \text{sign}(\theta_{k}^{a, ne_a}), \forall k \in ne_a.
\]
cont. prof of Thm 1.

With Lemma 1, assumption 6, we get with probability $1 - O(\exp(-cn^\epsilon))$ and some $\delta < 1$,

$$|G_b(\hat{\theta}^{a,ne_a,\lambda})| \leq \delta \lambda + |2n^{-1} < X_a - X\hat{\theta}^{a,ne_a,\lambda}, V_b > |.$$

Then, it remains to be shown that

$$P(|2n^{-1} < X_a, V_b > | \geq (1 - \delta)\lambda) = O(\exp(-cn^\epsilon)).$$
Theorem 2

Let the assumptions of Theorem 1 be fulfilled. For $\lambda = \lambda_n$ as a in Theorem 1, it holds for some $c > 0$ that

$$P(ne_a \subseteq \hat{ne}_a^\lambda) = 1 - O(\exp(-cn^\epsilon))$$

for $n \to \infty$.

Proposition 4 says that assumption 5 cannot be relaxed.

**Proposition 4**

Let the assumptions of Theorem 1 be fulfilled with $\vartheta < 1$ in Assumption 4. For $a \in \Gamma(n)$, let there be some $b \in \gamma(n) \backslash \{a\}$ with $\pi_{ab} \neq 0$ and $|\pi_{ab}| = O(n^{-(1-\xi)/2})$ for $n \to \infty$ for some $\xi < \epsilon$. Then

$$P(b \in \hat{ne}_a^\lambda) \to 0$$

for $n \to \infty$. 
Proof of Thm2

\[ P(\text{ne}_a \subseteq \hat{\text{ne}}_a^\lambda) = 1 - P(\exists b \in \text{ne}_a : \hat{\theta}_b^{a,\lambda} = 0) \]

Let \( \mathcal{E} \) be the event

\[ \max_{k \in \Gamma(n) \setminus \text{cl}_a} |G_k(\hat{\theta}^{a,\text{ne}_a,\lambda})| < \lambda. \]

As in Thm 1, \( \hat{\theta}^{a,\text{ne}_a,\lambda} \) is a solution of (3). Then,

\[ P(\exists b \in \text{ne}_a : \theta^a,\lambda_b = 0) \leq P(\exists b \in \text{ne}_a : \hat{\theta}^{a,\text{ne}_a,\lambda}_b = 0) + P(\mathcal{E}^c). \]

\[ P(\mathcal{E}^c) = O(\exp(-cn^\epsilon)) \]

by theorem 1 and

\[ P(\hat{\theta}^{a,\text{ne}_a,\lambda} = 0) = O(\exp(-cn^\epsilon)) \]

by lemma 2.
Edge set

Ideally, edge set can be given by

\[ E = \{(a, b) : a \in ne_b \land b \in ne_a\}. \]

An estimate of the edge set is

\[ \hat{E}^{\wedge,\wedge} = \{(a, b) : a \in \hat{ne}_b^{\lambda} \land b \in \hat{ne}_a^{\lambda}\}. \]

Or

\[ \hat{E}^{\wedge,\vee} = \{(a, b) : a \in \hat{ne}_b^{\lambda} \lor b \in \hat{ne}_a^{\lambda}\}. \]

Corollary 1

Under the conditions of Theorem 2, for some \( c > 0 \),

\[ P(\hat{E}^{\lambda} = E) = 1 - P(\exp(-cn^\epsilon)) \] for \( n \to \infty \).
How to choose the penalty?
For any level $0 < \alpha < 1$, the penalty

$$
\lambda(\alpha) = \frac{2\hat{\sigma}_a}{\sqrt{n}} \Phi^{-1}(\frac{\alpha}{2p(n)^2}).
$$

**Theorem 3**

Assumptions 1-6 be fulfilled. Using the penalty $\lambda(\alpha)$, it holds for all $n \in \mathbb{N}$ that

$$
P(\exists a \in \Gamma(n) : \hat{C}_a^\lambda \not\subseteq C_a) \leq \alpha.
$$

This constrains the probability of (falsely) connecting two distinct connectivity components of the true graph.