# Summary of *Model selection and estimation in the Gaussian graphical model*

*Stat 679 Graphical Models*

**December 14, 2015**


**Group members:**

Weilan Yang

Qingyang Liu

Hao Xin



**Abstract**

Estimating the concentration matrix, i.e. the inverse of covariance matrix, is vital in graphical model structure inference. The common practice may subject to the multiple testing problems, inconsistent parameter estimation and other issues. However, when we assume that the underlying probabilistic model is Gaussian, then it would be possible to perform likelihood based methods to estimate the concentration matrix, without even restoring the graph structure beforehand. When the graph is not complete, there are some entries in the concentrate matrix being zero. Lasso-like penalized method will detect these zero terms help reducing the number of parameters to be estimated, and also leads to the correct graph structure. Thus, in this approach model selection and parameter estimation are done at the same time, which is the main strength. The computation is eased by exploiting maxdet algorithm. BIC-type criterion is applied to the selection of tuning parameter in the penalized term.


## Introduction and Motivation

The main set up for the (undirected) graph structure learning problem is as follows: Assume $f(X_{1,2}, \cdots, X_p)$ is a multivariate distribution that satisfies Markov Property with respect to an unknown graph $G$, from which we observe $(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)})$ i.i.d observations from $f$, and our goal is to learn $G$ from $(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)})$.

Let's further assume that this $f$ follows a multivariate normal distribution $N_p(\mu, \Sigma)$, and the concentration matrix is denoted as $C = \Sigma^{-1}$. One of the main issue is to correctly detect the zero entries $c_{ij} = 0$ in the concentric matrix, since $c_{ij} \Leftrightarrow (X_i, X_j) \notin E \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_s$, where $E$ is the set of edges in $G$, and $X_s$ is the set of nodes except $X_i$ and $X_j$.

The three perspectives of Markov Property provides us with different approaches to the graph learning problem. In Gaussian case, Pairwise Markov Property, $(X_i, X_j) \notin E \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_s$, becomes $Cov(X_i, X_j | X_s) = \Leftrightarrow (X_i, X_j) \notin E$, where covariance is estimated by empirical conditional covariance. However, we need to do this for $\binom{p}{2}$ pairs of nodes. This leads to the multiple testing problems. Moreover, some of these tests are not independent. The second approach, using Local Markov Property, tries to determine the neighborhood set for a node by regressing all other nodes on this node, which converts it to the Lasso problem, and filter the zero terms effectively. However, in some case, it leads to the contradiction of $(X_i, X_j) \in E$ and $(X_j, X_i) \notin E$.

The third approach is to use Global Markov Property, which is the idea behind the method proposed by this paper. This likelihood based method employs an $l_1$ penalty on the off-diagonal entries $c_{ij}$ and thus leads to a sparse concentration matrix. Beside, the constraint of $C$ being positive definite is also satisfied. Moreover, a "non-negative garrote type" method is also introduced. The asymptotic properties and the relationship with other method can also be shown in detail. The computation of this method is turned into a *maxdet* problem, and an efficient algorithm is developed for the lasso type estimator. The simulation part compares two types of estimators with some other methods.

## Methods

### Lasso-type estimator

The MLE of this method is equivalent to the solution which minimizes

$$-log|C| + \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)' C (X_i - \mu)$$

However, we want some "sparsity" in the graph structure, i.e. filter out some of the near-zero entries so that trade-off some bias to obtain improved variance and thus achieve the lower MSE in the end. For this goal, we have to constrain the non-zero entries to be in some reasonable range. That is, for some $t$, to make the above term subject to $\sum_{i \neq j} |c_{ij}| \leq t$.

Using the Lagrangian form, along with rewriting the above term, we are actually minimizing

$$-log|C| + tr(C\bar{A}) + \lambda \sum_{i \neq j} |c_{ij}| \tag{1}$$

where $\bar{A}$ is the MLE of $\Sigma$ and $\lambda$ is a tuning parameter.

### Non-negative garrote-type estimator

Given an already valid estimator $\tilde{C}$ on $C$, usually just the MLE $\bar{A}$, we could also achieve the shrinkage by $c_{ij} = d_{ij}\tilde{c}_{ij}$, where a symmetric matrix $D = (d_{ij})$ is the solution which minimizes

$$-log|C| + tr(C\bar{A}) \quad \text{subject to} \quad \sum_{i \neq j} d_{ij} \leq t, d_{ij} \geq 0.$$

Similarly, using the Lagrangian form, along with some transformation, we are actually minimizing

$$-log|C| + tr(C\bar{A}) + \lambda \sum_{i \neq j} \frac{c_{ij}}{\tilde{c}_{ij}}, \quad \text{subject to} \quad c_{ij}/\tilde{c}_{ij} \geq 0 \tag{2}$$

and $C$ is positive definite.

### Illustration in a simple situation

Consider the situation where $p = 2$, where concentration matrix be 2 dimensional and we could easily derive the solution to (1) and (2) as below:

$$\hat{c}_{12}^{\text{lasso-type}} = \left( \frac{(1-r^2)[|r| - \lambda(1-r^2)]}{1 - [|r| - \lambda(1-r^2)]^2} \right)_+ sign(r) \tag{3}$$

$$\hat{c}_{12}^{\text{garrote-type}} = \left( \frac{(1-r^2)[|r| - \lambda(1-r^2)]}{|r| - [r^2 - \lambda(1-r^2)]^2/|r|} \right)_+ sign(r) \tag{4}$$

where $r$ is the MLE of $c_{ij}$.

The relationship between two types of estimator and MLE($r$) is shown on Figure 1. Both of them have the "cut-off" range which is determined by the tuning parameter, and the other range has an overall shrinkage that lowers the estimate down. However, as the MLE increases in absolute value, lasso-type estimator further departs the MLE, while the garrote estimator tends to get closer to the MLE.

We think the penalty on non-zero terms brings some bias to the model in trade of reduced variance, which is the effect of "cut-off" range. However, when it's outside this range, we may want it closer to the MLE so that the bias can be controlled. In this sense, garrote-type estimator should be more favorable, but the fact that it requires a initial estimator is somewhat restricting. We will further discuss this issue on the "future work" part.

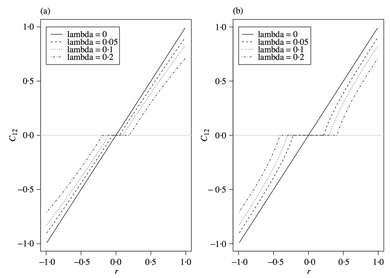


**Figure 1:** (a) Lasso-type estimator, (b) garrote-type estimator

## Future Improvements

### Strengths

The penalized-likelihood approaches brought up by this paper have two advantages over standard approach who use greedy stepwise regression for model selection and do parameter estimation based on the selected model. First, greedy stepwise regression is computationally more challenging than penalized-likelihood methods. Second, instead of determining neighborhood set for each $X_j$ locally through hypothesis testing at some level $\alpha$, which has long been recognized that this procedure does not correctly take account of the multiple comparisons involved, penalized-likelihood methods can globally determine graph structure avoid involving multiple dependent hypothesis tests. Many other methods like the greedy stepwise regression approach, model selection and parameter estimation are done seperately. However, penalized-likelihood estimators have precluded the discrete nature of such procedures which often leads to instability of the estimator. In most cases, since the estimators are based on maximum likelihood estimator the solutions of the penalized-likelihood estimators are continuous. Such continuity can ensure the stability of estimators.

Compared with the method proposed by Meinshausen & Bhlmann (2006), the paper's penalized-likelihood methods natrually incorporate the symmetry and positive-definiteness constraints in the estimation of the concentration matrix. In addition, since the loss function used by Meinshausen & Bhlmann (2006) is different from the quadratic approximation to the loglikelihood, the likelihood based approaches proposed by the paper are expected to be more efficient.

The paper proposed two type of penalized likelihood methods. One is LASSO-type method and the other is 'nonnegative garrote' type method. From the example provided in the second section of the paper, it seems garrote-type method has more advantages over LASSO-type method. As the MLE estimator's absolute value become larger and larger, approaching to 1, the garrote-type estimator tends to have relatively smaller penalty over the MLE estimator than LASSO estimator if both tuning parameter are well choosed. For those entries that MLE estimators are close to 0, the LASSO-type method tends to have smaller penalty than garrote-type estimator. Since the goal we have here is to achieve sparse structure and in the mean time provide more precise estimates as possible, if a good initial estimator is available, garrote-type estimator is more ideal than LASSO-type estimator. Also, the asymptotic properties for garrote-type estimator are better than those for LASSO-type estimator. The garrote-type estimator enjoys the so-called oracle property: it selects the right graph with probability tending to one and at the same time gives a root-n consistent estimator of the concetration matrix. From the simulation result provided in the six section of the paper, for garrote-type estimator, using MLE estimator as an initial estimator seems already good enough to compete with LASSO-type estimator.

### Weaknesses

An obvious weakness of the penalized-likelihood approaches in the paper is that they achieve sparsity by adding penalty term to the log-likelihood to get a shrinkage estimator for the concentration matrix globally. To achieve sparsity, we may want to add penalty to make those entries who close to 0 to be shrinked to 0. But for those entries who significantlly non-zero, we may want more accurate estimation of their true values. The shrinkage estimator simultaneously penalize all entries in the matrix at an almost same level, so for those entries who significantlly non-zero, penalized-likelihood estimator may lose much more accuracy than the MLE estimator.

Another weakness is that for garrote-type estimator, we need a good initial estimator. As shown in Figure 1, although garrote-type estimator penalize less when r far from 0, the distance between garrote-type estimator and MLE estimator is, in genral, larger than we expected.

### Possible future directions

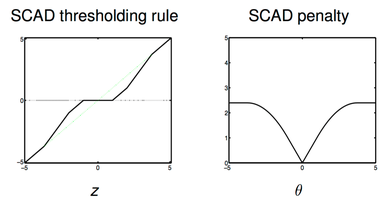


**Figure 2:** (a) Lasso-type estimator, (b) garrote-type estimator

As mentioned above, we do not want a shrinkage estimator simultaneously penalize all entries in the matrix at a same level. An ideal estimator we want could be that penalizes entries that are more close to 0 more severely than penalizing entries that are significantly non-zero. One straightforward idea would be that we use penalized-likelihood methods to find those 0 entries in the estimator and for other entries we use the MLE estimated values as our estimator. This idea is reasonable, but if we do this, our estimator would have a 'jump' respect to estimated values. For example, if we do this, it is very likely our estimated concentration matrix do not have entry value that falls between $(0, \alpha]$ ($\alpha$ is a value that close to 0) since all values smaller than $\alpha$ will be penalized to 0. Also, we have to find a reasonable thresholding for doing this. To avoid this situation, the idea of modifying LASSO penalty by SCAD penalty proposed by Jianqing Fan (1997) can be applied here.

$$P_{\tilde{\lambda}}^{scad}(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_{\tilde{\lambda},j}^{scad}(|\beta_j|), \quad \tilde{\lambda} = (\lambda, s),$$

A SCAD penalty in regression has properties similar to what we want here. From the Figure 2 we can find that SCAD penalty gives more sever penalty to larger coefficients than smaller coefficients in regression. We can actually borrow the idea from the SCAD penalty to define a SCAD-type penalty to relpace the LASSO-type penalty term used in the paper which penalize more on entries whose MLE estimator close to 0 and penalize little on significantlly non-zero entries.