The non-rejection rate for structural learning of gene transcriptional networks from E.coli microarray data

Alberto Roverato

Department of Statistics, University of Bologna, Italy

work in collaboration with

Robert Castelo

Dept. of Experimental and Health Sciences, Pompeu Fabra University, Spain

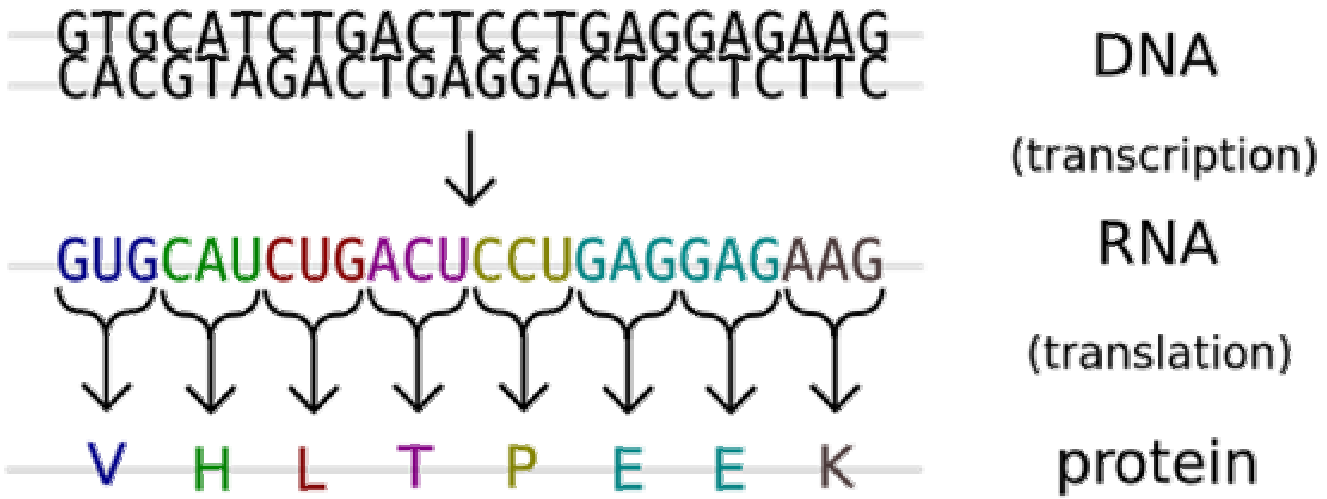
Mathematical Aspects of Graphical Models, Durham, July 2008

Outline

- Gene expression
- transcriptional regulatory networks
- structural learning of biological networks
- co-expression networks
- the non-rejection rate
- two different applications of the non-rejection rate
- concluding remarks
- references

Gene expression

1. **Gene expression** is the process by which information from a gene is made into a functional gene product, such as protein or RNA
2. when and in what quantities a gene is expressed, determines differential protein abundance, thereby inducing different cell functions



Transcription factors (TFs)

1. In general, each mRNA molecule makes a specific protein (or set of proteins)
   - structural protein: gives the cell particular structural properties
   - enzyme: micro-machine that catalyses certain reactions
   - **transcription factor (TF)**: protein which serves to regulate other genes
2. transcription factors bind to the promoter region of other genes turning them on, thereby initiating the production of another protein
3. a transcription factor may be either an activator or a repressor

Gene regulatory networks (GRNs)

1. A **gene regulatory network (GRN)** is a graph where
   - **vertices**: genes or, more generally, DNA segments
   - **edges**: direct regulatory interactions
2. transcription factors are the main players in regulatory networks
3. alternative name: **transcriptional regulatory network**
4. some specific features
   - feedback relationships and self regulation are possible
   - **regulatory networks are SPARSE**
   - presence of hubs
   - **GRNs are dynamic objects** which modify their interaction structure to allow the cell to respond effectively to changes of its internal and external environments

Biological pathway construction

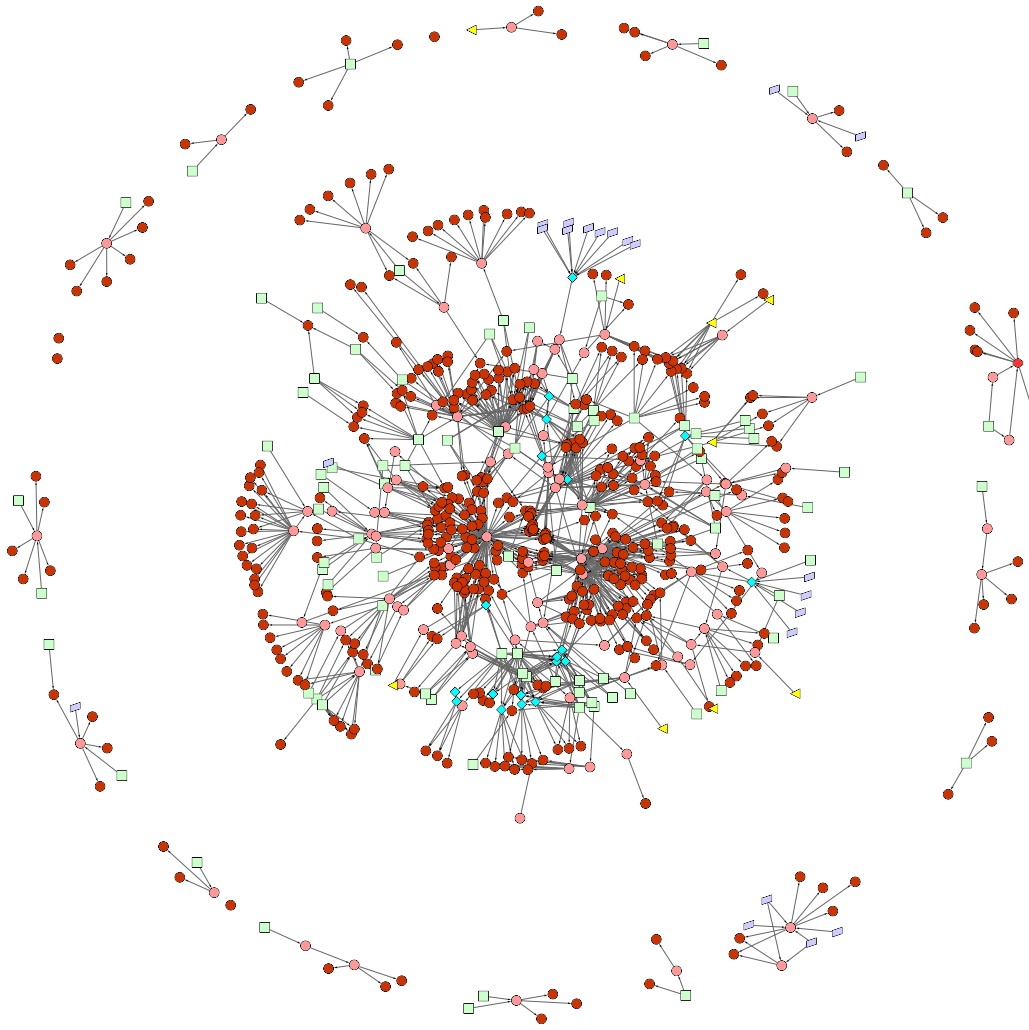
1. Pathway building is the process of identifying the entities and interactions of a network
2. two **Construction Processes**:
   - **Data-Driven (DDCP)**: relationship information between entities is learnt from specific experiments such as a microarray study
   - **Knowledge-Driven (KDCP)**: relationship information between entities is learnt by mining existing databases. Data repositories, which contain information regarding sequence data, metabolism, signaling, reactions and interactions are a major source of information for pathway building
3. KDCP can provide **benchmark networks** useful to validate statistical procedures

E.coli and RegulonDB

- RegulonDB is an internationally recognized reference database of *Escherechia coli* (E.coli) offering curated knowledge of the regulatory network and operon organization
- RegulonDB is currently the largest electronically-encoded dababase of the regulatory network of any free-living organism
- reference:
  Socorro, G.C. *et al.* (2008).
  RegulonDB (Version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation
  Nucleic Acids Research, 2008, vol 36, D120-D124

Gene regulatory network of E.coli

Map of the transcriptional regulatory network controlling metabolism in E. coli. There are genes coding for the TFs (pink circles), genes coding for enzymes (brown circles), external metabolites (green squares), certain internal fluxes (purple parallelograms), stimuli (yellow triangles) and other conditions (blue diamonds)
(Areejit and Sanjay, 2008)



RegulonDB: TF–gene interactions

1	2	3	4
AcrR	acrA	–	Binding of cellular extracts, Gene expression analysis, Binding of purified proteins, Human inference based on similarity to consensus sequences
AdiY ⋮	adiA ⋮	+ ⋮	Gene expression analysis ⋮

1. transcription factor
2. gene regulated by the TF
3. regulatory effect of the TF on the regulated gene
   (+ activator, – repressor, +– dual, ? unknown)
4. evidence that supports the existence of the regulatory interaction

Microarray data

- Microarray data measure the gene expression level by the abundance of mRNA produced
- the number of variables (genes), p, is very large
- the sample size, n, is small, compared to p

Structural learning of GRNs

1. Well defined statistical models
2. typically
   (a) data are assumed to be $i.i.d.$ observations from a multivariate normal distribution
   (b) exploit background information on the network structure, in particular network sparseness to overcome the $p >> n$ problem
3. some examples
   - Bayesian approach with sparsity inducing prior (Dobra *et al.*, 2004)
   - lasso estimate of the inverse covariance matrix (Friedman *et al.*, 2007)
   - shrinkage estimate of the covariance matrix (Schäfer and Strimmer, 2005)
   - limited order partial correlations (Wille and Bühlmann, 2006)
   - ⋮

Co-expression networks

- No formal definition: edges represent “associations” between genes, hopefully a direct regulatory interaction
- main task
  - identify associated genes
  - try to reduce the number of spurious associations
- **the goal is not to recover *all* direct regulatory interactions but rather to recover *some* transcriptional interactions with high confidence**
- typically the output of these procedures is a ranking of the edges of the complete graph
- performance of the procedure is evaluated with respect to a benchmark network

Some popular procedures

1. **Relevance networks**: associations are marginal dependence between genes (Butte and Kohane, 2000)
2. **relevance networks with correction for spurious associations**
   - ARACNE (Margolin *et al.*, 2006)
   - CLR (Faith *et al.*, 2007)
3. ⋮

Validation of procedures

- **ROC-curve**: it is not very useful because

$$\text{specificity} = \frac{\#\text{ missing edges correctly identified}}{\#\text{ missing edges}}$$

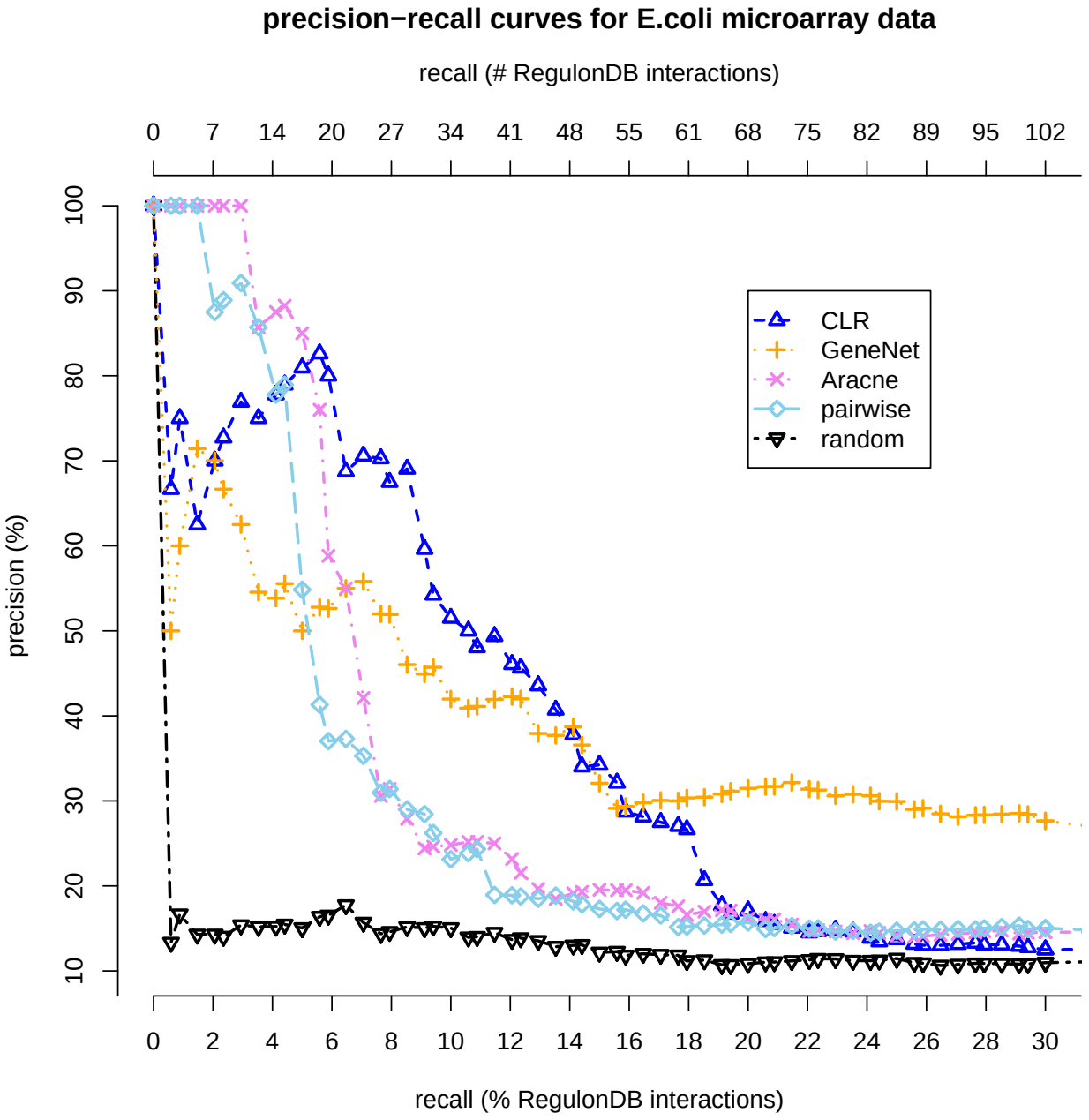
  is always close to one, because of sparseness;

- **Precision-Recall** curve: (**recall=sensitivity**)

$$\text{recall} = \frac{\#\text{ present edges correctly identified}}{\#\text{ present edges}}$$

$$\text{precision} = \frac{\#\text{ present edges correctly identified}}{\#\text{ identified edges}}$$

Example of precision-recall curve



Pooled datasets

- Typically these procedures are applied to pooled datasets from different experimental conditions so that
  1. sample size is larger
  2. more direct regulatory interactions can be identified

- does it make sense to apply graphical models in this context?

Gaussian graphical models

- Finite set $V = \{1, 2, \ldots, p\}$
- (sparse) undirected graph $G = (V, E)$
- random vector $X_V \sim N(\mu, \Sigma)$ Markov w.r.t. G
- $i.i.d.$ random sample of size n form X_V
- the subset $Q \subseteq V$ identifies the subvector X_Q
- **q-order partial correlation**: if $Q \subseteq V \backslash \{i, j\}$ with $q = |Q|$, then

$$\rho_{ij.Q}$$

is the partial correlation between X_i and X_j given X_Q

Full vs. limited-order partial correlations

- if $Q = V \backslash \{i, j\}$ then $\rho_{ij.Q}$ is a **full-order** partial correlation and we write

$$\rho_{ij.rest}$$

- in a frequentist approach, testing the hypothesis

$$H_0 : \rho_{ij.rest} = 0$$

  requires the computation of S^{-1}

- the sample covariance matrix S has full rank, with probability one, iif $n > p$ (Dykstra, 1970)

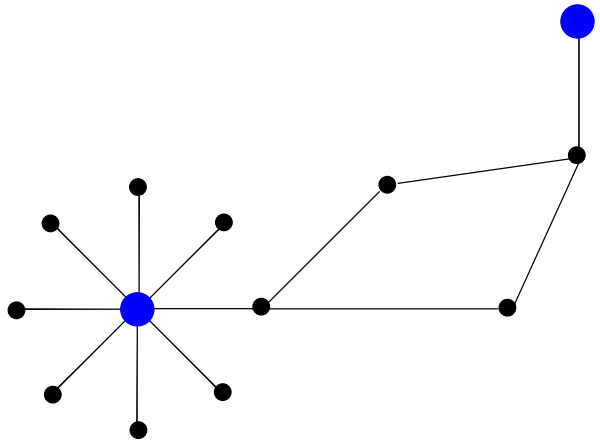
$\Longrightarrow$ **try to use limited-order partial correlations**

q–order p.c. and missing edges

- If $q < (n-2)$ then for any $Q \subset V$ such that $|Q| = q$ the hypothesis

$$H_0 : \rho_{ij.Q} = 0$$

can be verified with standard techniques;



- when G is sparse it seems natural to investigate the possible use of q-order partial correlations to identify missing edges of the graph (see Castelo and R., 2006)
- however...

Learning procedures based on q–order p.c.

- **Faithfulness assumption**: all the conditional independence relationships in X_V can be read off the graph G through the Markov property
- for every pair of variables, X_i and X_j, there are $\binom{p-2}{q}$ different q-order partial correlations
- the edge (i, j) should be removed when at least one of the $\binom{p-2}{q}$ q-order partial correlations is equal to zero
- consequently
  1. computational problems unless q is very small;
  2. multiple testing problem;
  3. what if faithfulness assumption fails?

Non-rejection rate

NON-REJECTION RATE for the pair (i, j): $E(T_{ij}^q)$

- Set a value $q < (n-2)$
- T_{ij}^q results from a two stage experiment:
  1. select randomly a subset $Q \subset V \backslash \{i, j\}$ with $|Q| = q$
  2. test the hypothesis $H_0 : \rho_{ij.Q} = 0$
-

$$T_{ij}^q = \begin{cases} 0 & \text{if } H_0 \text{ is rejected} \\ 1 & \text{otherwise} \end{cases}$$

Present edges and non-rejection rate

The non-rejection rate takes value between 0 and 1 and behaves differently for present and missing edges.

If $(i, j) \in E$ then

$$E(T_{ij}^q) = \beta_{ij}$$

- β_{ij} average second type error over all sets $Q \subset V \backslash \{i, j\}$ with $|Q| = q$
- β_{ij} converges to zero as $(n - q)$ increases

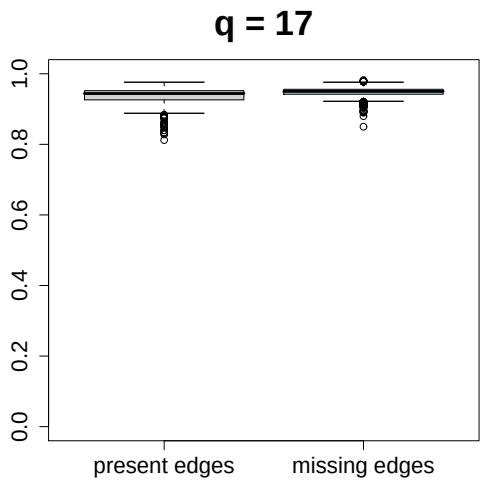
Missing edges and non-rejection rate

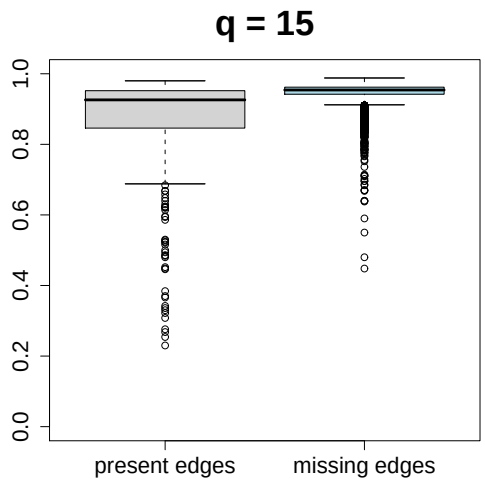
If $(i,j) \notin E$ then

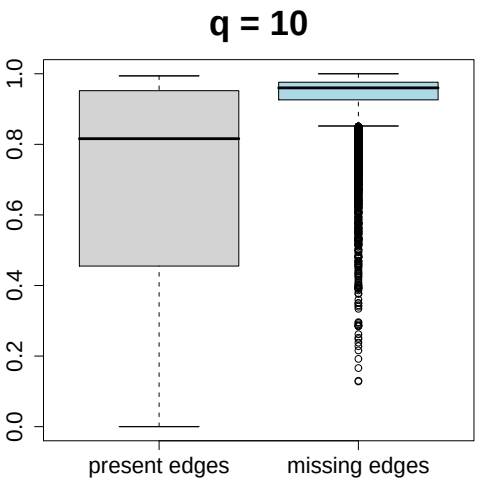
$$E(T_{ij}^q) = \beta_{ij}\,(1-\pi_{ij}) + (1-\alpha)\,\pi_{ij}$$

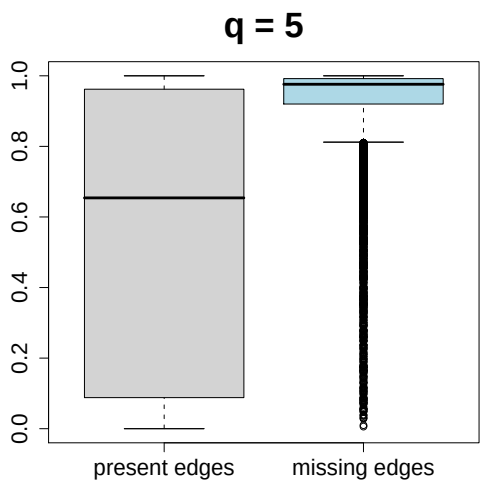
- α is the significance level of tests
- π_{ij} is the proportion of subsets $Q \subset V \backslash \{i,j\}$ with $|Q| = q$ which separate i and j in G
- if $q = p-2$ then $E(T_{ij}^q) = (1-\alpha)$.

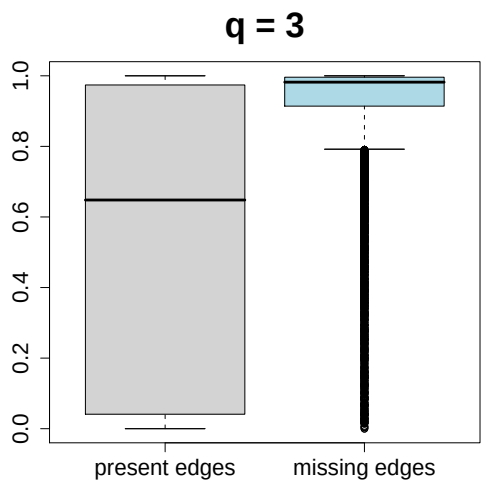
Example with $p = 150$ and $n = 20$

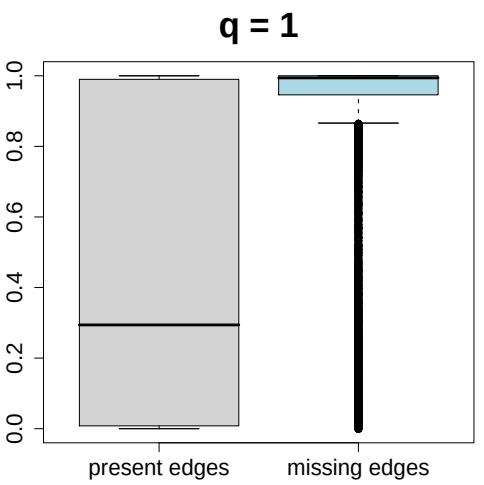












Interpretation of the non-rejection rate

- Role of $(n-p)$ in inference for partial correlations
- the quantity $n-p$ is split into two parts:

$$(n-p) = (n-q) + (q-p)$$

- $(n-q)$ has to be sufficiently large to guarantee the required power of statistical tests
- $(q-p)$ is always negative, and has to be sufficiently close to zero to exploit the sparseness of G
- there is a trade-off between these two requirements but but q can be chosen accordingly with the real dimension of the problem.

First possible use of the non-rejection rate

1. Specify a threshold β^*
2. return a graph $\widehat{G}$ obtained by removing from the complete graph all the edges whose estimated non-rejection rate is greater than β^*
3. **conservative procedure** that aims at keeping the number of wrongly removed edges small: β^* is set close to one
4. the selected graph $\widehat{G}$ will contain a large number of edges that are missing in G
5. typically, $\widehat{G}$ can be dealt with standard techniques so that the selected graph is the **starting point for further analysis**.

Example: simulated data

- Number of genes: $p = 164$
- sample size: $n = 40$
- value of q: $q = 20$
- block diagonal structure
- number of possible edges $13\,366$
- present edges 1206
- sparsity degree 9%

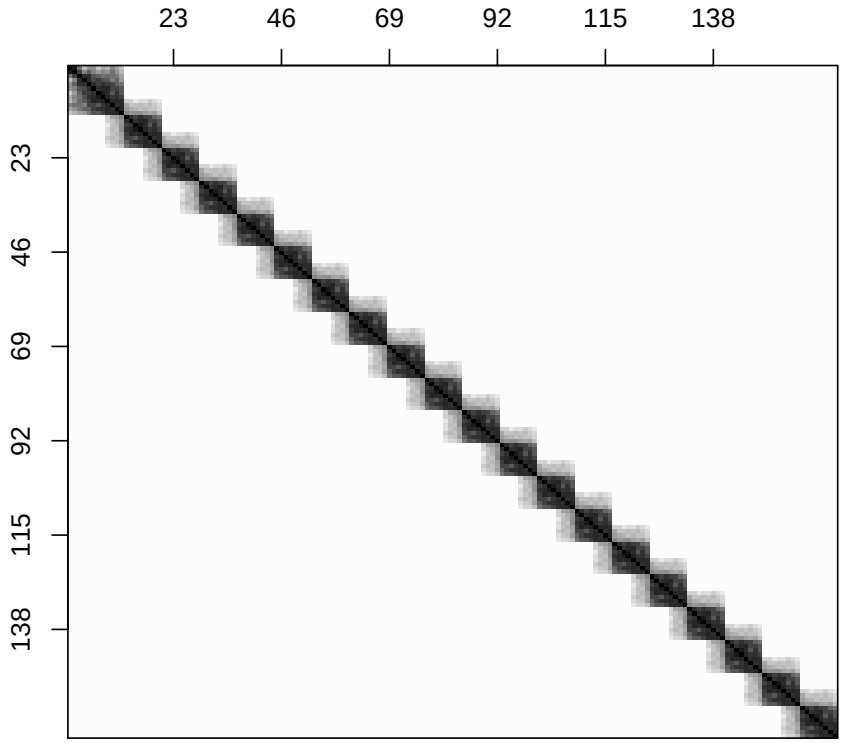
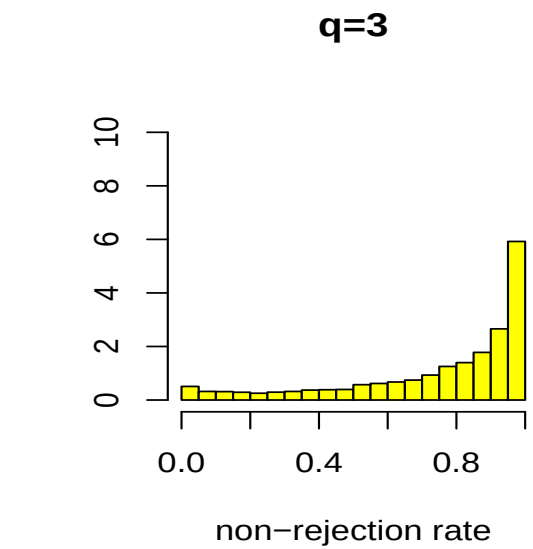
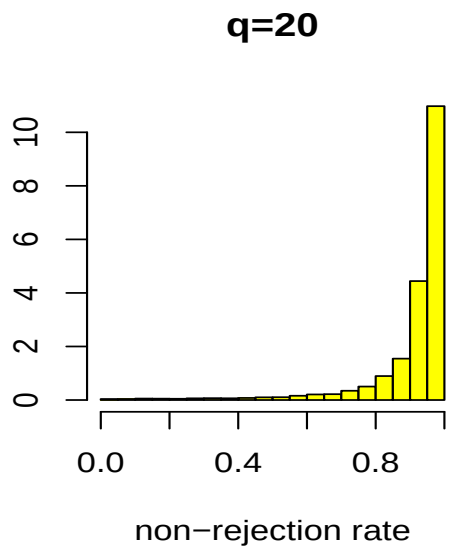
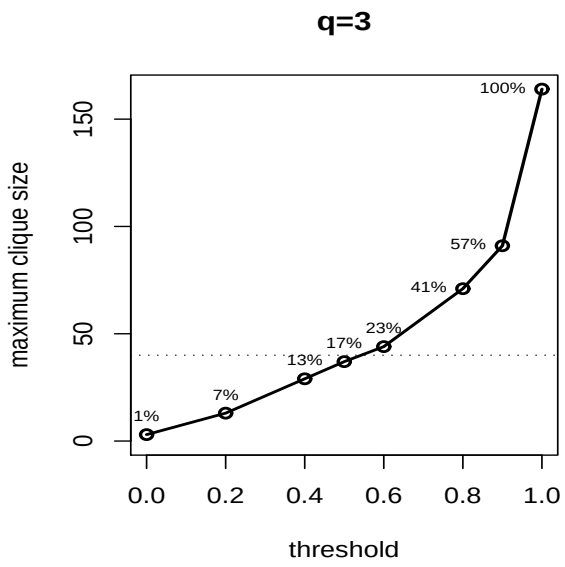


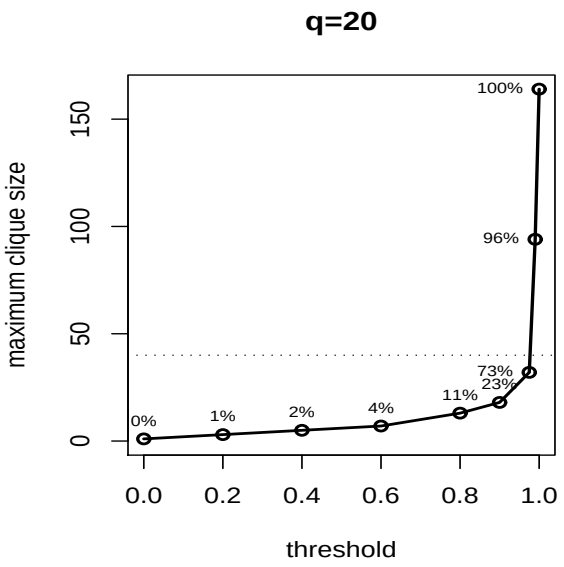
image of the partial correlation matrix

Visual check for sparsity

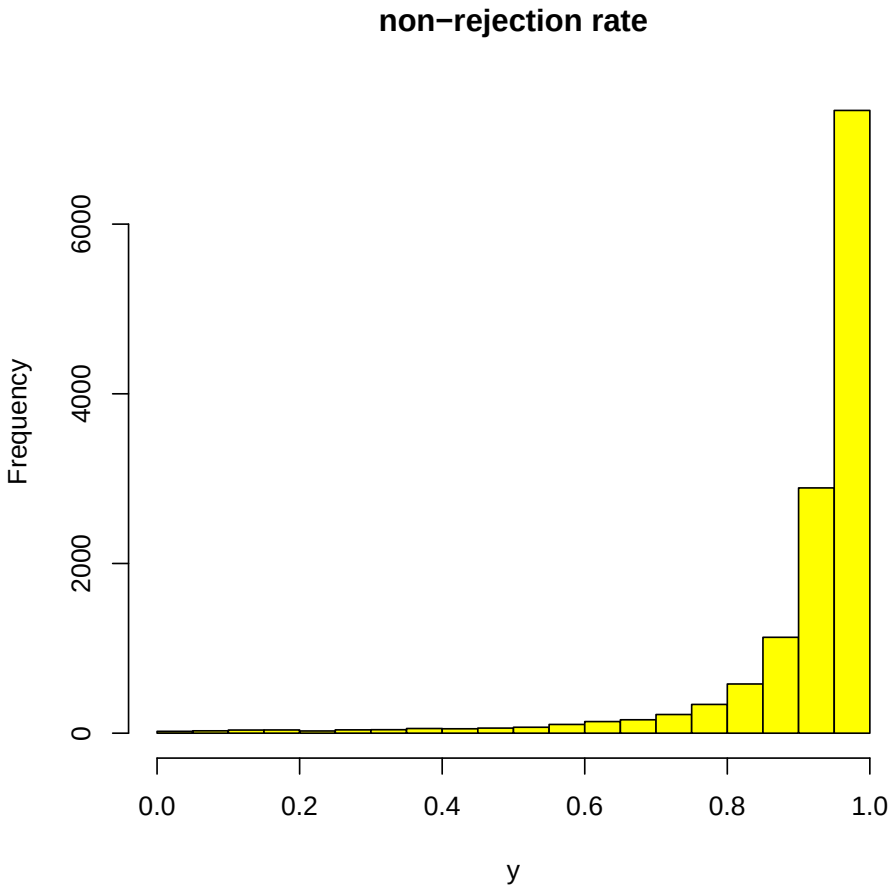


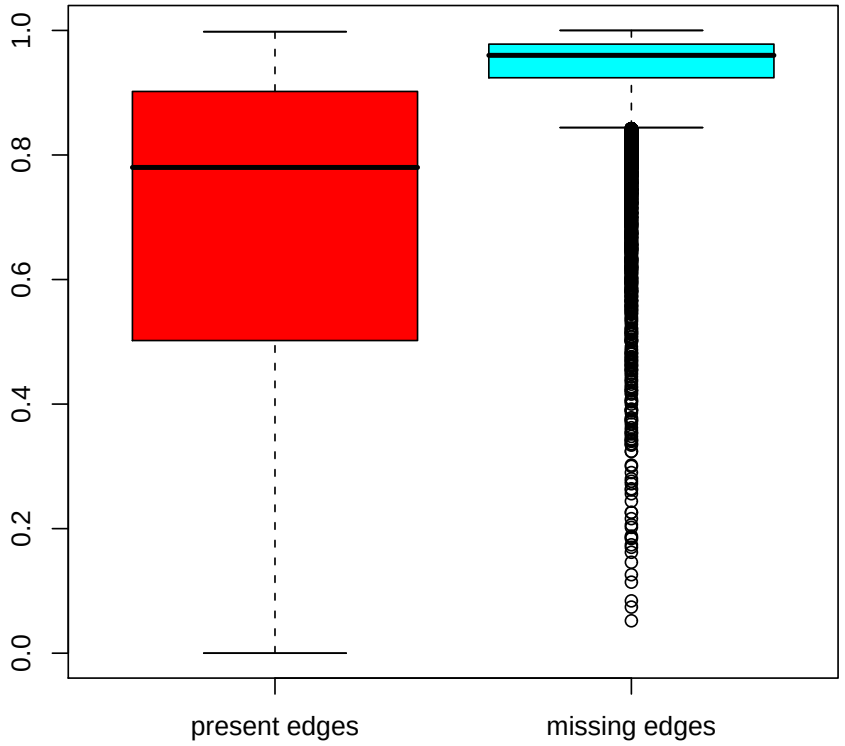




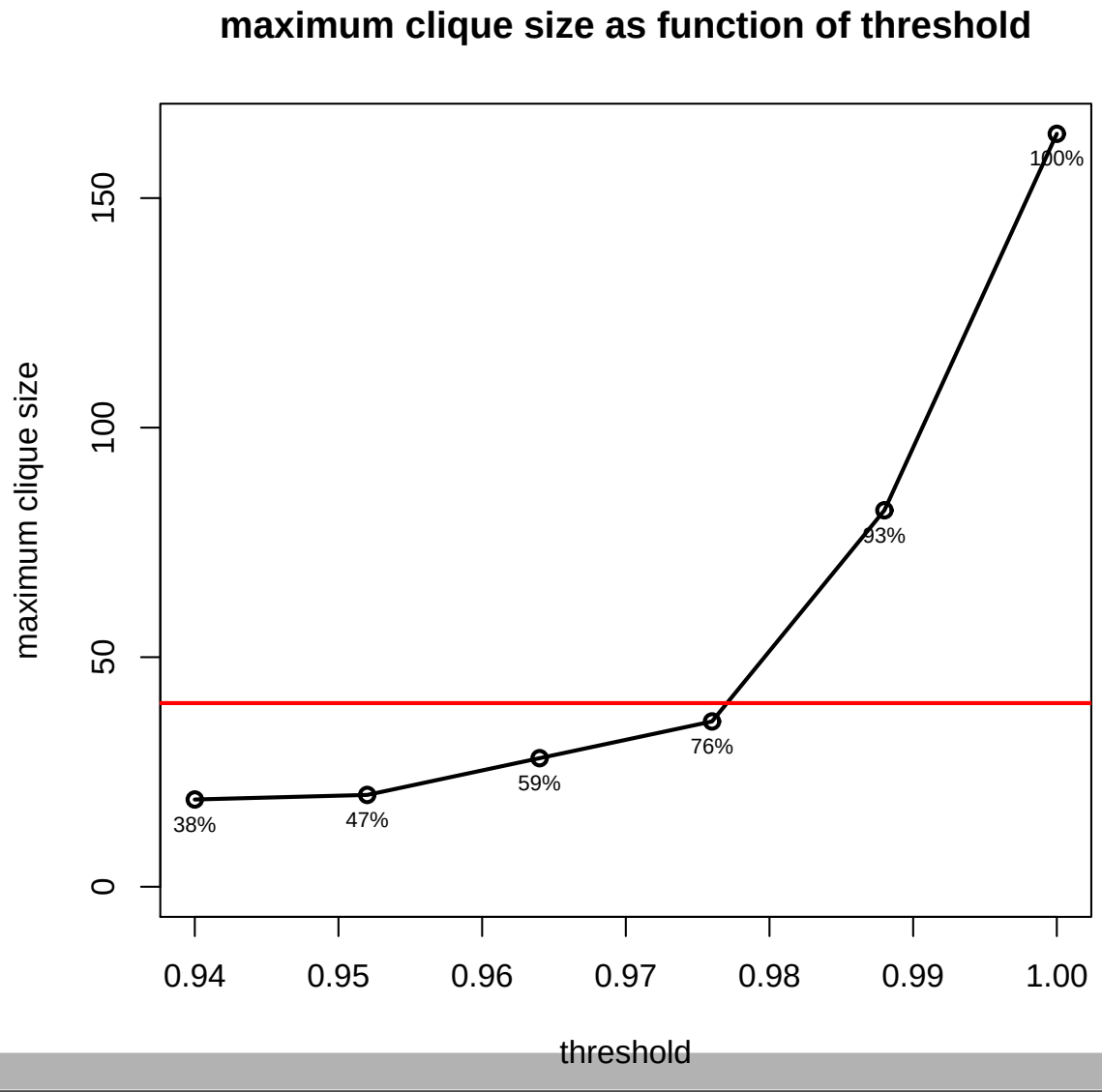


Distribution of non-rejection rates





Choosing the threshold: qp-clique plot

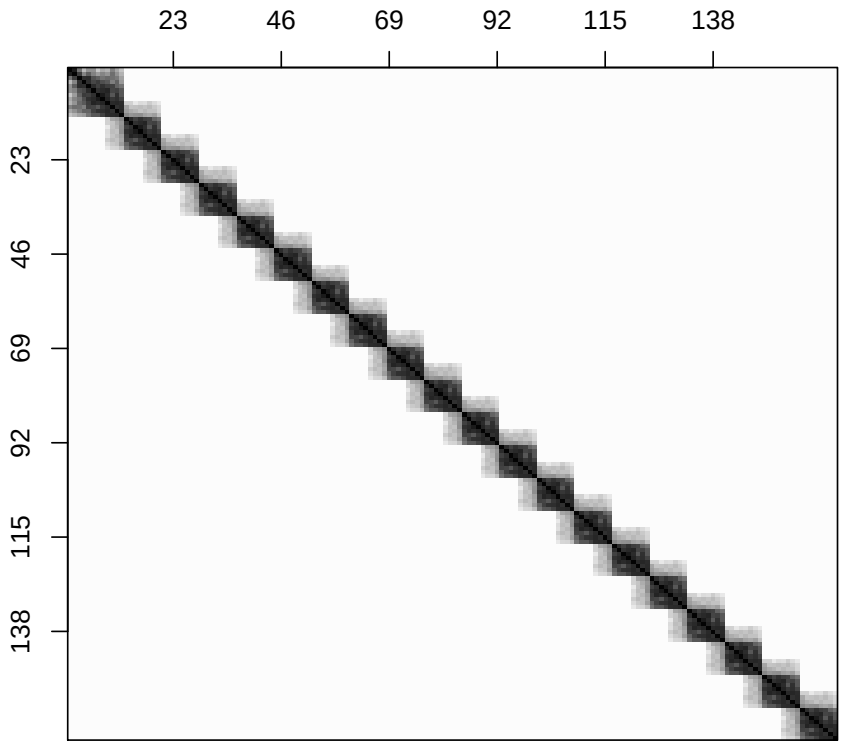


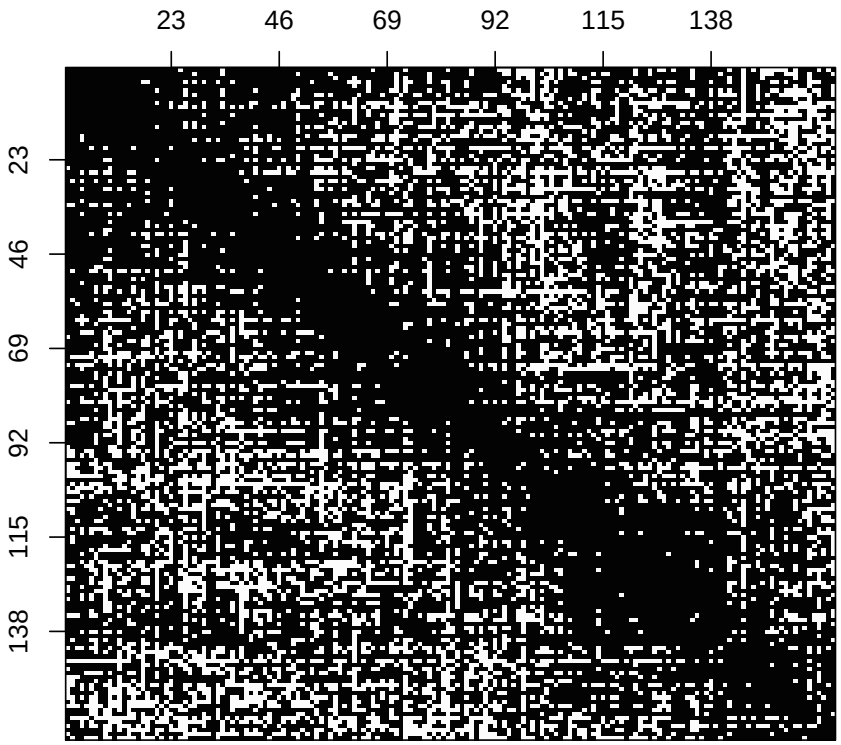
Performance of the procedure

ERROR to be controlled: removal of a present edges

threshold	0.96	0.976
max. clique size	20	36
% removed edges	$47.2\,\%$	$26.7\,\%$
relative error	$6.7\,\%$	$3.1\,\%$
absolute error	82	38

True graph vs. selected graph for thr.=0.976

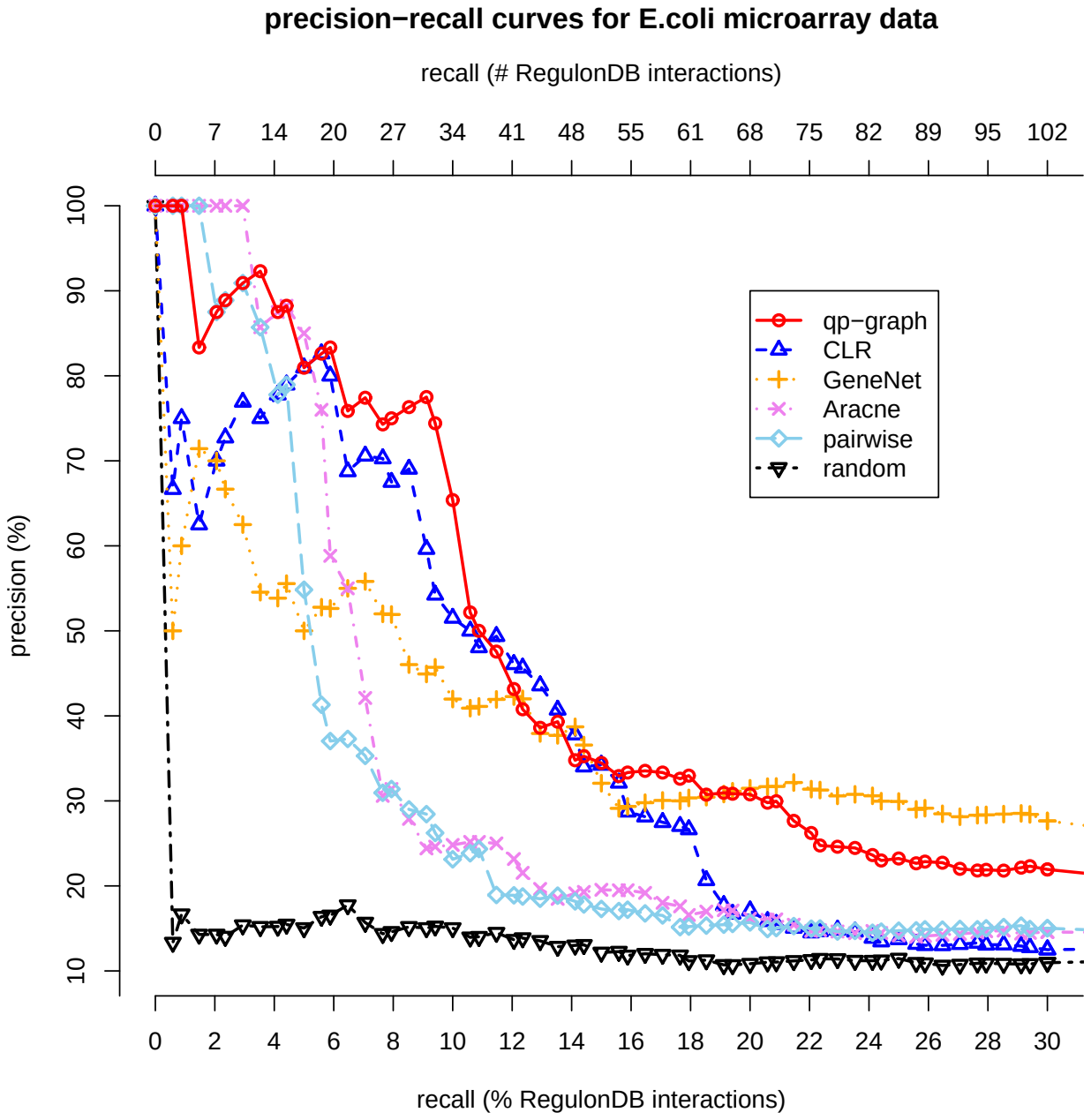




Second possible use of the non-rejection rate

- co-expression network approach
- main attention on non-rejection rates with small value (close to zero)
- microarray experiments from GEO (GSE1121) from a study by Cover et al. (2004) that investigated the changes of global gene expression in E. coli during an oxygen shift
- E. coli antisense Affymetrix chip with 7,312 probesets
- $n = 43$
- data are filtered and the number of genes reduced to 199
- a subset of 341 interactions is extracted from RegulonDB

Precision-recall curve



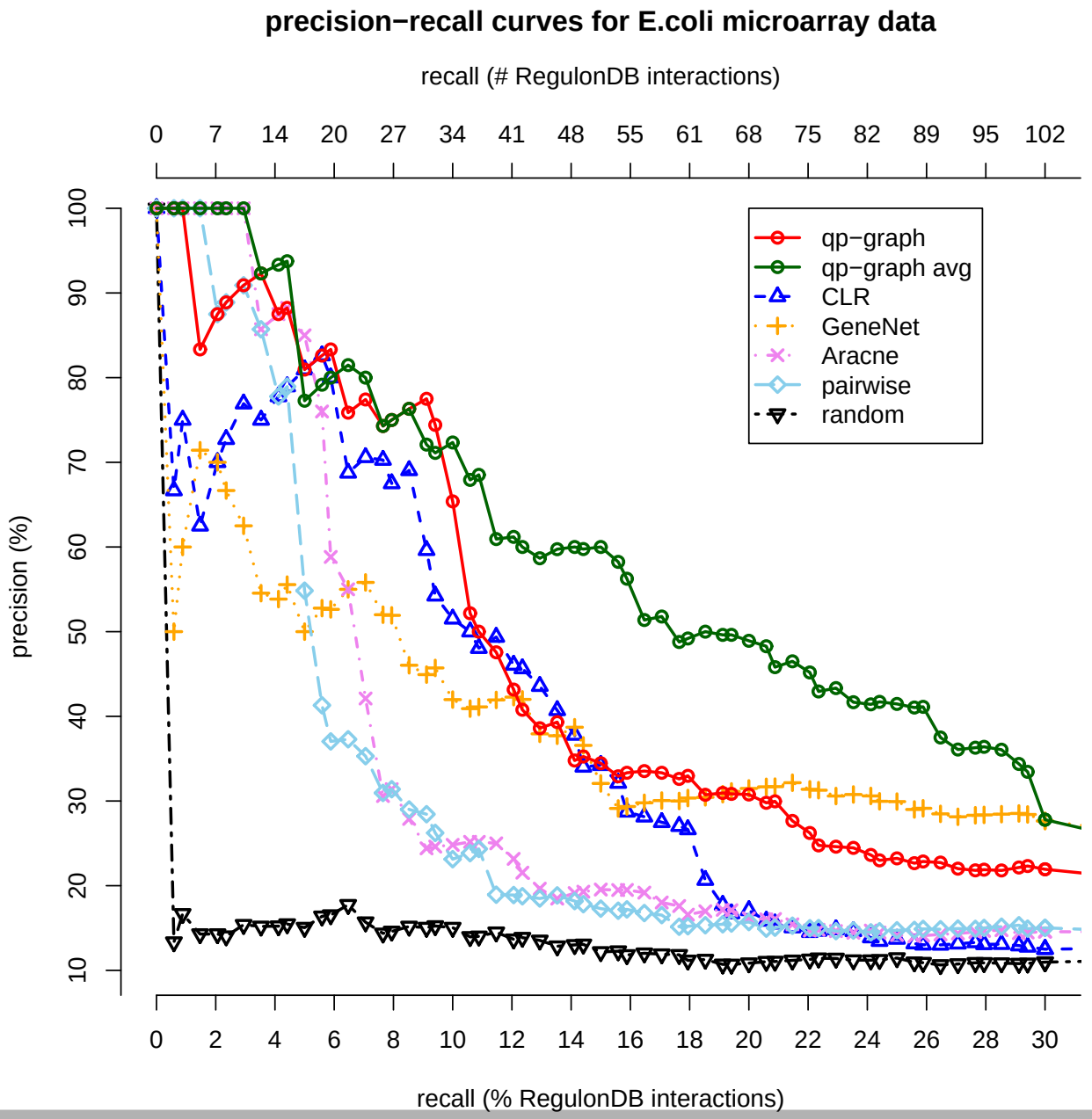
A closer look at the data

- Two different experimental conditions
  1. $n_1 = 21$ aerobic
  2. $n_2 = 22$ anaerobic
- generalization of the non-rejection rate to deal with multiple experimental conditions

Generalization to multiple datasets

- Set a value $q < (n_1 - 2), (n_2 - 2)$
- consider a binary random variable T_{ij}^q associated to the edge (i, j) resulting from a **three** stage experiment
  1. **select one of the two datasets with probability proportional to n_1 and n_2**
  2. select randomly a subset $Q \subset V \backslash \{i, j\}$ with $|Q| = q$
  3. test the hypothesis $H_0 : \rho_{ij.Q} = 0$
- T_{ij}^q is equal 0 if H_0 is rejected and 1 otherwise
- **NON-REJECTION RATE:** $\boldsymbol{E(T_{ij}^q)}$

Precision-recall curve



Estimation of the non-rejection rate

- For every pair of variables, estimation of the non-rejection rate requires to carry out $\binom{p-2}{q}$ statistical test
- Monte Carlo sampling of sets $Q \in V \backslash \{i, j\}$
- computation of non-rejection rates are easy to implement but computer intensive. R package "qp" can be downloaded from CRAN;
- non-rejection rates can be computed in parallel

Concluding remarks

1. The distinction between the structural learning approach and the co-expression network approach is important
2. graphical models can be useful in both approaches
3. formal procedures to deal with pooled datasets are called for
4. non-rejection rate is an empirical quantity, but it is an useful tool
   - easy to compute
   - no multiple testing problem
   - robust with respect to the *faithfulness* assumption
   - sparseness is not assumed but exploited when present

References

- Butte, A.J. and Kohane, I.S. (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput: 418-429.
- Castelo R. and Roverato A. (2006). Gaussian graphical model search from microarray data with p larger than n. *J Mach Learn Res*, 7, 2621-2650.
- Dobra, A. Hans, C. Jones, B. Nevins, J.R. and West. M. (2004). Sparse graphical models for exploring gene expression data. *J. Mult. Anal.* 90: 196Ű-212.
- Dykstra, R.L. (1970). Establishing the positive definiteness of the sample covariance matrix. *Ann. Math. Statist.*, 41, 6, 2153-2154.
- Faith, J. *et al.* (2007). Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. PLOS Biology, 5, 1, 54-66

References

- Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9, 3, 432-441.
- Margolin, A.A. Nemenman, I. Basso, K. Wiggins, C. Stolovitzky, G. Favera R.D. and Califano. A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks ina mammalian cellular context. *BMC Bioinformatics*, 7(Suppl 1):S7.
- Schäfer, J. and Strimmer K. (2005): A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.* 4: 32.
- Wille A. and Bühlmann. P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology*, 5(1): article 1.