Network sparsity selection and robust estimation via bootstrap with applications to genomic data

José Sánchez

Bioinformatics CF, University of Gothenburg

September, 2015

Joint work with:

- A. Jauhiainen, Karolinska Institutet.
- R. Jörnsten , University of Gothenburg and Chalmers University of Technology.
- S. Nelander, IGP and SciLifeLab, Uppsala University.

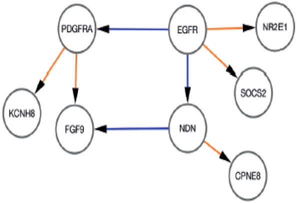
Outline

1. Introduction
2. Method
3. Simulation study
4. Future work

Network Modeling

Goals

- Construct regulatory networks and predictive models
- Identify disease-specific key regulators and their targets
- Relate the network structure to patient survival or clinical subtypes



- Data now available at multiple levels
- (Epi)Genetic variation: single-point mutations, copy number aberrations, loss of heterozygosity, methylation
- mRNA, miRNA
- clinical: age, survival,...

Estimation methods

Correlation based methods

Assume genes to be $N(\mu, \Sigma)$. The network given by the non-zeros of the thresholded empirical correlation matrix. Alternatively threshold a power of the empirical correlation matrix (Langfelder and Horvath, 2008).

Partial correlation based methods

Assume genes to be $N(\mu, \Sigma)$. The links for the gene network are given by the non-zeros of $\Theta = \Sigma^{-1}$ (Friedman et al., 2008).

Information theory based methods

Connectivity between genes is given by the mutual information. The links most pass a significance threshold and the data processing inequality (Margolin et al., 2006).

More on partial correlation based methdos

In general the number of genes is larger than the number of samples, $p >> N$, thus the empirical covariance matrix is not invertible. Assuming $X \sim N(\mu, \Sigma)$, maximize the L_1 penalized likelihood function for the precision matrix Θ

$$l(\Theta) = \ln\left[\det\left(\Theta\right)\right] - \operatorname{tr}\left(S\Theta\right) - \lambda\|\Theta\|_1$$

where $S = \frac{1}{N}X^TX$ is the empirical covariance matrix, $\|\Theta\|_1 = \sum_{i \neq j} |\theta_{ij}|$. The parameter $\lambda > 0$ controls the degree of sparsity in Θ.

More on partial correlation based methdos

In general the number of genes is larger than the number of samples, $p >> N$, thus the empirical covariance matrix is not invertible. Assuming $X \sim N(\mu, \Sigma)$, maximize the L_1 penalized likelihood function for the precision matrix Θ

$$l(\Theta) = \ln\left[\det\left(\Theta\right)\right] - \text{tr}\left(S\Theta\right) - \lambda\|\Theta\|_1$$

where $S = \frac{1}{N}X^T X$ is the empirical covariance matrix, $\|\Theta\|_1 = \sum_{i \neq j}|\theta_{ij}|$. The parameter $\lambda > 0$ controls the degree of sparsity in Θ.

Goal

Select sparsity level (λ).

What has been done using bootstrap?

Generate B bootstrap samples choosing randomly 90% of the samples and let $\hat{\theta}_{ij,b}$ be the b-th bootstrap estimate for link (i,j), then

$$h_{ij,\lambda} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left(\left|\hat{\theta}_{ij,b}\right|\right)$$

is an estimate of the probability of presence of link (i,j). Final sparsity and differentiality levels are selected thresholding n_{ij} by some $T > 0$.

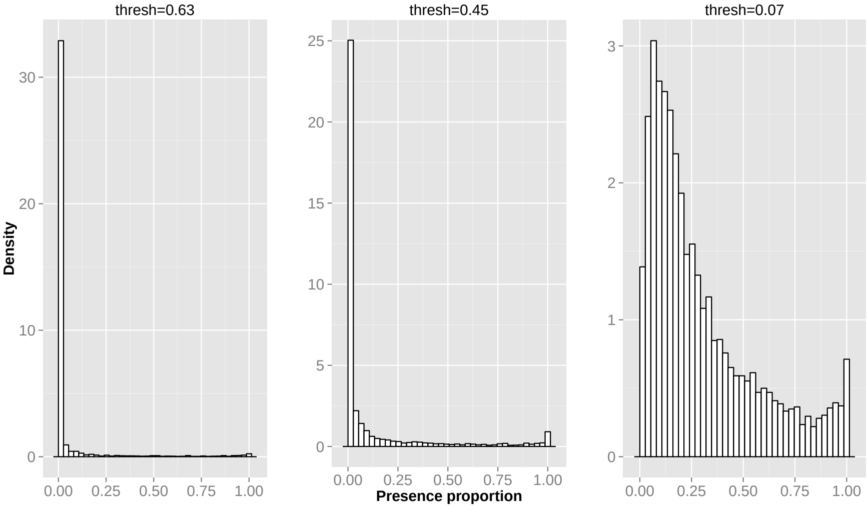
What has been done using bootstrap?

Generate B bootstrap samples choosing randomly 90% of the samples and let $\hat{\theta}_{ij,b}$ be the b-th bootstrap estimate for link (i,j), then

$$h_{ij,\lambda} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{I}\left(\left|\hat{\theta}_{ij,b}\right|\right)$$

is an estimate of the probability of presence of link (i,j). Final sparsity and differentiality levels are selected thresholding n_{ij} by some $T > 0$. A proposal (de Matos Simoes and Emmert-Streib, 2012) is to select T by data simulation from the null distribution (a random network), and test for the presence of each link.

Presence distribution



The Beta-Binomial mixture model

- Netwok estimates comprise two link populations: negatives (N) and positives (P).

The Beta-Binomial mixture model

- Netwok estimates comprise two link populations: negatives (N) and positives (P).
- Natural parameters of interest: FPR, which is the average rate N → FP; and TPR, the inverse of the average failure rate P → FN.

The Beta-Binomial mixture model

- Netwok estimates comprise two link populations: negatives (N) and positives (P).
- Natural parameters of interest: FPR, which is the average rate N → FP; and TPR, the inverse of the average failure rate P → FN.
- Let $p_0(i,j)$ be the edge-specific failure rate for N → FP, and $p_1(i,j)$ the rate for P → FN. We model $p_l(i,j) \sim$Beta(α_l, β_l).

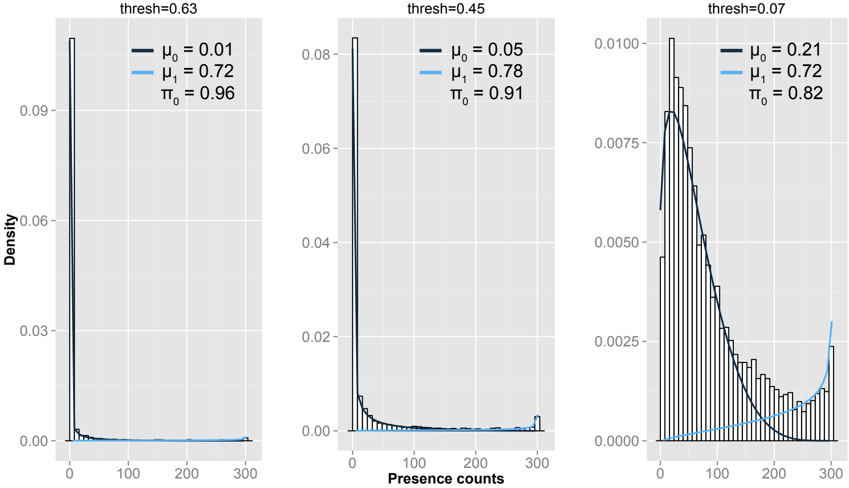
The Beta-Binomial mixture model

- Netwok estimates comprise two link populations: negatives (N) and positives (P).
- Natural parameters of interest: FPR, which is the average rate N → FP; and TPR, the inverse of the average failure rate P → FN.
- Let $p_0(i,j)$ be the edge-specific failure rate for N → FP, and $p_1(i,j)$ the rate for P → FN. We model $p_l(i,j) \sim$Beta(α_l, β_l).
- Let $x_{ij,\lambda} = \sum_{b=1}^{B} I\left(\left|\hat{\theta}_{ij,\lambda}^{b}\right|\right)$, then $x_{ij,\lambda}$ is an observation of $X_{ij,\lambda}$ with the following distribution

$$f_{X_{ij,\lambda}}(k) = \pi_0 \binom{B}{k} \frac{\mathrm{Be}(k+\alpha_0, k+\beta_0)}{\mathrm{Be}(\alpha_0,\beta_0)} + (1-\pi_0)\binom{B-k}{k} \frac{\mathrm{Be}(B-k+\alpha_1, B-k+\beta_1)}{\mathrm{Be}(\alpha_1,\beta_1)}$$

where Be is the Beta function defined as $\mathrm{Be}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}$.

Presence distribution estimated by EM



Sparsity selection

- From the Beta-Binomial model we estimate the average false positive rate as $\hat{\mu}_0 = \hat{\mu}_0(\lambda) = \frac{\alpha_0}{\alpha_0+\beta_0}$.
- We propose to select the sparsity level λ that keeps the FPR below some predetermined value t ($t = 0.05$, for example). The optimal estimated network size, N^*, becomes the one that corresponds to $\lambda^* = \max\{\lambda : \hat{\mu}_0(\lambda) < t\}$.

Final network estimate

- Let Δ_{ij} be the unobserved true class (P eller N) of edge (i, j), $\gamma_{ij}(\alpha_0, \beta_0, \alpha_1, \beta_1, \pi_0) = E(\Delta_{ij} | \alpha_0, \beta_0, \alpha_1, \beta_1, \pi_0; X_{ij,\lambda})$ and let $\hat{\alpha}_0$, $\hat{\beta}_0$, and $\hat{\pi}_0$ be the Beta-Binomial parameter estimates corresponding N^*.
- The E-Step of the Beta-Binomial mixture model we get

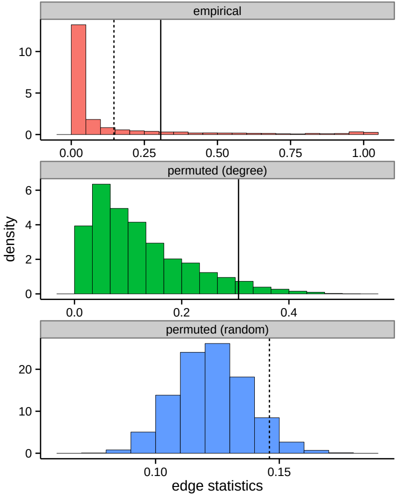
  $$\hat{\gamma}_{ij} = \frac{\hat{\pi}_0 \text{BetaBin}(x_{ij,\lambda}; \hat{\alpha}_0, \hat{\beta}_0)}{\hat{\pi}_0 \text{BetaBin}(x_{ij,\lambda}; \hat{\alpha}_0, \hat{\beta}_0) + (1 - \hat{\pi}_0)\text{BetaBin}(x_{ij,\lambda}; \hat{\alpha}_1, \hat{\beta}_1)}$$

  as a class expected value of edge (i, j).
- The final network is thus constructed by removing edges where $\hat{\gamma}_{ij} < 0.5$.

Final network estimate

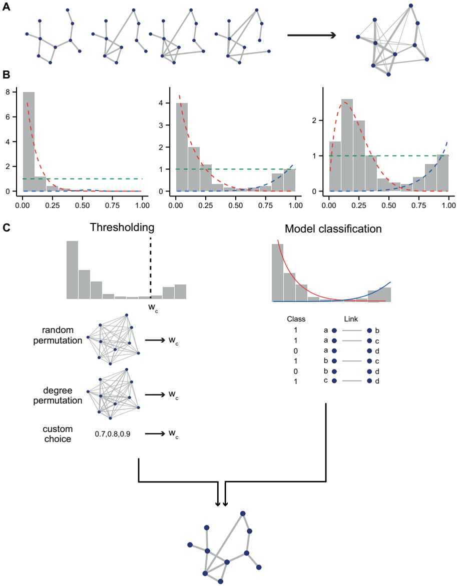
We present also two methods based on thresholding the aggregated boostrap networks.

- Edge presence for estimated networks, requires validation.
- Node-degree preserved permuted networks, more principled method.
- Permuted network, most generous.



Flow chart

- A Bootstrap estimation
- B Model fitting
- C Robust estimate construction



Data generation

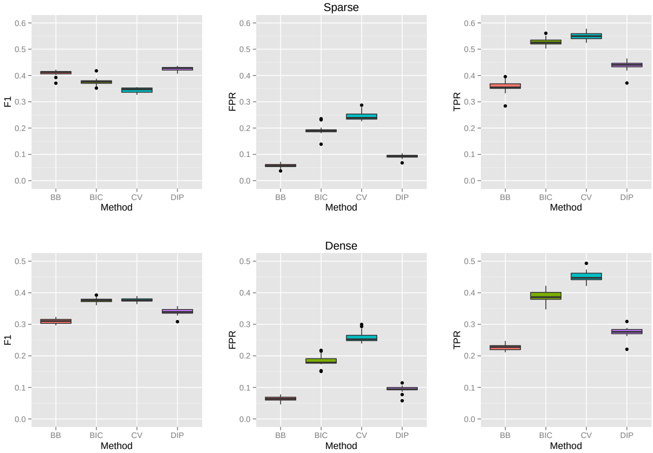
- Gene expression data set for 509 glioblastoma tumors and 10321 genes from The Cancer Genome Atlas.
- Select 100 genes randomly among the most strongly connected genes.
- Using glasso we construct two *true* networks: a *sparse* one with about 13% non-zeros and a *dense* one with about 21% non-zeros.
- Data is generated as realizations of multivariate normal distributions whose covariance matrices are given by the inverses of the true networks.

Sparsity selection methods

We compare our method, BB, with the following

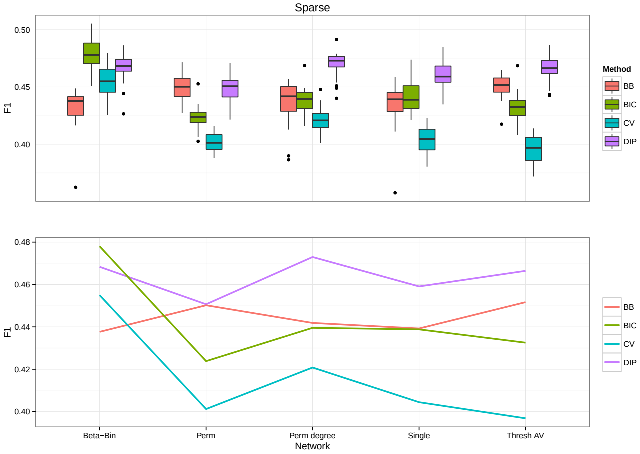
- The DIP statistic (Hartigan and Hartigan, 1985), defined as $\mathrm{D}(F) = \sup_x |F(x) - U(x)|$ for F any distribution function and U the uniform distribution.
- Bayesian information criterion, BIC, defined as $\mathrm{BIC}(\hat{\Theta}) = n \left[\ln |\hat{\Theta}| + \mathrm{tr}(S\hat{\Theta}) \right] + \ln(n) \sum_{i<j} \mathrm{I}(|\theta_{ij}|)$.
- Cross-validation.

Network size selection performance



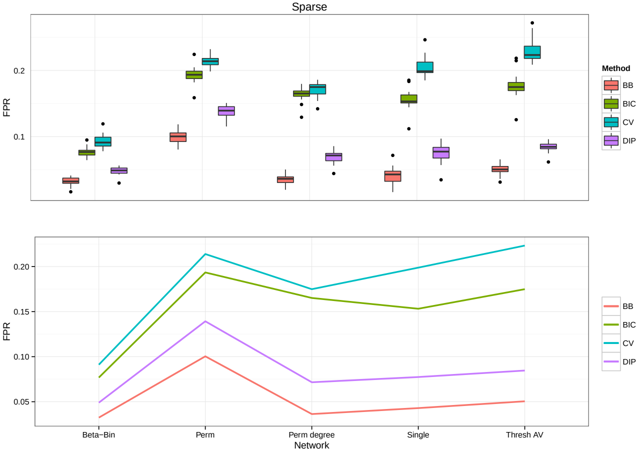
Beta-Binomial and DIP have the best FPR control, additionally, they have the best agreement with the true network in the sparse case.

Robust estimate performance



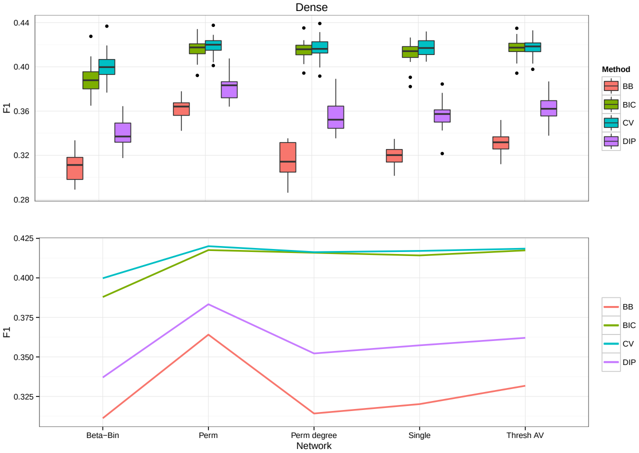
Beta-Binomial and DIP have better F1 measure across all network construction methods.

Robust estimate performance



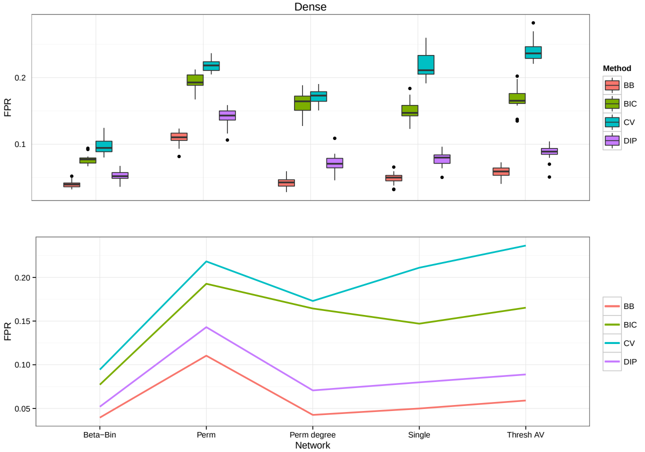
Beta-Binomial and DIP control the FPR at low levels.

Robust estimate performance



Beta-Binomial and DIP achieve the best F1 for permuted networks, which are in general denser.

Robust estimate performance

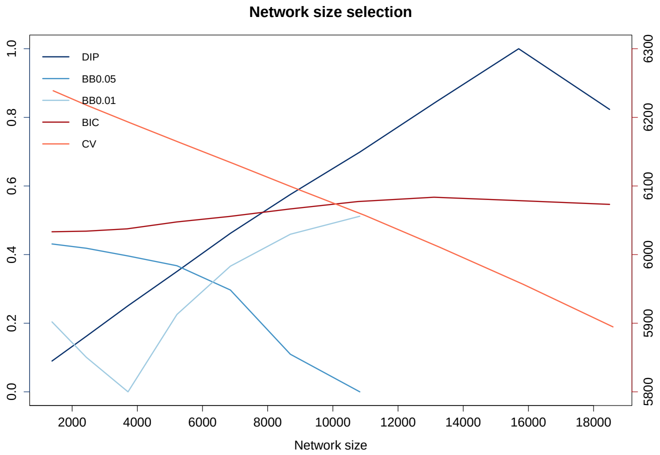


Beta-Binomial and DIP control the FPR at low levels.

Application to cancer genomic data

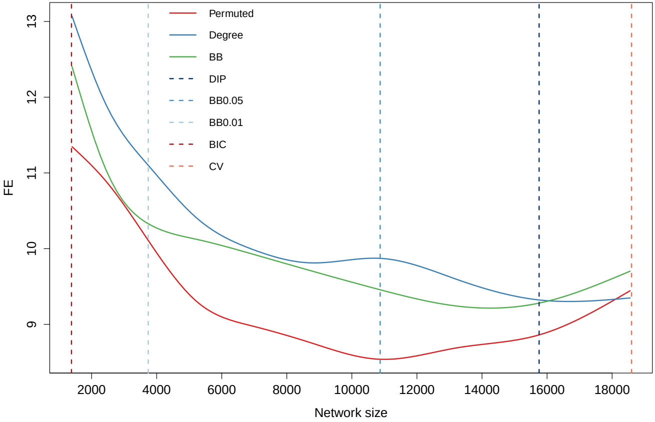
- Expression data from The Cancer Genome Atlas from 266 ovarian cancer tumors comprising 20500 genes.
- Remove genes without variability and keep pairs whose correlation is above 0.7, resulting in 5600 genes.
- Estimate the bootstrap network with glasso. Due to the screening process we only need to apply high treshold values, thus being able to estimate only smaller connected components (Danaher et al., 2014).
- We discard the smallest components, finally focusing on approximately 2000 genes.

Network size selection



Beta-Binomial and DIP select relatively small networks while CV and BIC seem to fail.

Networks overlap with biological pathways



Best FE are achieved for smaller networks with Beta-Binomial and degree preserved permutation post-processing.

Future work

We introduce a framework that uses bootstrap in a principled way to perform network sparsity selection and robust estimate construction. Through simulations we show that, given network estimation method, sample size, dimensionality and signal to noise ratio, sparsity selection should be guided by FP control rather than matching the (unknown) true network size. Our method successfully controls FPR in different sparsity settings and outperforms CV and BIC.

- An extension to the group and fused graphical lasso is ongoing, some preliminary results are available.
- Extension to local sparsity and group/fusing selection. Edge independency assumption can be violated here, this is a limitation our method we plan to explore.
- Computational burden can be reduce by letting the FPR guide the sparsity level for the bootstrap networks (line search).
- Our method is not restricted to FPR control. Following (Li et al., 2013), other accuracy measures such as FDR can guide sparsity selection. As future work we will compare with this method and with other accuracy measures.
- The network post-processing proposed here, Beta-Binomial, controls FPR. Control of FP inclusion can be improved by meand some other measure such as FDR.

P. Danaher, P. Wang, and D.M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

R. de Matos Simoes and F. Emmert-Streib. Bagging statistical network inference from large-scale gene expression data. *PLoS ONE*, 7(3):6e33624, 2012.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

John A Hartigan and PM Hartigan. The dip test of unimodality. *The Annals of Statistics*, pages 70–84, 1985.

P. Langfelder and S. Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(559), 2008.

S. Li, L. Hsu, J. Peng, and P. Wang. Bootstrap inference for network construction with an application to a breast cancer microarray study. *The annals of applied statistics*, 7 (1):391, 2013.

AA. Margolin, I. Nemenman, K. Basso, C. Wiggins, and et al. Stolovitzky, G. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(Supl. 1), 2006.

BINCO

- Collect frequency statistics from bootstrap.
- Determine if the count distribution is U-shaped and find $(V_1, V2]$, the interval where the distribution is monotonically decreasing.
- Obtain the null density estimate by
$$\left(\hat{\pi}, \hat{\theta}\right) = \text{argmin}_{\pi,\theta} L\left(f^\lambda, (1-\pi)h_\theta\right),$$
where h_θ is the Beta-Binomial distribution, L denotes the Kullback-Leibler distance in $(V_1, V_2]$.
- Compute the FDR for all models S_c^λ by
$$\text{FDR}(S_c^\lambda) = \frac{\sum_{x\geq c}(1-\pi)f_0^\lambda(x)}{\sum_{x\geq c} f^\lambda(x)},$$
where f^λ is the empirical estimate of the edge frequency distribution.
- Obtain $c^*(\lambda) = \min\{c : \text{FDR}(S_c^\lambda) \leq \alpha\}$ and $\hat{N}_E(S_c^\lambda) = |S_c^\lambda|(1 - \text{FDR}(S_c^\lambda))$.
- The optimal regularization parameter is given by $\lambda^* = \text{argmax}_\lambda \hat{N}_E(S^\lambda_{c^*(\lambda)})$.