A Practical Approach to Inferring Large Graphical Models from Sparse Microarray Data

Juliane Schäfer

Department of Statistics, University of Munich

Workshop: Practical Analysis of Gene Expression Data

Munich, 16-19 February 2004

Institut
für
Statistik
münchen

Acknowledgments

Coauthor:

Korbinian Strimmer
Department of Statistics, University of Munich, Germany

Thanks for discussion and comments:

Stefan Pilz, Leonhard Held, Jeff Gentry

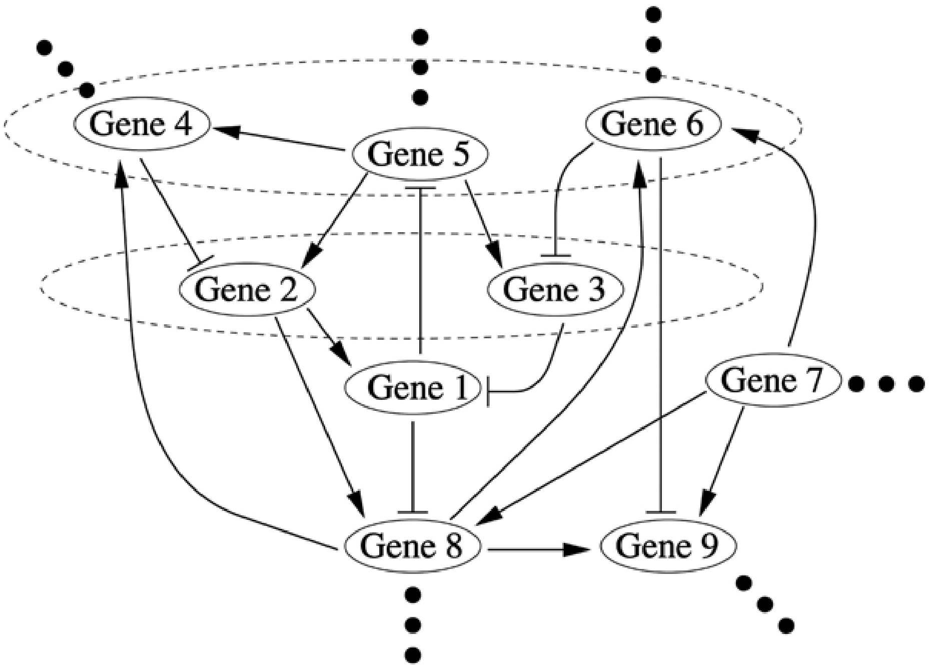
We thank the Deutsche Forschungsgemeinschaft (DFG) for financial support.

Contents

1. Motivation: Gene regulatory networks
2. Graphical Gaussian models
3. Coping with problems arising in application to microarray data
4. Simulation study to assess statistical properties of proposed procedures
5. Application to biological data
6. Discussion

Motivation: Gene regulatory networks

Cellular processes lead to complex dependency structure in gene expressions



Microarray experiment

Central dogma: $\boxed{\text{DNA}} \xrightarrow{\text{transcription}} \boxed{\text{mRNA}} \xrightarrow{\text{translation}} \boxed{\text{protein}}$

- explore transcript abundance, taken as a proxy for gene expression
- hybridization properties
- gene expression profile data: measurements under different conditions (certain points in time, treatments, tissues, etc.)

Reverse engineering problem

- Given a set of measurements (=multiple time series data), what can we deduce about the underlying network structure?

In particular:

Dimensionality problem: data feature space $>>$ sample size

- Challenging problem whose tractability is controversially discussed (e.g. Friedman et al. (2000) were the first to propose the use of Bayesian networks)

- What can we expect from available microarray data?

Graphical models

- Graphical models provide appropriate statistical framework:
  - association structure between multiple interacting quantities
  - distinguish between direct and indirect correlations
  - visualization in graph $G = (V, E)$
  - concept of conditional independence

- There are many different graphical models:
  - undirected vs. directed models
  - dynamic vs. static models

Some Definitions

Sample covariance matrix (with empirical mean $\hat{\mu}_i = \overline{y}_{\cdot i} = \frac{1}{N}\sum_{k=1}^{N} y_{ki}$)

$$\hat{\sigma}_{ij} = s_{ij} = \frac{1}{N}\sum_{k=1}^{N}(y_{ki} - \overline{y}_{\cdot i})(y_{kj} - \overline{y}_{\cdot j}) \quad (1 \leq i, j \leq G)$$

Empirical correlation coefficient matrix according to Bravais-Pearson

$$\hat{\rho}_{ij} = r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \quad (1 \leq i, j \leq G)$$

Genetic Correlations

Possible reasons for high pairwise correlation coefficient:

- direct interaction
- indirect interaction
- regulation by common gene

Not accounting for intermediates can lead to considerably biased conclusions (pseudo correlations, hidden correlations)!

We are mainly interested in direct interactions.

Graphical Gaussian models

We focus in this talk on a very simple class of graphical models:
Undirected graphical Gaussian models (Dempster, 1972; Whittaker, 1990)

- Starting point:
  - correlation structure, neither direction nor causality
  - multivariate Normal distribution with parameters μ and Σ assumed

- Based on the following:
  - Conditional distribution of genes i and j, given all the rest of the genes, is bivariate normal
  - Partial correlations as opposed to simple correlations

Graphical Gaussian models: Technical Details

- Partial correlations $\Pi = (\pi_{ij})$ are computed from the inverse of the $(G \times G)$ correlation matrix $(\omega_{ij}) = \Omega = P^{-1}$, with $P = (\rho_{ij})$

- the following are equivalent

  1. $\omega_{ij} = 0$
  2. genes i and j conditionally independent given the remainder of the genes
  3. partial correlation coefficient $\pi_{ij} = \rho_{ij|\text{rest}} = \dfrac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} = 0$

- Significance tests based on deviance difference between successive models (i. e. large sample tests based on limiting χ^2 distribution)

Problems arising in application to microarray data

- unstable partial correlation estimators for $G > N$
- multicollinearity: (nearly) linear dependencies in the data
- model selection: N is small, hence needs to be based on exact tests

$\hookrightarrow$ Application of GGMs so far restricted to assess relationships between small number of genes (Waddell & Kishino, 2000) or clusters of genes (Toh & Horimoto, 2002)

$\hookrightarrow$ Problem of interpretability

Small sample GGM framework needed!

Trick 1: Use pseudoinverse to invert correlation matrix

- failure of standard definition for inverse of a matrix for singular matrices
- generalization using singular value decomposition: $A = U\,\Sigma\,V^T$
- Pseudoinverse (Moore Penrose inverse): $A^+ = V\,(\Sigma^T\Sigma)^{-1}\,\Sigma\,U^T$
- $\sum(A^+A - I)^2$ minimized

This allows for computing partial correlations for $N < G$.

Trick 2: Use Bagging (Bootstrap aggregation)

General algorithm to improve estimates (Breiman 1996):

Step 1 Generate bootstrap sample y^{*b} with replacement from original data. Repeat process $b = 1, \ldots, B$ times idependently (e. g. $B = 1000$).

Step 2 Calculate for each bootstrap sample y^{*b} estimate $\hat{\theta}^{*b}$.

Step 3 Compute bootstrap mean

$$\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b}$$

Small Sample Estimates of Partial Correlation

1. $\hat{\Pi}^1$: use pseudoinverse for inverting $\hat{P}$ but do not perform bagging ($=$ observed partial correlation).

2. $\hat{\Pi}^2$: use bagging to estimate correlation matrix P, then invert with pseudoinverse ($=$ partial bagged correlation).

3. $\hat{\Pi}^3$: use bagging on estimate $\hat{\Pi}^1$, i. e. use pseudoinverse for inverting each bootstrap replicate estimate $\hat{P}^{*b}$ ($=$ bagged partial correlation).

Simulation study

To assess the statistical properties of the proposed procedures we need to perform a simulation study:

1. Generate random artificial network, i. e. true matrix of partial correlations Π
2. Compute corresponding matrix of correlations P
3. Simulate data from respective multivariate Normal distribution (with zero mean and variance one)
4. Estimate partial correlations $\hat{\Pi}^i$ from simulated data

Trick 3: Generating GGMs

Problem: true P must be positive definite, thus completely randomly chosen partial correlations do not necessarily correspond to valid graphical Gaussian model.

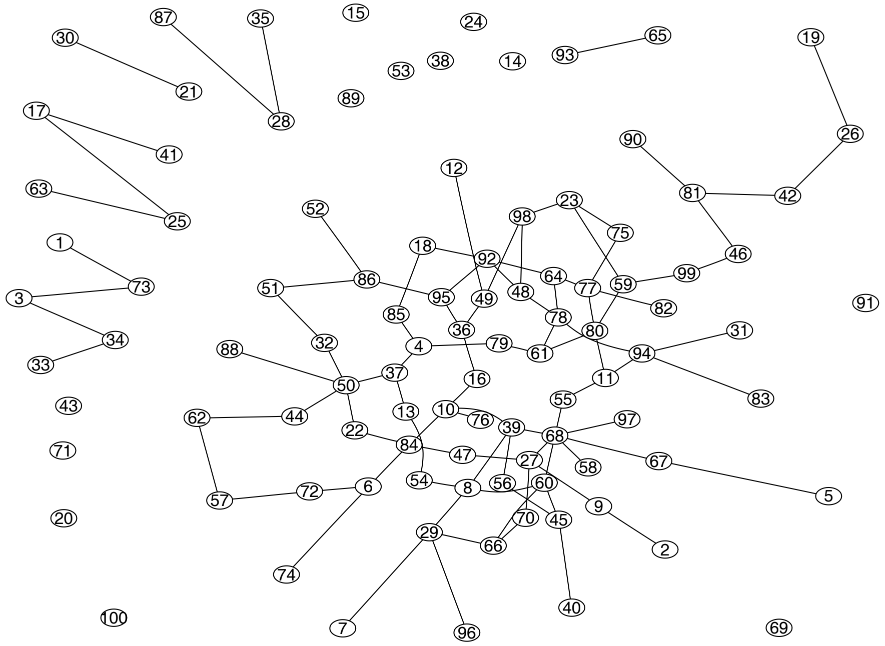
Solution:

1. generate random diagonally dominant matrix

2. standardize to obtain partial correlation matrix Π

$\longrightarrow$ resulting model is guaranteed to be valid

Random network with 100 nodes and edge fraction 0.02



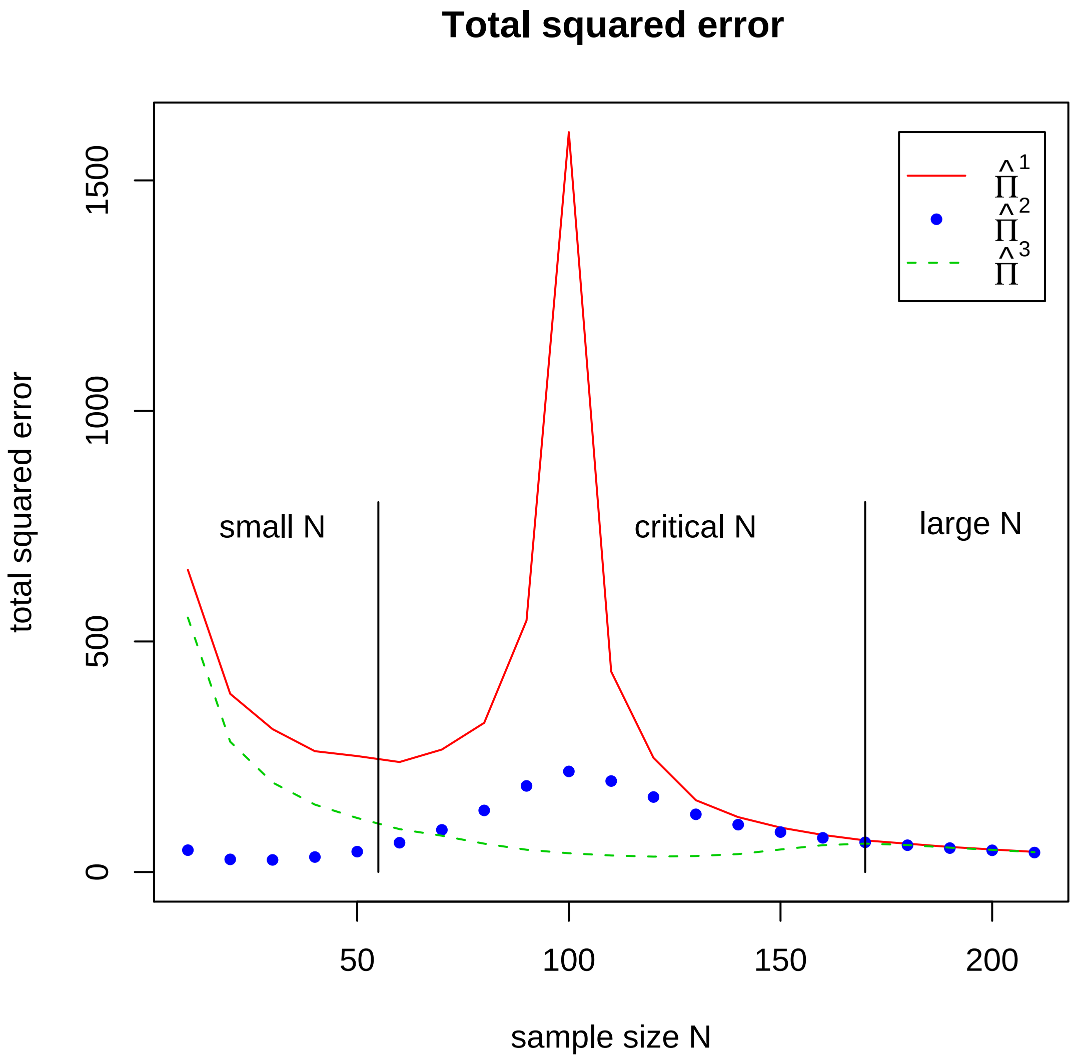
Evaluation of empirical mean squared error

$$\sum_{1\leq i<j\leq G}(\hat{\pi}_{ij}^{k} - \pi_{ij})^2 \quad (k = 1, 2, 3)$$

Example simulation setup:

- 100 nodes
- 2% non-zero partial correlations (biological networks are known to be sparse)

  $\hookrightarrow$ 99 true edges out of 4950 potential edges
- 1000 bootstrap replicates
- 50 simulation runs/sample size


Total squared error
total squared error
0
500
1000
1500
sample size N
50
100
150
200
small N
critical N
large N
$\hat{\Pi}^1$
$\hat{\Pi}^2$
$\hat{\Pi}^3$

Peaking phenomenon

- From a statistical point of view: VERY surprising!
- estimates expected to improve with increasing sample size

But:

- well known in small-sample regression and classification problems (Raudys & Duin, 1998; Skurichina & Duin, 2002)

Comparison of Point Estimates

- extremely bad performance of observed partial correlation $\hat{\Pi}^1$ in critical region (sample size $N \approx$ feature size G)

- Partial bagged correlation $\hat{\Pi}^2$ performs well for very small sample sizes (reason: bagged *sample* correlation matrix positive definite)

- Bagged partial correlation estimate $\hat{\Pi}^3$ best in critical region $N \approx G$

- the three methods coincide for $N >> G$ (note that this is where classical GGM theory applies)

Model selection

Determination of network topology

- try all potentially adequate graphical models and evaluate their goodness of fit
  $\hookrightarrow$ impossible in realistic applications due to enormous effort

- textbook methods (e. g. stepwise selection based on significance tests that are asymptotic χ^2-tests based on the deviance difference between successive models) are unreliable for small sample sizes

Alternative strategy used here:
multiple testing of all possible edges using exact correlation test

Null Distribution

Density under null hypothesis, i. e. $\rho = 0$, of Normal (partial) correlation coefficient (Hotelling 1953):

$$f_0(r) = (1 - r^2)^{(\kappa - r)/2} \frac{\Gamma(\frac{\kappa}{2})}{\pi^{\frac{1}{2}} \Gamma(\frac{\kappa - 1}{2})} \tag{1}$$

where κ is the degree of freedom.

For $\rho = 0$ the degree of freedom is equal to the inverse of the variance, i.e. Var$(r) = \frac{1}{\kappa}$, and to sample size minus one $(\kappa = N - 1)$.

For partial correlations: $\kappa = N - 1 - (G - 2) = N - G + 1$.

Negative for $N < G$!!!

Model Validation

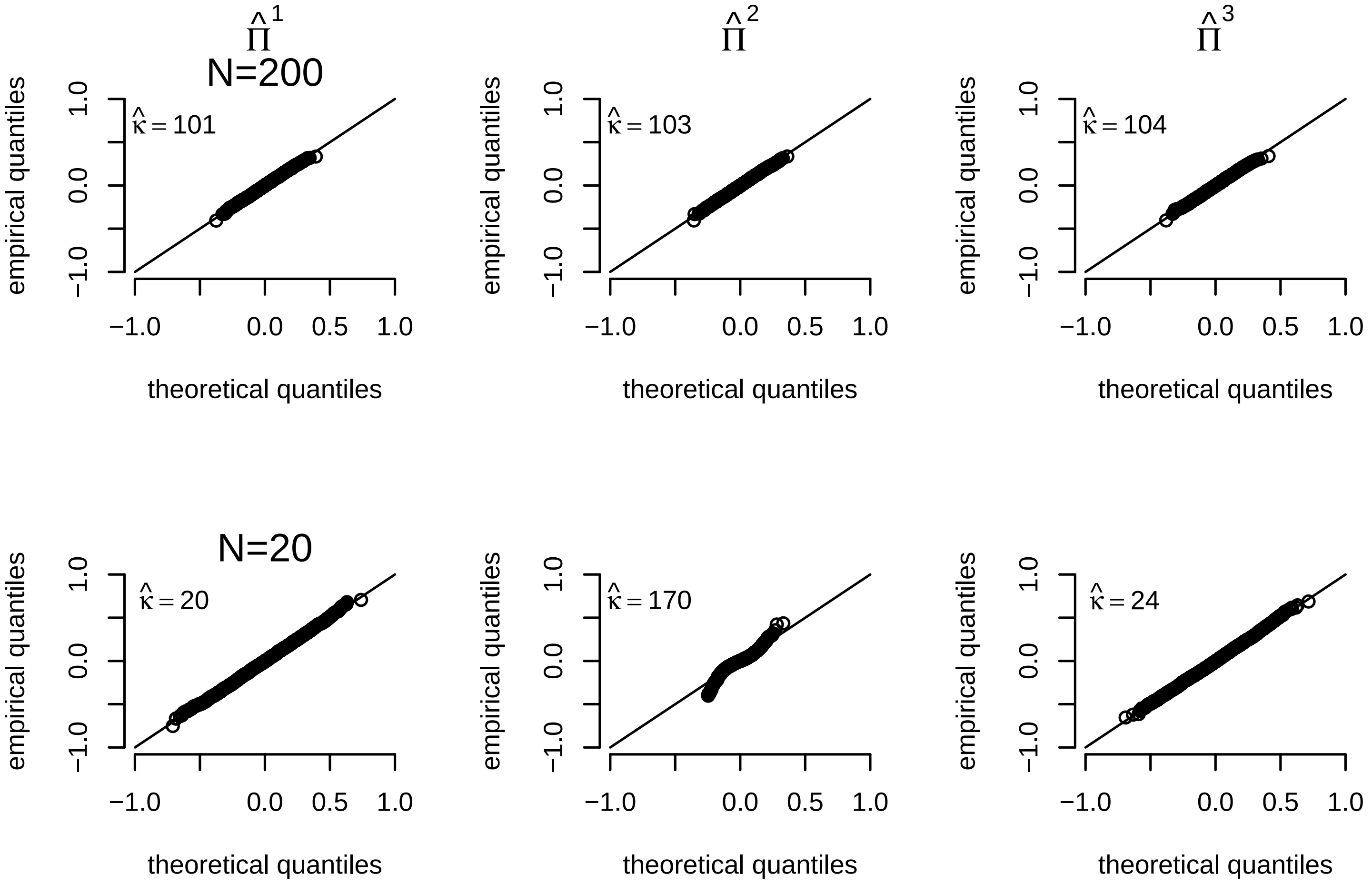
Do small sample estimates $\hat{\pi}_{ij}^1, \hat{\pi}_{ij}^2$, and $\hat{\pi}_{ij}^3$ of partial correlations under H_0 indeed follow this distribution?

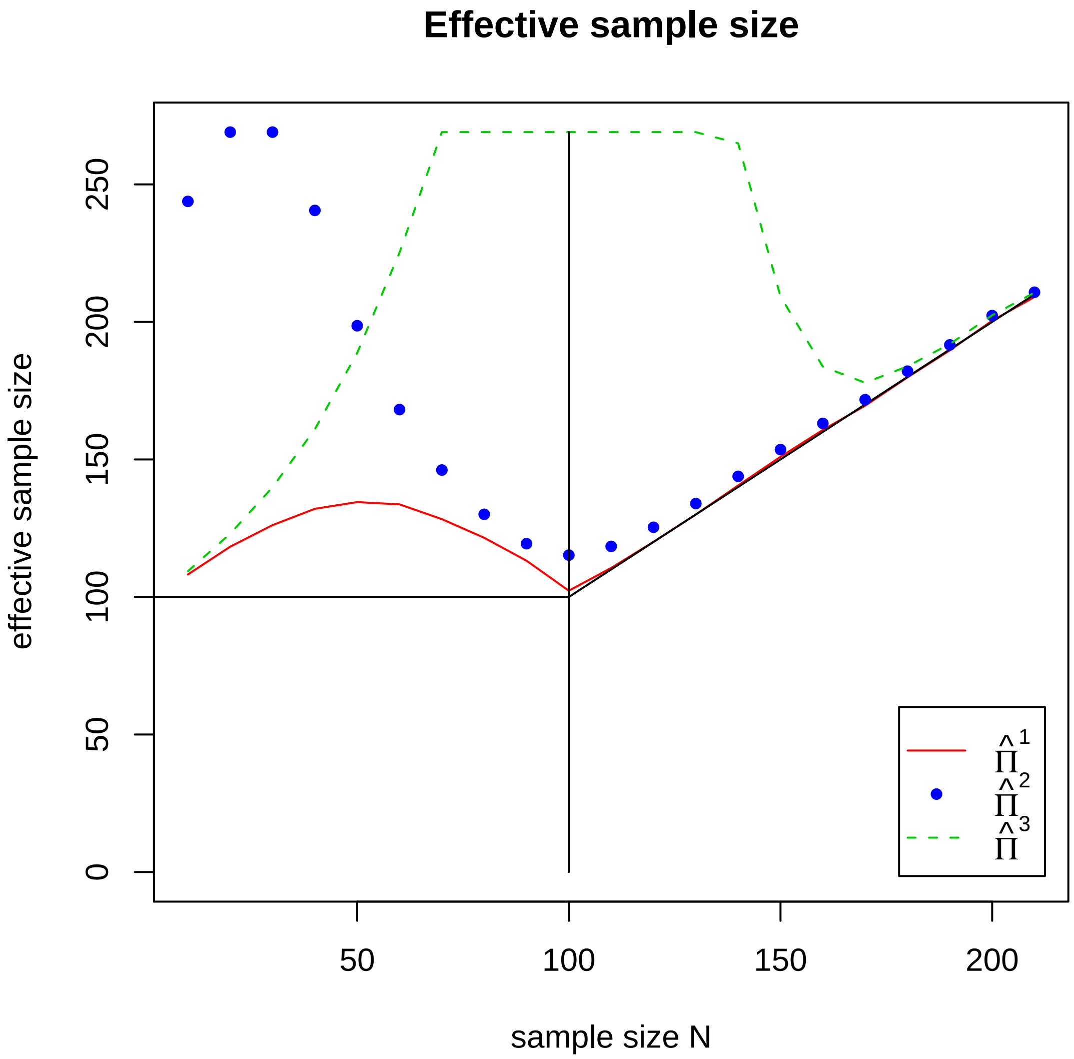
Trick 4: Estimate degree of freedom κ adaptively (details later).

Next two slides:

- QQ plots of all three point estimates for large (N=200, top row) and small (N=20, bottom row) sample size. Data simulated assuming $G = 100$ and no edges at all in underlying graph.

- plot of effective sample size $N_{\text{eff}} = \hat{\kappa} + G - 1$


$\hat{\Pi}^1$
$\hat{\Pi}^2$
$\hat{\Pi}^3$
N=200
N=20
$\hat{\kappa} = 101$
$\hat{\kappa} = 103$
$\hat{\kappa} = 104$
$\hat{\kappa} = 20$
$\hat{\kappa} = 170$
$\hat{\kappa} = 24$
empirical quantiles
theoretical quantiles
−1.0
0.0
0.5
1.0



Effective sample size

Results: Fit of Null-Model

- Empirical null distributions of estimates $\hat{\Pi}^i$ agree to a high degree with the theoretical distribution for the normal sample correlation.

- Estimated variance, degree of freedom and effective sample size differ among estimators and investigated region $(N << G, N \approx G, N >> G)$.

- Small total mean squared error and large effective sample size coincide

Inference of Edges

Trick 5: Exploit highly parallel structure of the problem and sparsity of biomolecular networks.

- Assume most edges to be zero.

- more specifically: observed partial correlations p across all edges follow mixture distribution:

$$f(p) = \eta_0 f_0(p; \kappa) + \eta_A f_A(p) \tag{2}$$

with $\eta_0 + \eta_A = 1$ and $\eta_0 >> \eta_A$.

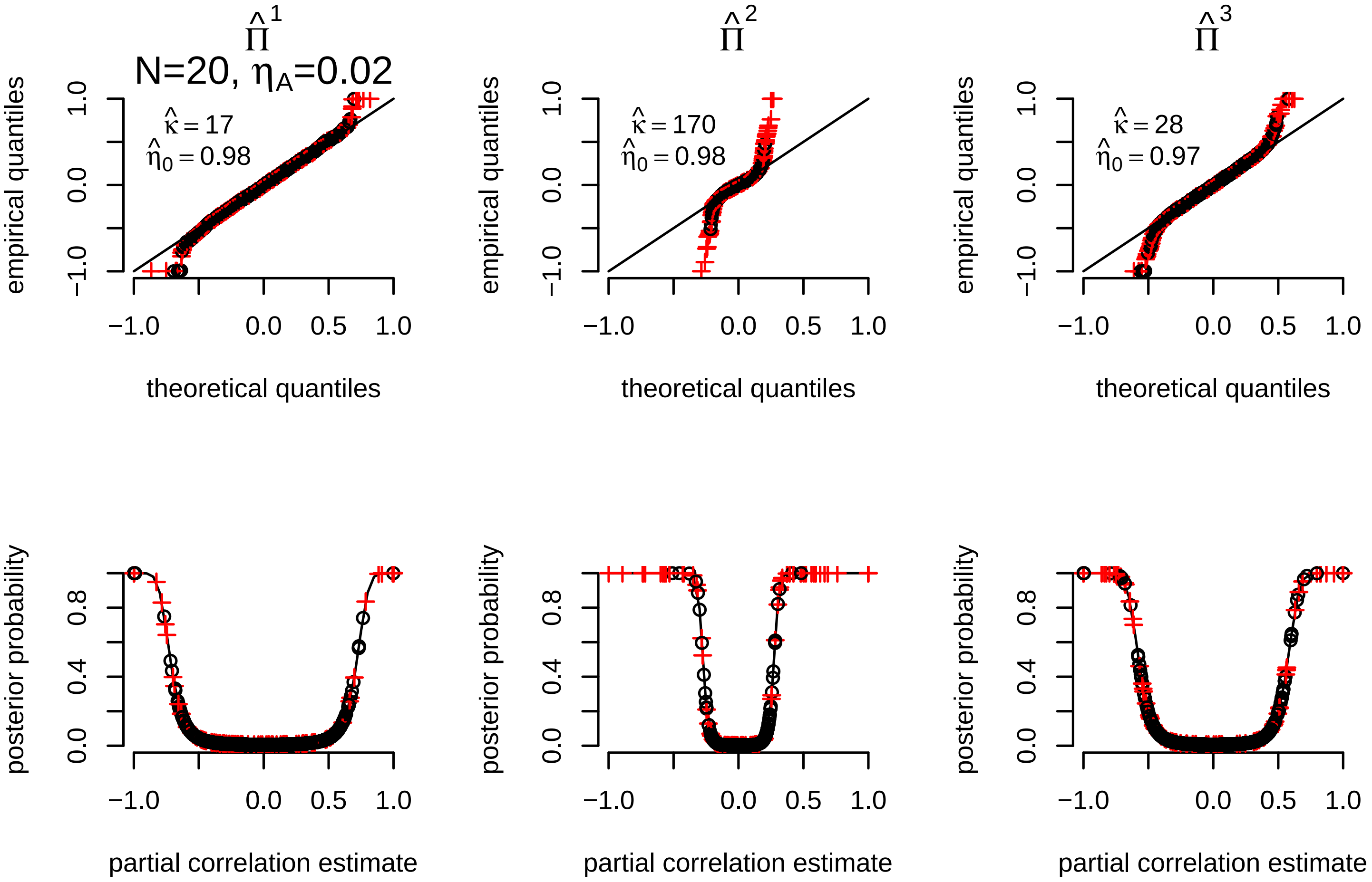
- alternative distribution f_A: uniform distribution from -1 to 1

Trick 5 in style of empirical Bayes methods for problems of differential expression (Sapir & Churchill, 2000; Efron *et al.*, 2001; Efron, 2003)

Fit of Mixture Distribution (next slide):

- QQ plots for all three estimates in small-sample example with $N = 20$, $G = 100$, and $\eta_A = 0.02$ (top row)
- supplementary: empirical posterior probability plots of an edge being present (bottom row)

$$\text{pr}(\text{non-zero edge}|\hat{p}) = \frac{\hat{\eta}_A f_A(\hat{p})}{f(\hat{p};\hat{\kappa})} \tag{3}$$


$\hat{\Pi}^1$
N=20, η_A=0.02
$\hat{\kappa} = 17$
$\hat{\eta}_0 = 0.98$
$\hat{\Pi}^2$
$\hat{\kappa} = 170$
$\hat{\eta}_0 = 0.98$
$\hat{\Pi}^3$
$\hat{\kappa} = 28$
$\hat{\eta}_0 = 0.97$
empirical quantiles
−1.0
0.0
1.0
theoretical quantiles
−1.0
0.0
0.5
1.0
posterior probability
0.0
0.4
0.8
partial correlation estimate

Model Selection Using FDR Multiple Testing

False discovery rate criterion (Benjamini & Hochberg, 1995): control expected proportion of false positives

1. Set of ordered p-values $p_{(1)}, p_{(2)}, \dots, p_{(M)}$ corresponding to all potential edges $e_{(1)}, e_{(2)}, \dots, e_{(M)}$
2. Let i_Q be largest i with $p_{(i)} < \frac{i}{M} \frac{Q}{\eta_0}$
3. Reject null hypothesis of zero partial correlation for edges $e_{(1)}, e_{(2)}, \dots, e_{(i_Q)}$

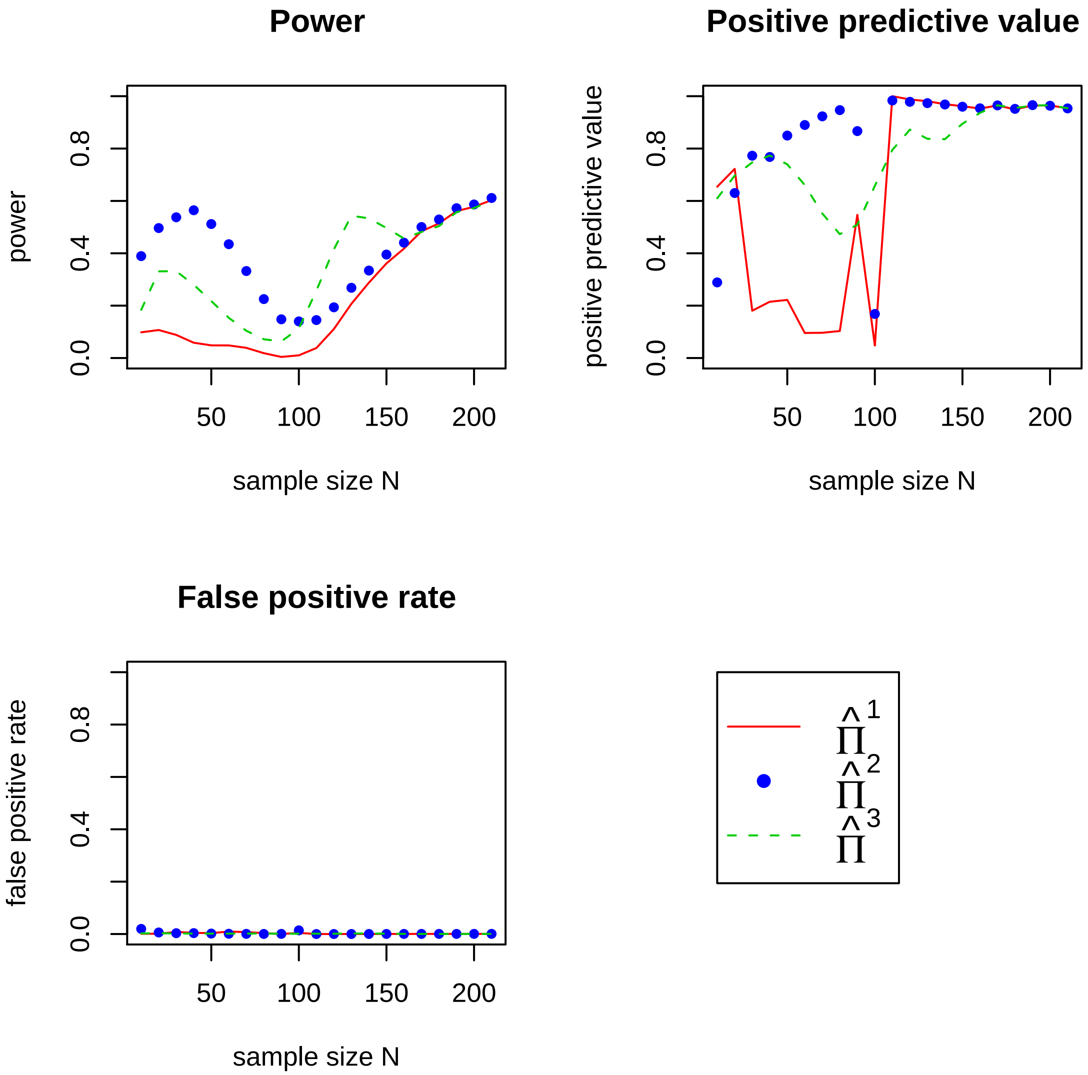
Approximation to proper model search!

Power analysis

Investigation of statistical properties of proposed model selection procedure for $\hat{\Pi}^1$, $\hat{\Pi}^2$, and $\hat{\Pi}^3$:

- FDR level $Q = 0.05$
- empirical power (sensitivity, true positive rate)
- empirical false positive rate (1-specificity)
- positive predictive value

Simulation setup: $G = 100$ and $\eta_A = 0.02$ with $N = 10, 20, \ldots, 210$



Power

Positive predictive value

False positive rate

Summary: Recipe of Analysis

1. choose suitable point estimate of partial correlation

2. estimate degree of freedom κ of underlying null distribution

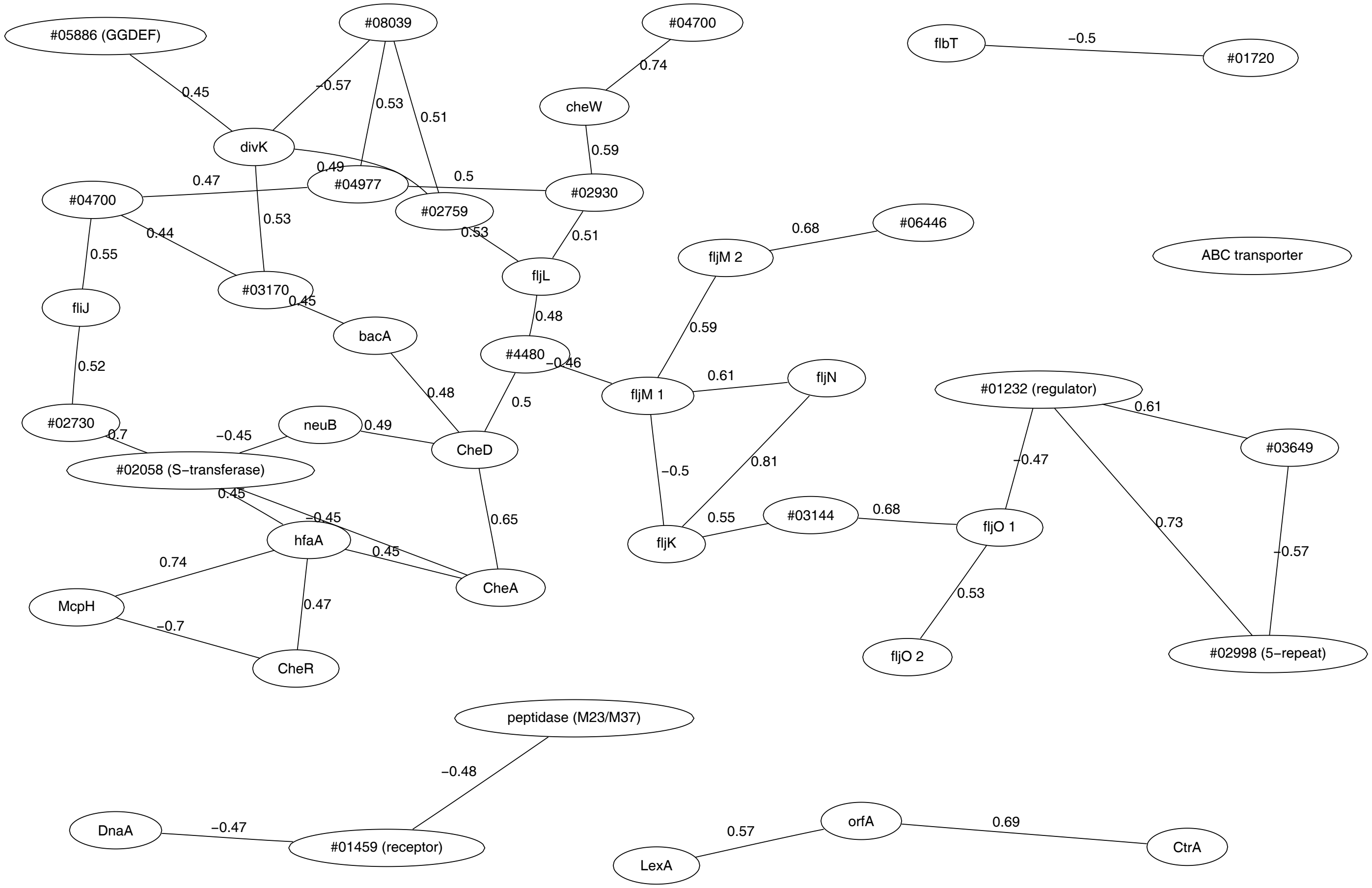
3. compute two-sided p-values and posterior probabilities, respectively, for all possible edges

4. apply multiple testing procedure using FDR criterion to determine graph topology (exploratory tool!)

5. visualize resulting network structure

Molecular Data

- cell cycle in *Caulobacter crescentus* (Laub *et al.*, 2000)
- 3062 genes and ORFs at 11 sampled time points
- reduced to 1444 (due to missing values) and further to 42 potentially interesting genes and ORFs (Wichert *et al.*, 2004)
- 47 significantly non-zero partial correlations


#05886 (GGDEF)
#08039
#04700
flbT
−0.5
#01720
0.45
−0.57
0.53
0.51
0.74
cheW
divK
0.59
0.49
0.47
0.5
#04977
#02930
#04700
#02759
0.53
0.44
0.53
0.51
0.68
#06446
0.55
fljM 2
ABC transporter
fljL
#03170
0.45
fliJ
bacA
0.48
0.59
#4480
−0.46
0.52
0.48
0.5
0.61
fljN
fljM 1
#01232 (regulator)
0.61
#02730
0.7
−0.45
neuB
0.49
CheD
#03649
#02058 (S−transferase)
−0.5
0.81
−0.47
0.45
−0.45
0.65
0.55
#03144
0.68
fljO 1
0.73
hfaA
fljK
−0.57
0.74
0.45
CheA
McpH
0.47
0.53
−0.7
CheR
fljO 2
#02998 (5−repeat)
peptidase (M23/M37)
−0.48
DnaA
−0.47
#01459 (receptor)
orfA
0.57
0.69
LexA
CtrA

Discussion

We have presented a novel framework for inferring large GGMs from small-sample data sets such as microarray (time series) data sets.

Key Insights:

- we may employ bagging to obtain improved point estimates of partial correlation
- we can exploit the sparsity of the network to estimate the null distribution from the point estimate of the correlation matrix
- heuristic (but fast) model selection can be done via multiple testing (using frequentist FDR method or empirical Bayes)

Discussion ctd.

Advantages:

- in contrast to other applications of GGMs to micorarray data the analysis can take place on the gene level (interpretability)
- our simulation results suggest that sensible estimation of sparse graphical models is possible in the proposed graphical Gaussian modeling framework, even for small samples.
- the inference procedure is computationally efficient
- software is available in R (GeneTS version 2.0)

Discussion ctd.

Further points to consider:

- critical review of model assumptions (i.i.d., normality)
- though estimation of κ somehow accounts for longitudinal autocorrelation in the data, data should be treated as proper time series
- heuristic network search may be improved
- imperfect fit of null distribution for $\hat{\Pi}^2$ may be modified to improve statistical testing for very small samples
- GGMs may serve as a starting point to build more sophisticated graphical models (Bayesian nets, dynamics etc).
- graphical model framework is suitable statistical approach to modeling, but inference and model selection remain challenging