A Note on the Lasso for Gaussian Graphical Model Selection

Nicolai Meinshausen
Seminar für Statistik, Leonhardstrasse 27
ETH Zürich, CH-8092 Zürich

October 3, 2005

Abstract

Inspired by the success of the Lasso for regression analysis (Tibshirani, 1996), it seems attractive to estimate the graph of a multivariate normal distribution by ℓ_1-norm penalised likelihood maximisation. The objective function is convex and the graph estimator can thus be computed efficiently, even for very large graphs. However, we show in this note that the resulting estimator is not consistent for some graphs.

1 Introduction

Let X be a p-dimensional random vector $X = (X^{(1)}, \ldots, X^{(p)})$ with a multivariate normal distribution,

$$X \sim \mathcal{N}(0, \Sigma). \tag{1.1}$$

Denote the concentration matrix by $K = \Sigma^{-1}$. The corresponding graph $\mathcal{G} = \mathcal{G}(K)$ is given by $\mathcal{G} = (\Gamma, E)$ with nodes $\Gamma = \{1, \ldots, p\}$. The edge set is given by all pairs $(a, b) \in \Gamma \times \Gamma$ so that $a \neq b$ and $K_{a,b} \neq 0$.

The distribution of X belongs to an exponential family, and the log-likelihood $\ell_n(K)$ of i.i.d. observations X_i, $i = 1, \ldots, n$ under concentration matrix $K = \Sigma^{-1}$ is given by

$$\ell_n(K) = -\sum_{i=1}^{n} \{X_i^T K X_i - A(K)\},$$

where $A(K) \propto -\log|K|$. The MLE of the concentration matrix is

$$\hat{K}^{MLE} = \arg\min_{M} \quad \{-\ell_n(M)\}., \tag{1.2}$$

where $\ell_n(M)$ is the log-likelihood of the observations $X_i, i = 1, \ldots, n$ under concentration matrix M. The MLE of the covariance matrix,

$$\Sigma^{MLE} = \{\hat{K}^{MLE}\}^{-1},$$

is componentwise identical to the empirical covariances between the corresponding pair of variables (Lauritzen, 1996).

It is often of interest to estimate the underlying graph of the distribution, that is estimate structural zeros of the inverse covariance matrix. The MLE leads to a full graph estimator, that is all edges are in general present in the MLE. Edge selection is usually achieved by adding a l_0-type penalty to the likelihood,

$$\hat{K}^{pen} = \arg\min_{M} \quad \{-\ell_n(M) + n\lambda\|M\|_0\}, \tag{1.3}$$

where $\|M\|_0 := \sum_{a \neq b} 1\{M_{a,b} \neq 0\}$ counts the number of edges present in the graph (Dempster, 1972; Edwards, 2000).

The convexity of the likelihood function is lost in the objective function in (1.3) as the added penalty term is not convex. The solution to (1.3) is thus usually found by greedy search, using forward selection or backward elimination of edges. This greedy search (let alone finding the exact solution) becomes infeasible for large graphs.

A remedy to the high computational complexity of (1.3) is to add a convex penalty term instead of the l_0-penalisation. For ordinary regression, Tibshirani (1996) proposed the Lasso, by adding a ℓ_1-norm penalisation instead of

the more common l_0-penalisation. An efficient algorithm for computing the exact solution to the resulting convex optimisation problem has been proposed recently in Efron et al. (2004). It seems interesting to examine if a ℓ_1-penalisation could lead to a reasonable graph estimator.

2 Lasso Graph Estimators

The Lasso-type estimator of K can be defined as the minimizer of the negative log-likelihood with an added penalty term proportional to the ℓ_1-norm of the non-diagonal elements of the concentration matrix,

$$\hat{K}^{\lambda} = \arg\min_{M} \quad \{-\ell_n(M) + n\lambda\|M\|_1\}, \tag{2.4}$$

where $\|M\|_1 := \sum_{a\neq b} |M_{a,b}|$. We show in the following that the Lasso-estimator (2.4) is not consistent for estimating the structure of the graph, regardless of how the penalty parameter λ is chosen.

An alternative approach to the Lasso-estimator would be to penalize the sum of the absolute values of the partial correlations instead of the diagonal entries of the concentration matrix. However, the concern about consistency is the same as for the present formulation of the Lasso estimator.

Solutions characteristics.

We characterize the solutions to (2.4). The distinctive feature of the solutions is a soft-thresholding of the estimated covariances $\hat{\Sigma}^{\lambda} = \{\hat{K}^{\lambda}\}^{-1}$.

Proposition 1 *For any λ and $\{a, b\} \in \Gamma \times \Gamma$ so that $\hat{K}^{\lambda}_{a,b} \neq 0$ and $a \neq b$, it holds that*

$$\hat{\Sigma}^{\lambda}_{a,b} - \hat{\Sigma}^{MLE}_{a,b} = \lambda \,\mathrm{sign}(\hat{K}^{\lambda}_{a,b}).$$

For any λ and $\{a, b\} \in \Gamma \times \Gamma$ so that $\hat{K}^{\lambda}_{a,b} = 0$ and $a \neq b$, it holds that

$$|\hat{\Sigma}^{\lambda}_{a,b} - \hat{\Sigma}^{MLE}_{a,b}| \leq \lambda.$$

Finally, for all $a \in \Gamma$, $\hat{\Sigma}^{\lambda}_{a,a} = \hat{\Sigma}^{MLE}_{a,a}$.

Proof. The concentration matrix $\hat{K}^\lambda$ is a solution to (2.4) if and only if the subdifferential of the objective function in (2.4) with respect to $K_{a,b}$ contains 0 for every pair $a, b \in \Gamma \times \Gamma$. If $K_{a,b} \neq 0$, the subdifferential of the objective function in (2.4) with respect to $K_{a,b}$ is identical to

$$\sum_{i=1}^{n} \{X_i^{(a)} X_i^{(b)} + (\partial/\partial K_{a,b}) A(K)\} + n\lambda \operatorname{sign}(\hat{K}_{a,b}^\lambda).$$

Using a general property of exponential distributions, the derivative of $A(K)$ with respect to $K_{a,b}$ is given by the expectation of the sufficient statistics $-X^{(a)}X^{(b)}$ under concentration matrix K, which is identical to $-(K^{-1})_{a,b}$. Using $\sum_{i=1}^{n}(X_i^{(a)}X_i^{(b)}) = n\hat{\Sigma}_{a,b}^{MLE}$ and $E_{\hat{K}^\lambda}(X_i^{(a)}X_i^{(b)}) = \hat{\Sigma}_{a,b}^\lambda$, the first part of the Proposition follows. For the second part, where $\hat{K}_{a,b}^\lambda = 0$, the subdifferential is given by the interval

$$[\hat{\Sigma}_{a,b}^{MLE} - \hat{\Sigma}_{a,b}^\lambda - \lambda, \hat{\Sigma}_{a,b}^{MLE} - \hat{\Sigma}_{a,b}^\lambda + \lambda],$$

and the claim follows. The third part about the diagonal elements follows analogously.

3 Consistency

Let $G(K) = (\Gamma, E)$ be again the graph associated with the concentration matrix $K = \Sigma^{-1}$, so that a pair of nodes in Γ is included in the edge set E if and only if the corresponding entry in K is non-zero. For an estimate $\hat{K}$ of the concentration matrix, the corresponding estimate $\hat{E}^\lambda$ of the edge set E is given by the set of non-zero (and non-diagonal) entries of $\hat{K}^\lambda$.

We show for an example that the estimator (2.4) cannot be consistent for graph estimation.

Example 1 Let $\Gamma = \{1, 2, 3, 4\}$ and the edge set E given by all $(a, b) \in \Gamma \times \Gamma$ so that $a \neq b$ and $(a, b) \neq (1, 4)$. The covariance matrix is then determined for some $0 < \rho < 1/\sqrt{2}$ by

$$\Sigma_{a,a} = 1 \quad \text{for} \quad a \in \{1, 2, 3, 4, 5\}$$

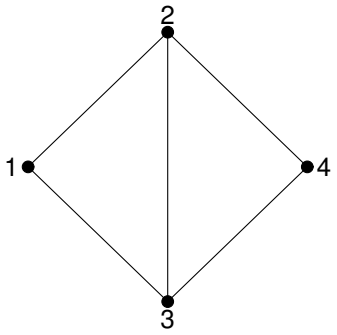

Figure 1: *The graph of Example 1.*

$$\Sigma_{a,b} = \rho \quad \text{for} \quad \{a,b\} \in \{\{1,2\},\{1,3\},\{2,4\},\{(3,4\}\}$$
$$\Sigma_{a,b} = 0 \quad \text{for} \quad \{a,b\} = \{2,3\}$$

A picture of the graph is given in Figure 1.

The population case.

We make first an argument why the Lasso estimator is not consistent for the population case, regardless of how the penalty parameter is chosen. Consider that the MLE-covariance matrix is identical to the true covariance matrix. In some sense, this is the best case one could hope for. The corresponding estimate is denoted by K^λ,

$$K^\lambda = \arg\min_M \quad \{E(-\ell_n(M)) + \lambda\|M\|_1\},$$

For $\lambda = 0$, the estimator is identical to the true concentration matrix, $K^{MLE} = K$, and $\Sigma^{MLE} = \{K^{MLE}\}^{-1} = \Sigma$. Nevertheless, this Lasso estimator K^λ produces the wrong graph for arbitrarily small values of λ, if $\rho > -1 + (3/2)^{1/2} \approx 0.23$. As $K^\lambda \to K$ for $\lambda \to 0$, it holds for sufficiently small values of λ that $\text{sign}(K^\lambda_{a,b}) = \text{sign}(K_{a,b})$ for all $(a,b) \in \Gamma \times \Gamma$ with $K_{a,b} \neq 0$. Then, by Proposition 1, it follows that that $\Sigma^\lambda_{2,3} = \lambda$ and

$$\Sigma^\lambda_{1,2} = \Sigma^\lambda_{1,3} = \Sigma^\lambda_{2,4} = \Sigma^\lambda_{3,4} = \rho - \lambda.$$

If indeed $K^{\lambda}_{1,4} = 0$, this would imply that the covariance between nodes 1 and 4 is given by

$$2(\rho - \lambda)^2/(1 + \lambda) = 2\rho^2 - (4\rho + 2\rho^2)\lambda + O(\lambda^2)$$

The true covariance is $2\rho^2$ and thus $|\Sigma^{\lambda}_{1,4} - \Sigma_{1,4}| > (4\rho + 2\rho^2)\lambda + O(\lambda^2)$. For small enough positive values of λ, the r.h.s. is larger than λ (using $\rho > 0.23$). This is in contradiction to Proposition 1 by which $|\Sigma^{\lambda}_{1,4} - \Sigma_{1,4}| \leq \lambda$. Hence $K^{\lambda}_{1,2} \neq 0$. The graph is thus not estimated correctly with the population estimate K^{λ} for small penalties λ even though adding no penalty at all ($\lambda = 0$) would produce the correct result.

The penalty has thus for the edge 1-4 an effect that is opposite to what is intended: instead of "pushing" the estimate of the partial correlation towards zero with increasing size of the penalty, the penalty rather pushes the estimate of the partial correlation away from zero.

Finite samples.

As the effect of the penalty is opposite to what is intended for the edge 1-4, it is not surprising that the random fluctuations due to a limited number of observations cannot make up for this effect. Using the results above, it is indeed shown in the following that the Lasso estimator is not consistent for finite sample sizes.

For notational simplicity, denote the estimated entry $\hat{K}^{\lambda}_{1,4}$ by $\hat{k}^{\lambda}$. The unpenalized estimate, $\hat{k}^0$, is unbiased, that is $E(\hat{k}^0) = 0$ for $n > 6$ (Lauritzen, 1996). There is always positive probability that the estimate is negative and, because of asymptotic normality of $\hat{k}^0$,

$$\lim_{n\to\infty} P(\hat{k}^0 < 0) = 0.5. \tag{3.5}$$

We show in the following that, if indeed $\hat{k}^0 < 0$, then an added penalty λ pushes the estimate even further away from zero. That is, for sufficiently small λ, it holds with probability converging to 0.5 for $n \to \infty$ that

$$\hat{k}^{\lambda} < \hat{k}^0 < 0, \tag{3.6}$$

and the edge 1-4 is in this case falsely included into the edge set estimate. Using (3.5), it follows that for all penalty sequences with $\lambda = o(1)$ for $n \to \infty$, the edge set is different from the true edge set with non-negligible probability.

Theorem 1 *Consider the graph in Example 1 with edge set E and edge set estimate $\hat{E}^\lambda$. If $\rho > -1 + (3/2)^{1/2} \approx 0.23$, it holds for any sequence $\lambda = \lambda_n$ with $\lambda = o(1)$ for $n \to \infty$, that the edge set is wrongly estimated with positive probability,*

$$\liminf_{n\to\infty} P(\hat{E}^\lambda \neq E) \geq 0.5.$$

Proof. We only need to prove that (3.6) holds with positive probability. This is true by the following Lemma 1, which completes the proof. □

Lemma 1 Let $\lambda = \lambda_n = o(1)$ for $n \to \infty$. If $\rho > -1 + (3/2)^{1/2} \approx 0.23$,

$$\liminf_{n\to\infty} P(\hat{k}^\lambda < \hat{k}^0 < 0) \geq 0.5.$$

Proof. The probability that $\hat{k}^0 = \hat{K}^0_{1,4} < 0$ is converging to 0.5 for $n \to \infty$. Using

$$\hat{k}^\lambda = \hat{k}^0 + \{\frac{\partial \hat{k}^\lambda}{\partial \lambda}(\lambda = 0)\}\lambda + O_p(\lambda^2),$$

it suffices to show that there exists some positive constant $c(\rho) > 0$ so that, with probability converging to 1 for $n \to \infty$,

$$\frac{\partial \hat{k}^\lambda}{\partial \lambda}(\lambda = 0) < -c(\rho).$$

Using $\hat{K}^0 \to_p K$ (Lauritzen, 1996), it follows that, with probability converging to 1 for $n \to \infty$, $\text{sign}(\hat{K}^0_{a,b}) = \text{sign}(K_{a,b})$ for all $\{a, b\} \neq \{1, 4\}$. Define matrix Δ as

$$\Delta_{a,b} = \begin{cases} 0 & a = b \\ 1 & \{a, b\} = \{2, 3\} \\ -1 & \text{otherwise} \end{cases}.$$

As $\lim_{n\to\infty} P(\hat{k}^0 = \hat{K}^0_{1,4} < 0) = 0.5$, it follows by convergence in probability of $\hat{K}^0$ to K for $n \to \infty$ that

$$\lim_{n\to\infty} P\{\text{sign}(\hat{K}^0) = \Delta\} = 0.5 \tag{3.7}$$

Suppose for the following that indeed $\text{sign}(\hat{K}^0) = \Delta$. By Proposition 1, the estimated covariance $\hat{\Sigma}^\lambda = \{\hat{K}^\lambda\}^{-1}$ is then, for sufficiently small values of λ, given by

$$\hat{\Sigma}^\lambda = \hat{\Sigma}^{MLE} + \lambda\Delta.$$

Using that $\hat{K}^\lambda = \{\hat{\Sigma}^\lambda\}^{-1}$, the derivative of $\hat{K}^\lambda$ with respect to λ is given under the above sign-condition by

$$\frac{\partial \hat{K}^\lambda}{\partial \lambda}(\lambda = 0) = -\hat{K}^0 \Delta\ \hat{K}^0.$$

It holds that $\hat{K}^0 \to_p K$. Note that, for $\rho > -1 + (3/2)^{1/2}$, there exists some $c(\rho) > 0$ so that

$$-(K\Delta K)_{1,4} < -2c(\rho).$$

Hence

$$\lim_{n\to\infty} P\{-(\hat{K}^0 \Delta\ \hat{K}^0)_{1,4} < -c(\rho)\} = 1.$$

Thus, using (3.7),

$$\liminf_{n\to\infty} P\Big\{\hat{k}^0 < 0\ \wedge\ \frac{\partial \hat{k}^\lambda}{\partial \lambda}(\lambda = 0) < -c(\rho)\Big\} = 0.5,$$

which completes the proof. □

References

Dempster, A. (1972). Covariance selection. *Biometrics 28*, 157–175.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–451.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B *58*, 267–288.