Model Selection and Network Construction For High-Throughput Biological Data

Xin Gao

Department of Mathematics and Statistics,
York University,
Toronto, Ontario, Canada, M3J 1P3.

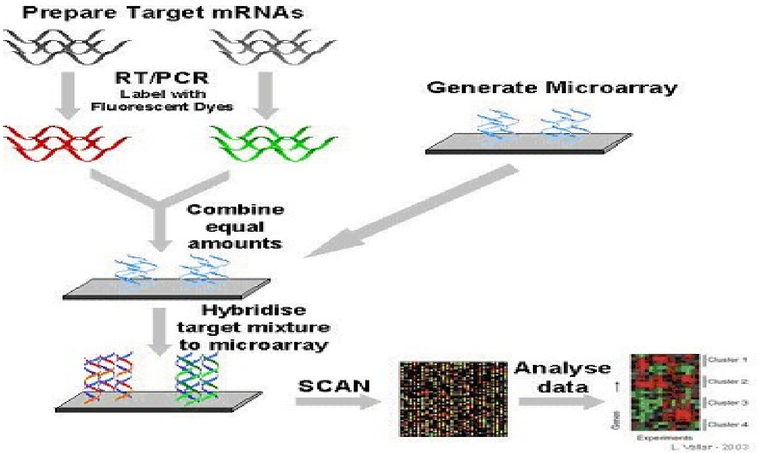
August 26, 2008

1 Model Selection via Penalized Likelihood Estimation

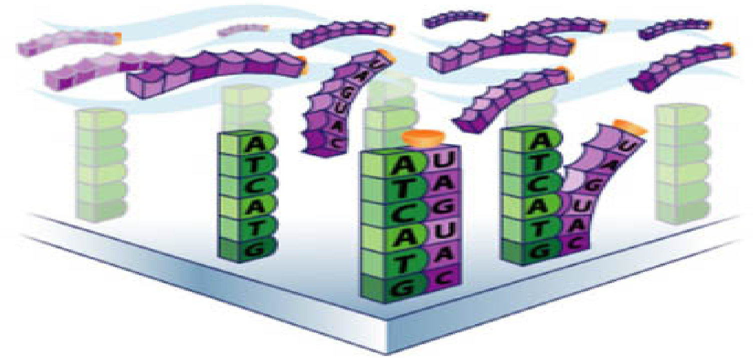
2 Model Selection when $P = O(N^k)$

3 Application in Chemogenomics

The high-throughput gene expression data

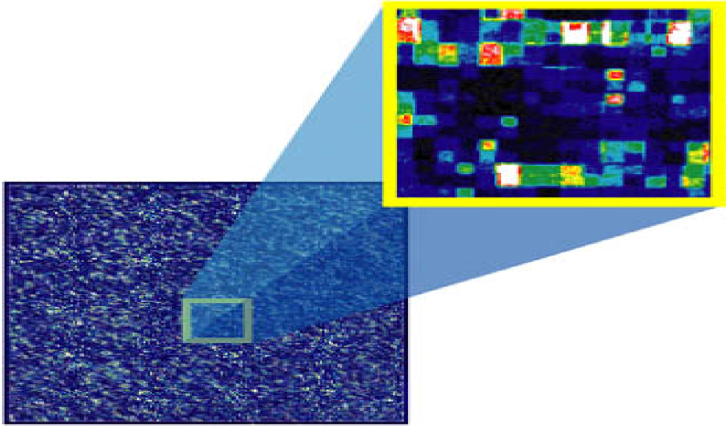


The chemical mechanism



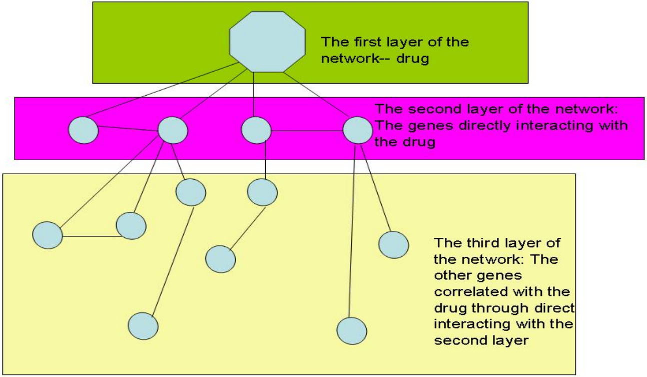
Sample RNA fragments (purple)
hybridized to DNA probe array (green)

The microarray image



Drug discovery and targeted genes

Architecture of drug and its associated genes


The first layer of the network-- drug
The second layer of the network: The genes directly interacting with the drug
The third layer of the network: The other genes correlated with the drug through direct interacting with the second layer

How to select the target and understand the network?

- In statistical modelling, a large number of predictors are usually introduced at the initial stage to attenuate modeling biases.
- To enhance predictability and obtain a parsimonious model, statisticians usually perform variable selection through stepwise deletion or subset selection.
- It is hard to understand the statistical properties of stagewise procedures as stochastic errors are inherited in every stage of the procedure.
- The best subset selection suffers the lack of stability (Breiman, 1996).

Simultaneous variable selection

- Using penalized likelihood method, the parameter estimation and variable selection can be done simultaneously.
- A good penalty function should possess the following properties:
  - Unbiasedness: When the true unknown parameter is large, the resulting estimator should be asymptotically unbiased to avoid unnecessary bias.
  - Sparsity: The estimator should be a threshold rule so that many small coefficients are automatically set to zero.
  - Continuity: The estimator should be continuous in data to avoid instability in model prediction.

Some insights on the requirements

- The minimizing criterion is: $Q(\theta) = -\sum_{i=1}^{n} \ell_i(\theta) + n\sum_{j=1}^{p} P_\lambda(|\theta_j|)$. In order to be asymptotically unbiased, $\lim_{n\to\infty} P'_\lambda(|\theta_{j,0}|) = 0$, for large $\theta_{j,0} \neq 0$.
- The penalty term should be heavy enough to shrink the small coefficients toward zero. (How heavy is enough? To address this question, study the proper order of the tuning parameter.)
- Consider the bridge regression with L_q penalty: $P_\lambda(|\theta|) = \lambda|\theta|^q$. The solution is continuous only when $q \geq 1$. On the other hand, when $q > 1$, (the penalty is not heavy enough) it does not produce a sparse solution. The L_1 penalty is the only member in this family which can produce both sparse and continuous solution.

Some commonly used penalty functions

- Hard thresholding:

$$P_\lambda(\theta) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda).$$

- L_1 penalty: (Donoho and Johnstone, 1994 and Tibshirani, 1996, 1997)

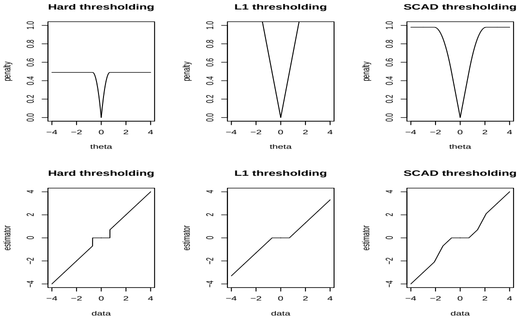
$$P_\lambda(\theta) = \lambda\theta.$$

- *SCAD* penalty: (Fan and Li, 2001)

$$p'_\lambda(\theta) = \lambda\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\}, \text{ for } \theta > 0,$$

where $(t)_+ = tI(t > 0)$ is the hinge loss function.

Comparisons of Different Penalty Functions



Advantages of SCAD penalty function

- It leads to nearly unbiased estimator when the true unknown parameter is large to avoid modelling bias.
- The resulting estimator is a thresholding rule, which leads to sparse estimators with small estimated coefficients automatically set to zero.
- The resulting estimator is continuous in data so that the procedure is stable in model selection.

Covariance selection in Gaussian graphical model

- Let $X = (X^{(1)}, ..., X^{(p)}) \sim N_p(\mu, \Sigma)$ with μ denoting the unknown mean and Σ denoting the nonsingular covariance matrix. We wish to estimate the concentration matrix $C = \Sigma^{-1}$.
- The zero entries c_{ij} in the concentration matrix indicates the conditional independence between the two random variables $X^{(i)}$ and $X^{(j)}$ given all other variables (Dempster, 1972, Whittaker, 1990, Lauritzen, 1996).
- The Gaussian random vector X can be represented by an undirected graph $G = (V, E)$, where V contains p vertices corresponding to the p coordinates and the edges $E = (e_{ij})_{1 \leq i < j \leq p}$ represent the conditional dependency relationships between variables $X^{(i)}$ and $X^{(j)}$.

Existing methods

It is of interest to identify the correct set of edges, and estimate the parameters in the concentration matrix simultaneously.

- MLE: In general, there would be no zero entries in the maximum likelihood estimate, which results in a full graphical structure.
- L_0-type penalty: (Dempster, 1972 and Edwards 2000) The L_0 penalty is discontinuous, the resulting penalized likelihood estimator is unstable.
- Stepwise forward selection or backward elimination of the edges: Ignores the stochastic errors inherited in the multiple stages of the procedure (Edwards, 2000). The computational complexity of this greedy search algorithm increases exponentially with the number of vertices.
- Neighborhood selection: (Meinshausen & Bühlmann, 2006) Neighborhood selection for each node and the results are

Our Approach

- We aim to develop a penalized likelihood estimation method for covariance selection which possesses the oracle property but also has simple asymptotic distribution to facilitate the derivation of the standard error estimates.
- With proper choices of the regularization parameter, the proposed SCAD type estimators possess an oracle property. Explicit expressions for the asymptotic covariance matrix of the estimators can be derived.
- We establish the consistency of the BIC criterion to select the true graphical structure under the penalized likelihood framework with the SCAD penalty and extend the result to more general penalized likelihood estimation setting.

Penalized Likelihood with SCAD Penalty

- Given a random sample $X_1, ..., X_n \sim N_p(\mu, \Sigma)$, the loglikelihood is

$$\ell(\mu, C) = \frac{n}{2}\log|C| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)' C (X_i - \mu).$$

- The MLE of (μ, Σ) is $(\bar{X}, \bar{A})$, where

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'.$$

- Assume that the observations are properly centered, then the sample mean is zero. The resulting SCAD estimator $\hat{C}$ should minimize the following objective function:

$$Q(C) = -\log|C| + \mathrm{tr}(C\bar{A}) + \sum_{i \neq j} p_\lambda(|c_{ij}|). \qquad (1)$$

Asymptotic properties

- **Lemma** Let $X_1, ..., X_n$ be independent and identically distributed random vectors following multivariate normal distribution $N_p(\mu, C_0^{-1})$. Let the objective function $Q(C)$ defined as in equation (1). If $\lambda \rightarrow 0$, as $n \rightarrow \infty$, then there exists a local minimizer $\hat{C}$ of $Q(C)$, such that $\left\|\hat{C} - C_0\right\| = O_p(n^{-\frac{1}{2}})$.

The sparsity and oracle properties

- Notation:$s = \{(i,j) : c_{ij,0} \neq 0,\ i \leq j\}$, $t = \{(i,j) : c_{ij,0} = 0,\ i \leq j\}$. Correspondingly, $C^{(1)} = \{c_{ij} : (i,j) \in s\}$ and $C^{(2)} = \{c_{ij} : (i,j) \in t\}$.
- **Theorem** If $\lambda \to 0$, and $\sqrt{n}\lambda \to \infty$, as $n \to \infty$, then with probability tending to 1, the root-n consistent local minimizer $\hat{C}$ in leads to the sparse estimator with $\hat{C}^{(2)} = 0$. Furthermore, the estimators of the nonzero partial correlations follows an asymptotic normal distribution $\sqrt{n}(\hat{C}^{(1)} - C_0^{(1)}) \to N(0, I_1^{-1}(C_0^{(1)}))$.
- For the true zero parameters, they are estimated as zero with probability tending to one. The nonzero components are estimated as well as when the correct model is known. The method outperforms MLE. (Super-efficiency!)

The size of tuning parameter

- It is worthy to note that the order of the regularization parameter is crucial.
- **Theorem** As $n \to \infty$, if $\lambda \to 0$, and $\sqrt{n}\lambda \to 0$, then the penalized likelihood estimator under the SCAD penalty has the same limiting distribution as the usual maximum likelihood estimator:

  $$\sqrt{n}(\hat{C} - C_0) \to argmin_{U=U'}(V),$$

  in distribution, where

  $$V(U) = tr(U\Sigma U\Sigma) + tr(UW),$$

  in which W is a random symmetric $p \times p$ matrix such as $vec(W) \sim N(0, \Lambda)$, and Λ is such that $cov(w_{ij}, w_{i'j'}) = cov(X^{(i)}X^{(j)}, X^{(i')}X^{(j')})$.

Selection of tuning parameter

- The performance of the proposed penalized likelihood method relies on the proper choice of the tuning parameter.
- For the regularization parameter λ, it is desirable to have a data-driven method to make the selection automatically.

Selection of graphical structure

- Define a full graphical model G_F with the full edge set $E_F = (e_{ij})_{1 \leq i < j \leq p}$.
- Define an arbitrary graphical model G with the corresponding edge set $E \subseteq E_F$.
- Define a true model G_T, with the edge set $E_T = (e_{ij})_{(i,j):c_{ij,0} \neq 0, i<j}$.
- Define an over-fitted model G if the corresponding edge set $E \supseteq E_T$ and $E \neq E_T$.
- Define an under-fitted model G with the edge set $E \not\supseteq E_T$.

Selection of graphical structure

- Giving a tuning parameter λ, the penalized likelihood approach yields the estimated parameters $(\hat{c}_{ij,\lambda})_{1 \le i \le j \le p}$.
- The resulting model is denoted as G_λ with the edge set $E_\lambda = (e_{ij})_{(i,j):\hat{c}_{ij,\lambda} \neq 0}$.
- We define $\Omega_- = \{\lambda \in \Omega : E_\lambda \not\subseteq E_T\}$, $\Omega_0 = \{\lambda \in \Omega : E_\lambda = E_T\}$, and $\Omega_+ = \{\lambda \in \Omega : E_\lambda \supseteq E_T \text{ and } E_\lambda \neq E_T\}$. The three subsets of Ω_0, Ω_-, Ω_+ lead to the true, under and over-fitted models, respectively.

Selection of graphical structure

- Given a λ, the associated BIC criterion is defined as:

$$BIC_\lambda = -\log|\hat{C}_\lambda| + \mathrm{tr}(\hat{C}_\lambda A) + \frac{log(n)}{n} \sum_{1 \le i < j \le p} I(\hat{c}_{ij,\lambda} \neq 0).$$

- On the other hand, suppose we know the correct model beforehand and perform the maximum likelihood estimation under the correct model, the associated BIC criterion is denoted as

$$BIC_{G_T} = -\log|\hat{C}_{G_T}| + \mathrm{tr}(\hat{C}_{G_T} \overline{A}) + \frac{log(n)}{n} \sum_{1 \le i < j \le p} I(c_{ij,0} \neq 0),$$

where $\hat{C}_{G_T} = (\hat{C}_{G_T}^{(1)}, 0)$, with $C^{(2)}$ knowing to be 0.

Selection of graphical structure

- Construct a sequence of reference tuning parameters $\lambda_n = log(n)/\sqrt{n}$, which satisfies the requirement that as $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$.
- Under such working sequence of tuning parameters, with probability tending to one, the resulting method will not only identify the correct set of true edges but also yield root-n consistent estimators for all the nonzero partial correlation coefficients. This guarantees the following result:
- **Lemma** $Pr(BIC_{\lambda_n} = BIC_{G_T}) \to 1$ as $n \to \infty$.

Consistency of BIC criterion

- Consider the under-fitted model which is essentially a misspecified model with at least one of the nonzero parameters being mistakenly set to zero.
- Given $\lambda \in \Omega_-$, let $C^{(a)}$ denote $(c_{ij})_{(i,j)\in E_\lambda}$, and let $C^{(b)}$ denote $(c_{ij})_{(i,j)\notin E_\lambda}$.
- The penalized likelihood $\hat{C}_\lambda = (\hat{C}_\lambda^{(a)}, 0)$ is the local minimizer of

$$Q_\lambda(C) = -L(C^{(a)}, 0) + \sum_{(i,j)\in E_\lambda} p_\lambda(|c_{ij}|). \qquad (2)$$

Consistency of BIC criterion

- Lemma If $\lambda \rightarrow 0$, as $n \rightarrow \infty$, then $Pr(inf_{\lambda \in \Omega_- \cup \Omega_+} BIC_\lambda > BIC_{\lambda_n}) \rightarrow 1$
- This Lemma implies that the λs that fail to identify the true model yields BIC always lower than λ_n. Consequently, the λ value which minimizes the BIC criterion will identify the true model. This establishes the consistency of the BIC criterion used under the penalized likelihood framework with the SCAD penalty.
- **Theorem** If $\lambda \rightarrow 0$, as $n \rightarrow \infty$, then $Pr(G_{\hat{\lambda}_{BIC}} = G_T) \rightarrow 1$, where $\hat{\lambda}_{BIC}$ is the regularization parameter that minimizes the BIC criterion.

Real data analysis

- We analyzed a biological data to construct the gene network among a set of nine genes. The cDNA expression levels of the nine genes were measured in 60 cancer cell lines. The data set was obtained from *http://discover.nci.nih.gov/datasetsNature2000.jsp*.
- The edges in the graphic model represent the dependency relationships among the genes, which is useful in elucidating the underlying regulatory network.
- The Lasso method detects the least number of edges while the neighborhood selection method detects the greatest number of edges. Our proposed SCAD method generates a network which is similar to that of nonnegative garrote and the neighborhood selection methods.
- We used the fivefold crossvalidation to compute the KL

Average KL loss estimated by fivefold cross validation on the gene expression data

Method	*Scad*	*Lasso*	*Garrote*	*MB*
KL distance	7.284	7.370	7.564	7.575

Model Selection in High Dimensional Space

- The discussion above is focused on the situation that P, the number of covariates is fixed, and N, the sample size $\rightarrow \infty$.
- With the advent of high-throughput biological data, scientists often encounter the problem of model selection with $P = O(N^k)$.
- The traditional approach becomes inadequate and the consistency of the traditional model selection criterion does not hold anymore.

Extended Bayesian Information Criterion

- Let $\{y_i : i = 1, \dots, n\}$ be independent observations. Suppose that the density function of y_i is $f(y_i|\theta)$, where $\theta \in \Theta \in R^P$, P being a positive integer. The likelihood function is denoted as $L(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$.
- Let s be a subset of $\{1, \dots, P\}$. Denote by $\theta(s)$ the parameter θ with those components outside s being set to 0 or some prespecified values.
- The BIC proposed by Schwarz (1978) selects the model that minimizes

$$BIC(s) = -2\log L(\theta(\hat{s})) + \nu(s)\log n,$$

where $\nu(s)$ is the number of components in s.

Extended Bayesian Information Criterion

- Note that the cardinality of the model space is 2^P, where $P \to \infty$. Let S be the model space and $p(s)$ be the prior probability of model s. Assume that, given s, the prior density of $\theta(s)$ is given by $\pi\{\theta(s)\}$.
- The posterior probability of s is obtained as

$$p(s|Y) = \frac{p(s) \int f(Y|\theta(s)\pi\{\theta(s)\}d\theta(s)}{\sum_{s \in S} p(s) \int f(Y|\theta(s)\pi\{\theta(s)\}d\theta(s)}$$

- We select the model which minimizes the posterior probability. Under the assumption of equal prior, and use Laplace assumption, it can be shown that such model minimizes the BIC.

The problem of the equal prior

- Example: $P = 1000$. the class of model S_1 containing one covariate has size 1000. The class of model S_2 containing two covariate has size 1000*999/2.
- Under the equal prior assumption, the probability assigned to S_2 is 999/2 times that assigned to S_1. Therefore, models with larger number of covariates are given higher probabilities. This is against the principle of parsimony.

The problem of the equal prior

- Chen and Chen (2008) proposed EBIC which assigns the prior as follows: The model space is partitioned into $\cup_{j=1}^{P} S_j$, with each S_j containing models with the same dimension. Let $\tau(S_j)$ be the size of S_j. For instance, if S_j is the collection of all models with j covariates, $\tau(S_j) = C_j^P$. For each s in the subspace S_j, assign an equal probability, i.e., $P(s|S_j) = 1/\tau(S_j)$. Then instead of assigning probabilities $P(S_j)$ proportional to $\tau(S_j)$, as the ordinary BIC, we assign probabilities $P(S_j)$ proportional to $\tau^{\xi}(S_j)$, for some ξ between 0 and 1.
- $$EBIC == -2\log L(\theta(\hat{s})) + \nu(s)\log n + 2\gamma\log\tau(S_j),$$

  with $\gamma = 1 - \xi$. The last term is an extra penalty term to penalize the dimension of a model.

Application

- In chemogenomics field, scientists often conducts two aspects of measurements. They obtain the activity level measurement of certain drugs in different cell lines. In the mean time, they obtain the gene expression profile of the cell lines.
- This type of data can be used to discover novel drug target and elucidate the pathways that the drug participate in.
- Large set of potential predictors (usually greater than 20,000), which cannot be tackled by traditional statistical methods.

Data Analysis

- Consider the data

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iP}\beta_P + \epsilon_i,$$

  where $p = 1, \ldots, P \approx 20000$ denotes the set of covariates, and β_p denotes the coefficients for the pth covariate, ϵ_i are iid distributed noises, $i = 1, \ldots, n$ denoting the replicates of the responses variable.
- Among the P covariates, only a small subset of covariates have nonzero coefficients.
- The first step of predictor screening can be based on some biological information. The top ranked total L covariates are retained for the next step of analysis.

Data Analysis

- Next step: tournament approach (Chen and Chen, 2008) proposed to partition the subset of covariates into M manageable blocks. For each block of subset, the LASSO regression is applied so that a fixed number Q of significant covariates are selected. Then the total of MQ selected covariates are combined to form the reduced covariate set, which will be fully examined by the model selection technique.
- The LASSO regression (Tibshirani, 1996) is essentially a penalized regression with L_1 penalty. The LASSO regression estimate the model by minimizing the following objective function:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{p=1}^{P} |\hat{\beta}_p|.$$

Data Analysis

- The least angel regression (Efron, Tibshirani, Friedman 2003) can realize the LASSO regression in a very computationally efficient way and also provides the full sequence of nested models obtained when λ increases from 0 to ∞.
- Given the sequence of the nested models, we utilize the extended BIC criterion (Chen and Chen 2008) to determine the best model.
- Using the EBIC criterion, we found those predictors that are directly interacting with the drug. From our point of view, that constructs the first layer of the network.

Data Analysis

- How about other predictors? Their correlation with the drug activity is through indirect interactions.
- As for each variable X_p, we can regress it on all the remaining covariates $X_1, \ldots, X_{p-1}, X_{p+1}, X_P$, and obtain the following model:

$$X_p = \theta_1^p X_1 + \cdots + \theta_{p-1}^p X_{p-1} + \theta_{p+1}^p X_{p+1} + \cdots + \theta_P^p X_P.$$

It is known that (Meinshausen & Buhlmann, 2006),

$$\theta_q^p = -\frac{C_{pq}}{C_{pp}},$$

where the concentration matrix $C = (C_{pq})$, $p, q = 1, \ldots, P$.

Data Analysis

- The EBIC criterion ensures that with probability tending to one, for each gene, we are selecting the correct set of neighborhood.
- Data: U133A NCI60 cell line. Overall there are a total of 60 cell lines but only 47 cell lines with the drug activity measurements. There are a total of 22283 genes being measured on the array and could be the potential predictors for the drug activity.

Data Analysis

- We selected the top 2000 genes with the largest expression ratio and smallest expression ratio between the two cell lines CNS.SNB19 versus CNS.U251.
- We then have a reduced set of 2000 predictors. We divide the whole set into 10 equal sized subset of 200 predictors each. On each of the block of 200 predictor, we run LASSO to obtain the most significant 10 predictors and pool them together to get the most significant 200 potential predictors.

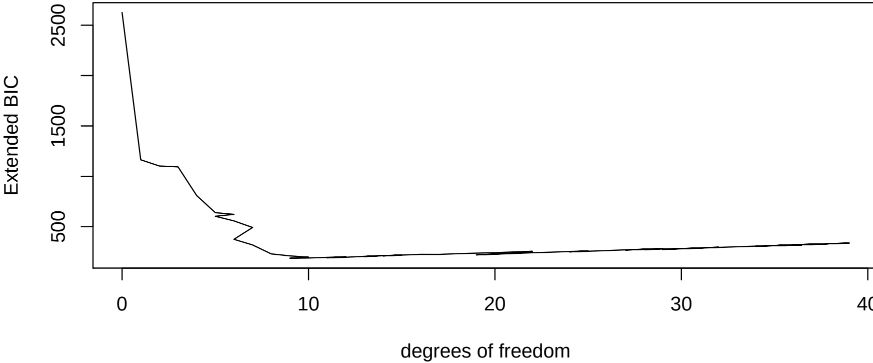
The sequence of predictors entering into the regression model

Step num	Row num	Affyprobe	gene name
1	8781	209287_s_at	CDC42EP3
2	2851	203324_s_at	CAV2
3	5041	205514_at	ZNF415
4	3500	203973_s_at	CEBPD
5	12807	213426_s_at	CAV2
6	18994	219630_at	NA)
7	11156	211756_at	PTHLH
8	1812	202284_s_at	CDKN1A
9	1904	202376_at	SERPINA3
10	19710	220346_at	LRRTM4
11	15077	215704_at	NA
12	1819	202291_s_at	MGP
13	12804	213423_x_at	TUSC3
14	1838	202310_s_at	COL1A1
15	15019	215646_s_at	CSPG2

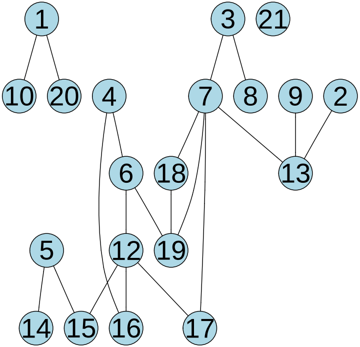
The sequence of nested model with the EBIC

Step	Df	Rss	EBIC
0	0	1.312957e+03	2625.9146
1	1	5.748809e+02	1164.0961
2	2	5.376123e+02	1102.4968
3	3	5.272411e+02	1093.8716
4	4	3.792328e+02	809.3863
5	5	2.887006e+02	639.3970
6	6	2.748882e+02	622.4723
7	5	2.707608e+02	603.5174
8	6	2.424549e+02	557.6057
9	7	2.045441e+02	492.1657
10	6	1.510993e+02	374.8946
11	7	1.178586e+02	318.7947
12	8	6.852696e+01	230.2356
13	9	5.369994e+01	210.4397
14	10	4.245803e+01	197.5930
15	9	4.183606e+01	186.7120
16	10	3.837976e+01	189.4364
17	11	3.674101e+01	195.5949
18	12	3.538412e+01	202.1324

The model selection curve



The subgraph of the third layer of the network. The total graph of the 200 predictors have 1575 edges.



Conclusion

- The asymptotic behavior of the model selection technique: consistency for both P fixed and $P \to \infty$ situations requires careful selection of tuning parameters and the choice of penalty term.
- Through model selection technique, we propose a systematic approach of constructing network structures for chemogenomic data.
- Static network versus dynamic network incorporating temporal information; Nonlinear modelling versus linear modelling; consistency of the overall graph versus the consistency of each local neighborhood