Bioinformatics Advance Access published October 12, 2004

BIOINFORMATICS

An Empirical Bayes Approach to Inferring Large-Scale Gene Association Networks

Juliane Schäfer[1] *and Korbinian Strimmer*[1]

[1]*Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany*

Received on December 28, 2003; last revised September 18, 2004.

ABSTRACT

Motivation: Genetic networks are often described statistically by graphical models (e.g. Bayesian networks). However, inferring the network structure offers a serious challenge in microarray analysis where the sample size is small compared to the number of considered genes. This renders many standard algorithms for graphical models inapplicable, and inferring genetic networks an "ill-posed" inverse problem.

Methods: We introduce a novel framework for small-sample inference of graphical models from gene expression data. Specifically, we focus on so-called graphical Gaussian models (GGMs) that are now frequently used to describe gene association networks and to detect conditionally dependent genes. Our new approach is based on (i) improved (regularized) small-sample point estimates of partial correlation, (ii) an exact test of edge inclusion with adaptive estimation of the degree of freedom, and (iii) a heuristic network search based on false discovery rate multiple testing. Steps (ii) and (iii) correspond to an empirical Bayes estimate of the network topology.

Results: Using computer simulations we investigate the sensitivity (power) and specificity (true negative rate) of the proposed framework to estimate GGMs from microarray data. This shows that it is possible to recover the true network topology with high accuracy even for small-sample data sets. Subsequently, we analyze gene expression data from a breast cancer tumor study and illustrate our approach by inferring a corresponding large-scale gene association network for 3,883 genes.

Availability: The authors have implemented the approach in the R package "GeneTS" that is freely available from http://www.stat.uni-muenchen.de/~strimmer/genets/, from the R archive (CRAN), and from the Bioconductor web site.

Contact: korbinian.strimmer@lmu.de

INTRODUCTION

Biological processes in the cell such as biochemical interactions and regulatory activities lead to complicated interaction patterns among genes and gene products. It is one of the aims of systems biology to provide suitable mathematical models of these networks. In this regard *graphical models* (Whittaker, 1990; Lauritzen, 1996) have emerged as useful tools because they allow the stochastic description of net-like association and dependency structures in complex high-dimensional data. At the same time, graphical models offer an advanced statistical framework for inference.

Consequently, many in part very complicated graphical models such as Bayesian networks (e.g., Friedman et al., 2000; Segal et al., 2003; Friedman, 2004), auto-regressive models (e.g., Yeung et al., 2002; De Hoon et al., 2003), state-space models (e.g., Murphy, 2002; Rangel et al., 2004), and graphical Gaussian models (e.g., Kishino and Waddell, 2000; Toh and Horimoto, 2002a; Wu et al., 2003; Dobra et al., 2004) have already been applied to genomic data, and put to use in expression analysis.

Unfortunately, as promising graphical models are for the analysis of gene interaction, their practical application is currently strongly limited by the amount of available experimental data. At first, this may seem paradoxical given today's high-throughput facilities. Note however, while these tools now allow to investigate experimentally a greatly increased number of features (genes), the number of available samples has not, and can not, similarly be expanded. As a result, in a typical microarray data set the number of genes G will exceed by far the number of sample points N. This poses a serious challenge to any statistical inference, and also renders estimation of genetic networks an extremely hard problem. This is corroborated by a recent study on the popular Bayesian network method where Husmeier (2003) demonstrated that this approach tends to perform poorly on sparse microarray data.

Motivated by these challenges, great efforts are now being undertaken to further extend the theory of graphical models to allow their large-scale application on

small-sample data (e.g., Wong et al., 2003; Dobra et al., 2004). In this paper we would also like to contribute to this development by proposing a practical empirical Bayes framework for inferring graphical models from sparse microarray data. More specifically, we focus here on improving inference of one of the simplest classes of graphical models, the so-called *graphical Gaussian models* (GGMs). These are similar to the more widely known Bayesian networks in that they allow to *distinguish direct from indirect interactions* (i.e. whether gene A acts on gene B directly or through a third agent C). As any graphical model, they also provide a notion of *conditional independence* of two genes. However, in contrast to Bayesian networks GGMs contain only undirected rather than directed edges. This makes graphical Gaussian interaction modeling on the one hand conceptually more simple, and on the other also potentially more widely applicable (e.g. there are no problems with feedback loops as in Bayesian networks).

GGMs have first been proposed as model for the association structure among genes by Kishino and Waddell (2000). However, a number of difficulties arise when the graphical Gaussian modeling concept is applied to the analysis of microarray data. First, standard GGM theory (Whittaker, 1990) can only be applied when $N > G$, because otherwise the sample covariance and correlation matrices are not positive definite, which in turn prevents the computation of partial correlations. Moreover, there are often additional linear dependencies between the variables, which leads to the problem of multicollinearity. This, again, renders standard theory of graphical Gaussian modeling inapplicable to microarray data. Second, the statistical tests widely used in the literature for selecting an appropriate GGM (e.g., deviance tests) are valid only for large sample sizes, and hence are inappropriate for the very small sample sizes present in microarray data sets. In this case, instead of asymptotic tests an exact model selection procedure is required.

Therefore, to avoid these dimensionality problems, graphical Gaussian modeling has been so far restricted to assess relationships among either a rather small number of genes (Waddell and Kishino, 2000; Bay et al., 2002; Wang et al., 2003) or among a small number of clusters of genes (Toh and Horimoto, 2002a,b; Wu et al., 2003). However, the resulting partial correlation coefficients for the clusters and the corresponding conditional dependence properties are difficult to interpret. For instance, not all the genes of one cluster will interact with all the genes of another cluster. Furthermore, information regarding quality and strength of the association on the gene level is lost when only clusters of genes are considered.

A Novel Small-Sample Framework for Inferring Graphical Gaussian Models

To resolve these issues, we propose here a novel framework for inferring GGMs from small samples. This centers around three new regularized small-sample point estimates of partial correlation. A second key element of this framework is a small-sample edge inclusion test where the degree of freedom of the null distribution is estimated adaptively from the data. This procedure exploits the parallel structure of microarray data in a similar fashion as an empirical Bayes approach suggested by Efron et al. (2001) to identify differentially expressed genes. Finally, multiple testing using the false discovery rate method is employed for heuristic but computationally efficient model selection.

The rest of the paper is organized as follows. In the next section (*Methods*) we introduce the mathematical and statistical background of graphical Gaussian models and present all details of the new small-sample framework for inferring an appropriate model. Subsequently, in *Results* we investigate using extensive computer simulations the question of model validity and the accuracy and power of network selection using the proposed approach. As an example, we then illustrate our framework by applying it to a large-scale breast cancer data set (West et al., 2001) with 3,883 genes and 49 samples. Finally, we discuss the advantages as well as potential drawbacks of our framework and point out further directions of research.

METHODS

Graphical Gaussian Models

Graphical Gaussian models, also known as covariance selection models, are *undirected* graphical models (Dempster, 1972; Whittaker, 1990; Edwards, 1995). Under this approach the observed data matrix X with N rows (=samples) and G columns (=genes) is considered to be drawn from a multivariate Normal distribution $N_G(\mu, \Sigma)$ with some mean vector $\mu = (\mu_1, \ldots, \mu_G)^T$ and positive definite covariance matrix $\Sigma = (\sigma_{ij})$, where $1 \leq i, j \leq G$. Via $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ the covariance matrix Σ can be further decomposed into variance components σ_i^2 and the Bravais-Pearson correlation matrix $P = (\rho_{ij})$.

A high correlation coefficient between any two genes may be indicative of either (i) direct interaction, (ii) indirect interaction, or (iii) regulation by a common gene. However, for the construction of a gene association network only the direct interactions are of interest as only these correspond to edges between two nodes (genes) in the resulting graph.

In the GGM framework the strength of direct pairwise correlation is characterized by the *partial* correlation matrix $\Pi = (\pi_{ij})$. These coefficients describe the correlation between any two genes i and j conditioned on all the remainder of the genes. For instance, the partial

correlation π_{12} between genes 1 and 2 is simply the correlation cor(ϵ_1, ϵ_2) of the residuals ϵ_1 and ϵ_2 resulting from linearly regressing gene 1 and gene 2 against genes 3 to G, respectively. Standard graphical model theory (e.g. Edwards, 1995) shows that the matrix Π is related to the *inverse* of the standard correlation coefficients P. This leads to a straightforward procedure to compute Π via the relations

$$\Omega = P^{-1} = (\omega_{ij}) \tag{1}$$

and

$$\pi_{ij} = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}. \tag{2}$$

Note that in the inversion step (Eq. 1) it is equally valid to use the covariance matrix Σ instead of the correlation matrix P.

The partial correlation coefficients allow for a number of further interpretations. As the multivariate Normal distribution is closed under marginalization and conditioning, the partial correlation π_{ij} is the correlation coefficient of the conditional bivariate distribution for genes i and j. Furthermore, assuming normality it can be shown that two variables are *conditionally independent* given the remaining variables if and only if the corresponding partial correlation vanishes. Equivalently, the conditional independence graph of a jointly normal set of random variables is determined by the location of zeros in the inverse correlation matrix Ω (Whittaker, 1990).

In order to reconstruct a GGM network from a given data set the following procedure is typically employed. First, an estimate of the correlation matrix P is obtained, usually via the unbiased sample covariance matrix $\hat{\Sigma} = (\hat{\sigma}_{ij}) = \frac{1}{N-1}(X-\bar{X})^T(X-\bar{X})$ followed by standardization. Second, estimates of partial correlation coefficients are computed from the sample correlation matrix using Eqs. 1 and 2. Third, statistical tests are employed to determine which entries in the estimated partial correlation matrix $\hat{\Pi}$ are significantly different from zero. Finally, the inferred correlation structure is visualized by a graph, with edges corresponding to non-zero partial correlation coefficients.

However, this algorithm is only applicable if the sample size N is larger than the number of variables G. Otherwise, the sample covariance matrix is not positive definite and cannot be inverted (e.g. Friedman, 1989; Hastie and Tibshirani, 2004). This in turn prevents the direct computation of the partial correlation coefficients. Unfortunately, this is the case for typical microarray data where one has a data situation with $N \ll G$. In addition, the small sample size also renders most standard statistical tests for GGMs invalid, as these usually rely on a large sample size N for asymptotic validity.

Estimating Partial Correlation From Small Samples

In order to obtain reliable *small-sample point estimates* of partial correlation coefficients we propose two conceptually simple but effective variations of the standard graphical Gaussian modeling framework. First, when inverting the estimated correlation matrix $\hat{P}$ we employ the Moore-Penrose pseudo-inverse. Second, we use bootstrap aggregation (bagging) to stabilize the estimator.

The Moore-Penrose pseudo-inverse (Penrose, 1955) is a generalization of the standard matrix inverse that can also be applied to singular matrices and that is based on the singular value decomposition (SVD). The correlation matrix P can be decomposed into $P = U\,D\,V^T$ where D is a square diagonal matrix of rank $m \leq \min(N, G)$ containing all non-vanishing singular values. The pseudo-inverse P^+ is then defined as $P^+ = U\,D^{-1}\,V^T$ and requires only the trivial inversion of D. It can be shown that the pseudo-inverse P^+ is the shortest length least-squares solution of $PP^+ = I$, and hence reduces to the standard matrix inverse where possible.

Bootstrap aggregation (Breiman, 1996) is a simple and very general approach to improve upon an unstable estimator $\hat{\theta}(y)$ for a given set of data y. The algorithm proceeds as follows:

1. Generate a bootstrap sample y^{*b} with replacement from the original data. Repeat this process for each $b = 1, \ldots, B$ independently (e.g. with $B = 1000$).
2. For each data sample y^{*b} calculate the estimate $\hat{\theta}^{*b}$.
3. Compute the bootstrap mean $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}$ to obtain the bagged estimate.

In a nutshell, bagging is essentially a variance reduction method. Another interpretation of the bagged estimate is as an approximate Bayesian posterior mean estimate (Hastie et al., 2001).

Both these techniques combined allow to construct a small-sample estimator of the partial correlation matrix $\Pi = (\pi_{ij})$. In particular, in this paper we consider the following three possibilities:

$\hat{\Pi}^1$**:** Use the pseudo-inverse for inverting the sample correlation matrix $\hat{P}$ to obtain an estimate of Π, without performing any form of bagging (= "observed partial correlation").

$\hat{\Pi}^2$**:** Use bagging to estimate the correlation matrix P, then invert the bagged correlation matrix with the pseudo-inverse to obtain an estimate of Π (= "partial bagged correlation").

$\hat{\Pi}^3$**:** Apply bagging to the estimator $\hat{\Pi}^1$, i.e. use the pseudo-inverse for inverting each bootstrap replicate estimate $\hat{P}^{*b}$, then average the results (= "bagged partial correlation").

By construction all three of these estimators can be applied to cases where the sample size is smaller than the number of variables. However, they differ drastically with respect to accuracy and power. This is investigated in detail below using computer simulations (see section *Results*).

Null Distribution of Sample Partial Correlation

To address the statistical testing problem of non-zero partial correlation

$$H_0 : \pi_{ij} = 0 \quad \text{versus} \quad H_1 : \pi_{ij} \neq 0, \tag{3}$$

we require the *sampling distribution* of $\hat{\pi}_{ij} = p_{ij}$ under the null hypothesis $\pi_{ij} = 0$ (for convenience, we drop the subscripts i and j in the following).

From Hotelling (1953) the distribution of the sample normal correlation coefficient $\hat{\rho} = r$ is known exactly. For $\rho = 0$ we have

$$f_0(r;\kappa) = (1-r^2)^{(\kappa-r)/2} \frac{\Gamma(\frac{\kappa}{2})}{\pi^{\frac{1}{2}}\Gamma(\frac{\kappa-1}{2})}\,, \tag{4}$$

where κ is the degree of freedom. For the standard correlation coefficient the degree of freedom $\kappa = N-1$ is determined by the sample size N. For $\rho = 0$ the variance of r also equals the inverse of κ, i.e. $\mathrm{Var}(r) = \frac{1}{\kappa}$.

The sample normal *partial* correlation coefficient $\hat{\pi} = p$ is distributed precisely as the standard correlation coefficient $\hat{\rho} = r$, only that κ is reduced by the number of eliminated variables (Hotelling, 1953). Thus, if there are G variables (of which $G-2$ have to be eliminated in order to compute the pairwise partial correlation coefficients) the resulting degree of freedom is $\kappa = N - G + 1$. Note that this relationship implies that N cannot be smaller than G if κ is to remain positive.

In a small-sample setting we cannot use the standard partial correlation estimate $\hat{\Pi}$ (Eq. 2) but rather have to rely on alternative estimators such as $\hat{\Pi}^1$, $\hat{\Pi}^2$, $\hat{\Pi}^3$ suggested above. Unfortunately, we cannot analytically derive the sampling distributions of these estimators. However, it can be shown numerically (see section *Results* for details) that their respective simulated sampling distributions still assume the distributional form of Eq. 4, albeit with a smaller variance and hence with $\kappa > 0$ even for $N < G$. Note that in this case the degree of freedom κ is not a simple function of N and G but rather has itself to be estimated from the data.

Robbins-Efron-Type Inference of Empirical Test Distribution

In principle, given an appropriate choice of κ, Eq. 4 allows to compute p-values for estimated partial correlation coefficients and thus to perform statistical testing with regard to the presence of edges in a GGM network.

As we do not have repeated estimates of the partial correlation coefficient per individual edge it is not trivial to estimate the degree of freedom κ. However, we can utilize the highly parallel structure of the edge testing problem and the fact that biomolecular networks are typically sparse (Yeung et al., 2002). In a network considering G genes there is a large number $E = G(G-1)/2$ of possible edges. Only a small fraction η_A of these will correspond to true edges, whereas for the remaining majority the corresponding true partial correlation coefficients will vanish.

Therefore we may assume that the observed partial correlation coefficients p *across all edges* in the network follow a mixture distribution

$$f(p) = \eta_0 f_0(p;\kappa) + \eta_A f_A(p)\,, \tag{5}$$

where η_0 and η_A are the priors for the null and alternative distribution, f_0 and f_A, respectively, with $\eta_0 + \eta_A = 1$ and $\eta_0 >> \eta_A$. The null distribution f_0 is given by Eq. 4. For reasons of simplicity we assume here for the distribution of partial correlation coefficients of the true edges f_A a simple uniform distribution from -1 to 1. Note that for f_A other more complicated distributions could easily be conceived, including non-parametric estimates.

Fitting this mixture distribution to the observed partial correlation coefficients (via optimizing the corresponding likelihood function or an EM-type algorithm) allows to infer the parameters $\hat{\eta}_0$ and $\hat{\kappa}$. It is then straightforward to compute two-sided p-values for each possible edge in the corresponding network using the exact null distribution f_0 with $\hat{\kappa}$ as plug-in estimate. Alternatively, one may also be interested in computing

$$\mathrm{Prob}(\text{non-zero edge}|p) = \frac{\hat{\eta}_A f_A(p)}{f(p;\hat{\kappa})}\,, \tag{6}$$

i.e. the empirical posterior probability of an edge being present.

This approach, though new for edge detection in graphical models, is directly inspired by similar approaches to detect differentially expressed genes (Sapir and Churchill, 2000; Efron et al., 2001; Efron, 2003). There, the mixture distribution models differentially expressed genes assuming that the majority of investigated genes is not differentially expressed.

A key element of this procedure is that it turns a seemingly disadvantage in the analysis, namely the large

number of genes G in a microarray data set, into an advantage: with growing G the number of zero-edges $\eta_0 E$ becomes larger, and hence it gets easier to estimate the null distribution from the data. Note that this "Robbins-Efron-type" inference (see Efron, 2003) enables one to determine the sampling distribution f_0 from a large-dimensional point estimate (!). A further benefit of using an empirical null distribution in a large-scale testing situation is that it also additionally accounts for hidden correlations and unobserved covariates (Efron, 2004).

Finally, we note that using the estimated degree of freedom $\hat{\kappa}$ we can define an effective sample size $N_{\text{eff}} = \hat{\kappa} + G - 1$. This reflects the relationship between sample size and κ for the standard Normal partial correlation coefficient, but also extends to the case when other estimators such as $\hat{\Pi}^1$, $\hat{\Pi}^2$, and $\hat{\Pi}^3$ are employed.

Selection of Graphical Gaussian Model Using False Discovery Rate Multiple Testing

One simple strategy for choosing a GGM network consistent with the data is to test each of the $E = G(G-1)/2$ potential edges individually for presence in the final network, i.e. whether the corresponding partial correlation coefficient is significantly different from zero (Whittaker, 1990; Drton and Perlman, 2004). This proceeds as follows. First, a list of p-values $p_1, p_2, \ldots, p_E$ is calculated, one for each edge. Subsequently, because of the parallel testing situation a multiple testing procedure needs to be applied.

Here we employ the method of false discovery rate (FDR) multiple testing (Benjamini and Hochberg, 1995). FDR controls the expected proportion of false positives out of the total number of rejections rather than the chance of any false positives. This makes it ideal for screening purposes (Storey and Tibshirani, 2003). The basic algorithm is as follows:

1. Construct the set of ordered p-values $p_{(1)}, p_{(2)}, \ldots, p_{(E)}$ with corresponding edges $e_{(1)}, e_{(2)}, \ldots, e_{(E)}$.
2. Then let i_Q be the largest i for which $p_{(i)} \leq \frac{i}{E}\frac{Q}{\eta_0}$.
3. Finally, reject the null hypothesis of zero partial correlation for all edges $e_{(1)}, e_{(2)}, \ldots, e_{(i_Q)}$.

It can be shown that this procedure controls the false discovery rate at level Q (Benjamini and Hochberg, 1995; Storey, 2002). Moreover, FDR is justified both from a frequentist as well as from a Bayesian perspective (Efron et al., 2001; Storey, 2002; Efron, 2003). Note that the above decision rule also requires the specification of η_0, the fraction of true zero partial correlations. This parameter is either set to one, the most conservative choice as done by Benjamini and Hochberg (1995), or it may be estimated adaptively from the data (Benjamini and Hochberg, 2000; Storey, 2002). In our case a suitable estimate $\hat{\eta}_0$ is available from the fit of Eq. 5.

Using a multiple testing procedure for GGM selection has the advantage that it is practical and computationally efficient also for large numbers of genes. Nevertheless, we are well aware that this is a heuristic and only an approximation to an exhaustive GGM search. Unfortunately, the number of possible network topologies grows super-exponentially with the number of nodes. Thus, an exhaustive network enumeration is necessarily limited to toy cases only. Other heuristic searches such as backward and forward selection (Whittaker, 1990) do not necessarily guarantee a better fit for large G than multiple testing (Drton and Perlman, 2004). However, stochastic searches such as Bayesian MCMC sampling of GGMs may prove more effective, see Wong et al. (2003) and Dobra et al. (2004) for recent developments.

Recipe of Analysis and Computer Program

In a nutshell, our suggested framework for inferring large graphical Gaussian models from small-sample data comprises the following steps:

1. Choose a suitable point estimator of partial correlation (one of $\hat{\Pi}^1$, $\hat{\Pi}^2$, $\hat{\Pi}^3$), see simulation study and our recommendations in the section *Results*.
2. Compute partial correlation estimates for each possible edge.
3. Estimate the degree of freedom κ by fitting the mixture distribution from Eq. 5.
4. Compute two-sided p-values and posterior probabilities for each edge.
5. Use FDR multiple testing for selection of edges to be included in the GGM.
6. Visualize the resulting network structure.

We have implemented this approach in the R package "GeneTS" (versions 2.0 and later). It is distributed under the terms of the GNU General Public License and freely available from http://www.stat.uni-muenchen.de/~strimmer/genets/, from the R package archive (http://cran.r-project.org) and from the Bioconductor web page (http://www.bioconductor.org)

Visualization of the inferred networks requires additional installation of the Bioconductor R packages "Rgraphviz" by Jeff Gentry and "graph" by Robert Gentleman (see Figs. 1 and 7 for examples).

RESULTS

In order to investigate the statistical properties of the proposed framework to inferring graphical Gaussian models from small samples we conducted a series of extensive computer simulations. Subsequently, we reanalyzed molecular data from a microarray study of breast

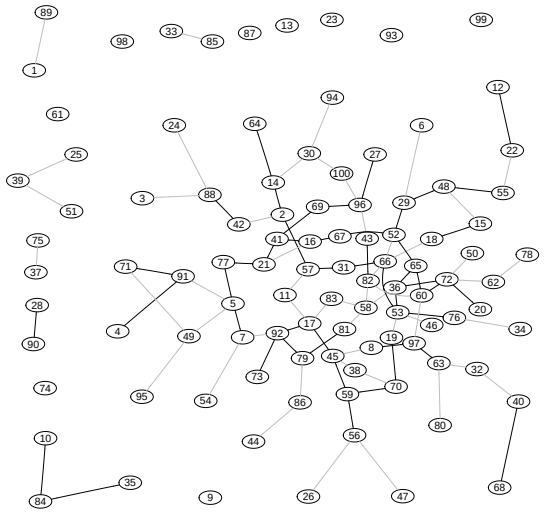

Fig. 1. Simulated sparse network with $G = 100$ nodes and 99 edges (corresponding to an edge fraction of $\eta_A = 0.02$). Note that in this figure branch lengths are purely due to the layout of the graph and do not indicate the strength of the correlation between two connected nodes. Grey lines indicate negative partial correlation, whereas edges with positive correlation are drawn in black.

Table 1. Definition of quantities used for assessing GGM network reconstruction.

Quantity	Definition
Number of true edges:	$TP + FN = \eta_A E$
Number of zero-edges:	$TN + FP = \eta_0 E$
Significant edges:	$TP + FP = S$
False positive rate:	E($FP/(\eta_0 E)$) $= \alpha_I$
False negative rate:	E($FN/(\eta_A E)$) $= \alpha_{II}$
True negative rate: (specificity)	$1 - \alpha_I$
True positive rate: (sensitivity, power)	$1 - \alpha_{II}$
Positive predictive value:	PPV $=$ E$(TP/S \mid S > 0)$
False discovery rate:	FDR $=$ E$(FP/S \mid S > 0)$

cancer tissue samples (West et al., 2001) and inferred a corresponding large-scale gene association network.

Simulation Setup

In our analysis of simulated data we used the following approach to generate random graphical models and data. It allows to control parameters of interest such as the number of nodes G, the fraction of non-zero edges η_A, and the sample size N of the simulated data.

First, partial correlation matrices Π were generated by an algorithm that guarantees that the resulting matrices are always positive definite. This method proceeds as follows:

1. Start with an empty, symmetric $G \times G$ matrix (with zero diagonal elements).
2. Choose randomly the off-diagonal positions corresponding to the $\eta_A E$ non-zero edges, and fill in preliminary correlation values drawn from the uniform distribution between -1 and 1.
3. Compute column-wise sums of the absolute values of the matrix entries, and set the corresponding diagonal element equal to this sum plus a small constant (say, 0.0001). This ensures that the resulting matrix is diagonally dominant, and thus positive definite.
4. Standardize the matrix so that the diagonal entries all equal 1 to obtain the simulated true partial correlation matrix Π which in turn represents the true GGM.

An example of a simulated network with $G = 100$ nodes and $\eta_A = 0.02$ is shown in Fig. 1. This choice of G and η_A implies that there are 99 true edges out of 4,950 potential edges. Note that even for small values of η_A the resulting "sparse" network still looks quite dense. This is because the number of available edges E grows with the square of the number of variables G.

Second, simulated data of the desired sample size N were generated as follows. From Π the true pairwise correlation matrix P was computed via reverse application of Eqs. 2 and 1. As Π is positive definite, so is its inverse and the corresponding matrix P. Subsequently, samples of length N were drawn from the multivariate normal distribution with mean zero and the correlation structure P.

In the next step, the simulated data was used to obtain point estimates $\hat{\Pi}^1, \hat{\Pi}^2$, and $\hat{\Pi}^3$. These were in turn compared to the original true matrix Π. As measure of the accuracy of the point estimates we employed the squared error loss $L(\hat{\Pi}^i, \Pi) = ||\hat{\Pi}^i - \Pi||_F^2 = \sum_{i,j} (\hat{\pi}_{ij}^k - \pi_{ij})^2$. The expected loss (risk), or mean squared error (MSE), was estimated by averaging $L(\hat{\Pi}^i, \Pi)$ over multiple simulation runs.

Then, after fitting the mixture distribution, we used the estimate $\hat{\kappa}$ to compute the effective sample size via $\hat{N}_{\text{eff}} = \hat{\kappa} + G - 1$. Finally, to assess the network reconstruction by multiple testing of edges we counted true

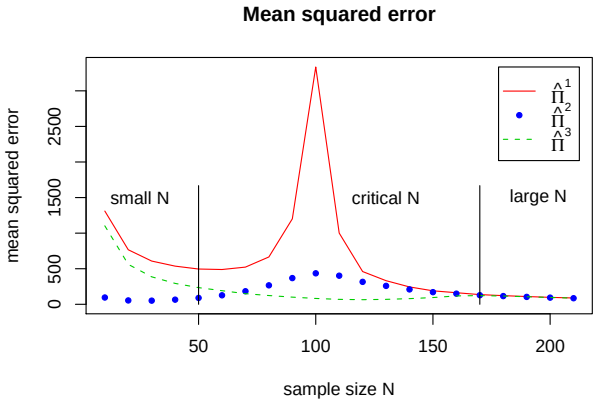

Fig. 2. Mean squared error of the three small-sample estimators $\hat{\Pi}^1, \hat{\Pi}^2$, and $\hat{\Pi}^3$ in dependence of sample size for $G = 100$ genes. The areas designated "small N", "critical N", and "large N" are defined relative to the number of genes G.

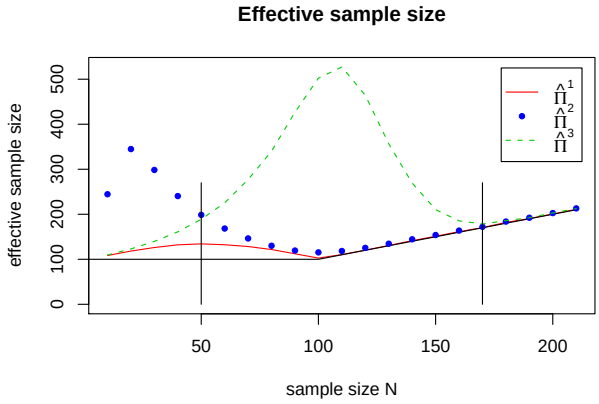

Fig. 3. Effective sample size N_{eff} for the three investigated small-sample point estimators of partial correlation in dependence of true sample size for $G = 100$ genes.

positives TP (correctly identified true edges), false positives FP (spurious edges, i.e. not recognized zero-edges), true negatives TN (correctly identified zero-edges), and false negatives FN (not recognized true edges). From this information we estimated the true negative rate (specificity) and the true positive rate (sensitvity), see Table 1 for the definitions of these quantities. We also computed estimates of the positive predictive value, i.e. the expected fraction of true edges among all significant edges.

Analysis of Simulated Data

Accuracy of Point Estimates First, we investigated the accuracy of the point estimators $\hat{\Pi}^1, \hat{\Pi}^2$, and $\hat{\Pi}^3$ to recover the true partial correlation matrix Π in dependence of the sample size N.

We varied the network parameters so that $N = 10, 20, \ldots, 210$, $G = 20 - 210$, and $\eta_A = 0.01 - 0.2$. For a fixed network size with a given proportion of non-zero edges η_A we randomly generated GGMs and simulated data as described above. The number of bootstrap replicates for bagging was set at $B = 1000$, and we conducted $R = 50$ simulations for each setting of N and G.

Fig. 2 shows as an example the graphs resulting from simulations run with $G = 100$, $\eta_A = 0.02$, and $N = 10, 20, \ldots, 210$. The same qualitative results were also obtained with all other investigated combinations of G and η_A (data not shown). The most striking result from these simulations is the existence of three different regions ($N \ll G$, $N \approx G$, and $N \gg G$) where all three estimators exhibit very different properties.

For large samples $N \gg G$ the point estimators $\hat{\Pi}^1, \hat{\Pi}^2$, and $\hat{\Pi}^3$ mainly agree with each other, with the same low error. Note that this is the only region where "classical" graphical Gaussian modeling theory is valid.

On the other end, for very small $N \ll G$ the best point estimate is clearly obtained by $\hat{\Pi}^2$. This can be explained as follows: $\hat{\Pi}^2$ is the only one of the three investigated estimators that is based on a positive definite estimate of the correlation matrix, as averaging over bootstrap sample correlation matrices $\hat{P}^{*b}$ acts as implicit regularization procedure (cf. Friedman (1989)). Also note that $\hat{P}$ is unbiased and hence $E(\hat{P}) = P \approx \frac{1}{B}\sum_{b=1}^{B} \hat{P}^{*b}$. The benefit is that the subsequent matrix inversion to obtain $\hat{\Pi}^2$ can proceed with little loss of accuracy.

In the "critical N" zone with $N \approx G$ a striking dimensionality resonance effect (Raudys and Duin, 1998; Skurichina and Duin, 2002) is observed. The mean squared error of $\hat{\Pi}^1$ increases dramatically around $N \approx G$, with *decreasing* error when the sample size decreases. This "peaking phenomenon" is well known in small-sample regression and classification problems and is due to the use of the pseudo-inverse (Raudys and Duin, 1998). It can be understood as follows. For $N \approx G$ the eigenvalues of the sample correlation matrix are strongly distorted in comparison with those of the true correlation matrix, so that the largest and smallest eigenvalues are strongly biased (e.g. Friedman, 1989). This causes the corresponding SVD directions in the pseudo-inverse to become highly overestimated. Regularization of the correlation matrix (for example by bagging) reduces this error dramatically (Skurichina and Duin, 2002). This

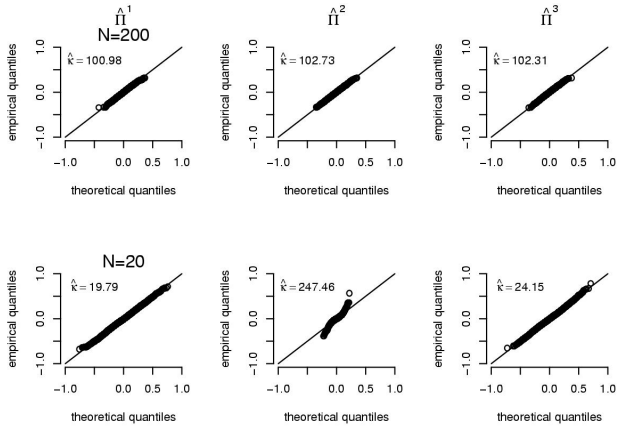

Fig. 4. Quantile-quantile plots of the observed null distribution of $\hat{\Pi}^1$, $\hat{\Pi}^2$, and $\hat{\Pi}^3$ for $G = 100$ genes. *Top row:* large sample size ($N = 200$). *Bottom row:* small sample size ($N = 20$).

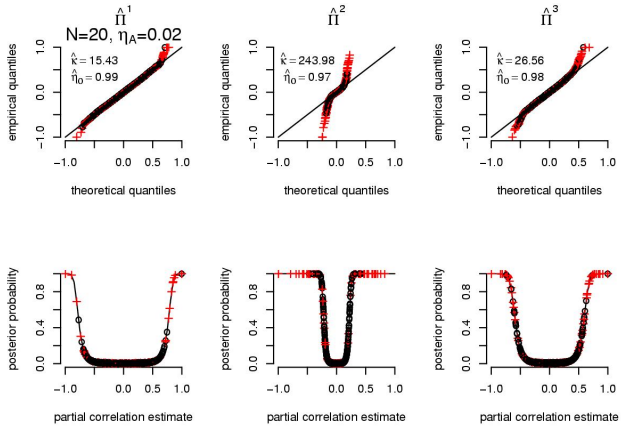

Fig. 5. *Top row:* Quantile-quantile plots for the observed mixture distributions with $N = 20$, $G = 100$, and $\eta_A = 0.02$. *Bottom row:* The corresponding empirical posterior probability plots.

can be immediately seen by comparing $\hat{\Pi}^1$ with the two bagged estimators $\hat{\Pi}^2$ and $\hat{\Pi}^3$ that both demonstrate a very good performance in the "critical N" zone and exhibit a considerably lower error than $\hat{\Pi}^1$.

Effective Sample Size Next, we conducted a further simulation study, similar in set-up as above to study the dependency of the effective sample size $N_{\text{eff}} = \hat{\kappa} + G - 1$ from the actual sample size. The results from an example run with $G = 100$ nodes are shown in Fig. 3.

A number of things can be learned from this figure. First, the effective sample size N_{eff} is always greater than the number of variables G, regardless of the actual sample size. This is noteworthy especially for small sample sizes $N \ll G$. Second, whenever the effective sample size N_{eff} is large then the mean squared error is small (see Fig. 2). This is particularly pronounced for estimator $\hat{\Pi}^2$ in the "small N" zone and for estimator $\hat{\Pi}^3$ in the "critical N" zone. Finally, as a large N_{eff} implies a large $\hat{\kappa}$ we note that the variance of the null distribution decreases with growing effective sample size. This is an important criterion when choosing an appropriate estimator (see subsection below for some suggestions).

Validation of Null Distribution In further studies we verified that under the null hypothesis of no partial correlation the three proposed small-sample estimators $\hat{\Pi}^1$, $\hat{\Pi}^2$, and $\hat{\Pi}^3$ do indeed follow the theoretical distribution suggested in Eq. 4. This is important to avoid systematic bias in the statistical testing of edges.

In Fig. 4 we show example quantile-quantile plots comparing the empirical with the theoretical null distribution for large (N=200, top row) and for small ($N = 20$, bottom row) sample size. In each case the data were simulated assuming $G = 100$ genes and an empty "network" with no edges as underlying model.

The first row of Fig. 4 shows that, as expected, for large N the observed correlation coefficients all fit the theoretical null distribution very well. The estimates of the degree of freedom κ are also broadly equivalent across the three estimators $\hat{\Pi}^1$, $\hat{\Pi}^2$, and $\hat{\Pi}^3$. Note that $\hat{\Pi}^1$ is identical to the classic partial correlation estimator for $N = 200$, and accordingly the corresponding estimate $\hat{\kappa}$ matches the theoretically expected value $\kappa = 101$.

For comparison, in the second row of the same figure, the quantile-quantile plots are shown for the much smaller sample size $N = 20$. For both $\hat{\Pi}^1$ and $\hat{\Pi}^3$ clearly the observed null distributions still fit the theoretical distributions well. The plot for $\hat{\Pi}^2$ indicates a stronger curtosis and slightly broader tails of the empirical compared to the theoretical distribution. Nevertheless, the fit between theoretical and empirical distribution is still good.

One further point to note is that for small samples the variability of partial correlation estimates and the estimated degrees of freedom $\hat{\kappa}$ differ considerably among the investigated estimators. For $N = 20$ and $G = 100$ the estimator $\hat{\Pi}^2$ exhibits the by far smallest variance and largest $\hat{\kappa}$.

Fit of Mixture Distribution Subsequently, we also checked the fit of the mixture distribution (Eq. 5) in the presence of true non-zero correlations. Results from a small-sample simulation with with $N = 20$, $G = 100$, and $\eta_A = 0.02$ are displayed in Fig. 5.

The top row of Fig. 5 shows the quantile-quantile

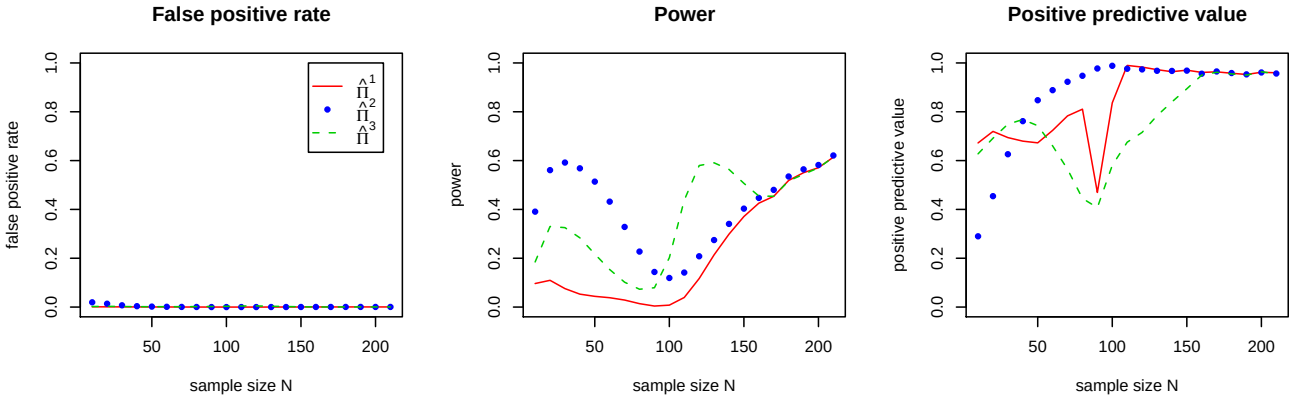

Fig. 6. Power, positive predictive value, and false positive rate for recovering the true GGM network. See Table 1 for the definitions of the investigated quantities, and the main text for the the simulation setup with $G = 100$ genes.

plots of the observed distribution of partial correlation coefficients versus the theoretical null distribution. We observe broader tails of the empirical as compared to the theoretical distribution. This is expected as in this case the empirical distribution is a mixture of the null distribution and the alternative distribution for the non-zero correlations belonging to the true edges (indicated in the plots by cross symbols). The proportion of zero-edges η_0 are estimated accurately, and the estimates of the degree of freedom κ of the null distribution are similar to the corresponding estimates for $N = 20$ in Fig. 4.

The bottom row of Fig. 5 depicts the corresponding empirical posterior probability plots (Eq. 6). The probability of an observed partial correlation to correspond to a true correlation is approximately one for large correlation strengths and quickly vanishes for smaller absolute values. Only the tails of the empirical mixture distribution contain the statistically significant edges. The width of characteristic U-shape of the posterior probability plot is determined by the degree of freedom κ of the null distribution. This shows that using an estimator with a small variance is advantageous as this allows to identify statistically significant edges even with relatively small absolute value of partial correlation.

Sensitivity and Specificity of GGM Selection Finally, we spent a large amount of computational effort on simulations to investigate the statistical properties of GGM selection using FDR multiple testing.

We conducted simulations with N ranging from 10 to 210 in steps of 10, $G = 100$ and $\eta_A = 0.02$. For $N \leq 110$ we performed $R = 500$ repetitions (i.e. simulation of GGM network and data) per sample size, whereas for reasons of computational economy only $R = 50$ repetitions were done for $N > 110$. The GGMs were inferred by multiple testing of $E = 4950$ edges with the desired FDR level fixed at $Q = 0.05$.

For each inferred network we counted the number of true positive features (i.e. the number of correctly recognized true edges) as well as the number of true negatives (i.e. the number of correctly identified zero-edges). From these raw statistics, and repeated simulations of networks and data, we obtained estimates of the false positive rate, of power, and of the positive predictive value (PPV) for $\hat{\Pi}^1$, $\hat{\Pi}^2$, and $\hat{\Pi}^3$ at a given sample size N. The precise definitions of these terms are given in Table 1. Fig. 6 summarizes our results.

All three small-sample estimators, $\hat{\Pi}^1$, $\hat{\Pi}^2$, and $\hat{\Pi}^3$, exhibit the same low empirical false positive rate regardless of N. For large $N > 170$ they also all agree in power and in positive predictive value. However, they differ drastically in the small-sample case $N < G$ and for $N \approx G$. In terms of power the bagged estimators both $\hat{\Pi}^2$ and $\hat{\Pi}^3$ consistently outperform the simple estimator $\hat{\Pi}^1$ that fares rather poorly particularly for $N < G$. In the latter region $\hat{\Pi}^2$ exhibits the overall highest power, whereas for $N \approx G$ and sample sizes slightly above G the estimator $\hat{\Pi}^3$ performs best.

The largest PPV is generally obtained by using the estimator $\hat{\Pi}^2$. However, for very small sample size the PPV of $\hat{\Pi}^2$ drops sharply; this is likely due to the imperfect fit with the theoretical null distribution (cf. Fig. 4).

A further noteworthy result from all our simulations is that close to $G = N$ there is generally very little power to infer the true network structure. This may again be a consequence of the "dimensionality resonance" phenomenon discussed above.

Finally, we would like to note that all these simulations and the resulting estimates are quite conservative. This is because we generated true GGMs in such a way that they contained edges with both strong as well as weak true correlation. The latter are notoriously difficult to detect (cf. Fig. 5) and this consequently depresses the test results.

Choice of Small-Sample Estimator

From the above analysis of simulated data it is clear that the estimators $\hat{\Pi}^1, \hat{\Pi}^2$, and $\hat{\Pi}^3$ perform very differently. As a summary we suggest the following guideline for choosing a suitable estimator:

$\hat{\Pi}^1$**:** Should only be used for $N \gg G$, otherwise it lacks statistical power. Note that in this "large N" region the other two estimators perform equally well but are computationally slower due to bagging.

$\hat{\Pi}^2$**:** Best used for small-sample applications with $N < G$. Here the main advantages of $\hat{\Pi}^2$ are its small variance (large effective sample size) and its high accuracy as a point estimate. It exhibits the overall best power in the "small N" zone. Furthermore, it is computationally less expensive than $\hat{\Pi}^3$. However, note its low PPV for very small N.

$\hat{\Pi}^3$**:** Is best used in the "critical N" zone where it offers small error and large effective sample size. For N slightly larger than G this estimator also provides the overall best power, though in terms of PPV this estimator performs less well than $\hat{\Pi}^2$.

As a result, this particularly promotes $\hat{\Pi}^2$ as estimator of choice for the inference of GGM networks from small-sample gene expression data.

Molecular Data

Breast Cancer Data Set We now illustrate the utility of the proposed empirical Bayes framework of inferring GGM networks from small samples by application to a large-scale biological data set. More specifically, we reanalyzed gene expression data from a breast cancer study described in West et al. (2001).

Preprocessing and Calibration This data set comprises 49 tissue samples and gene expression was measured for 7129 genes/probes using Affymetrix hu6800 chips. We downloaded the corresponding CEL data from the Duke University Center for Genome Technology (http://data.cgt.duke.edu/West/PNASCel1.zip). We then calibrated and normalized the raw data to obtain robust multi-array average (RMA) expression measures (Irizarry et al., 2003). This was done using the "affy" package in Bioconductor version 1.3 (http://www.bioconductor.org).

Subsequently, we removed all sequences that varied only minimally or on low levels. Specifically, we screened out genes whose expression levels across all samples varied less than two-fold (corresponding to a RMA difference less than 1.0, as RMA is a measure on the log-base 2 scale) or whose maximum RMA intensity value was less than 9.0. As a result of the prescreening gene expression data for 3,883 genes across 49 samples remained for further analysis.

Inference of Global Association Network In order to infer the global association structure and the corresponding GGM network for all 3,883 genes we employed the small-sample estimator $\hat{\Pi}^2$ with $B = 10,000$ bootstrap replications. The computation of the estimate of the partial correlation matrix –a 3,883 times 3,883 matrix with entries for 7,536,903 possible edges– required approximately 20 hours on a standard Intel Pentium 4 workstation running under the Linux operating system.

The subsequent fit of the mixture distribution (Eq. 5) resulted in an estimated degree of freedom $\hat{\kappa} = 4601.98$ with $\hat{\eta}_0 = 0.9924$. Using the FDR method with a desired level $Q = 0.05$ we determined 88,822 significantly non-zero coefficients, corresponding to a p-value cutoff of 0.0006 and a threshold of partial correlation $\hat{\pi} > 0.051$. Note that for this size of network most of the coefficients are very close to zero, so even small values are statistically significant. This is also reflected in the large value of $\hat{\kappa}$.

From a statistical perspective we caution that particularly in an extreme small-sample setting not all statistically significant edges will necessarily correspond to true edges (low PPV). To be on the conservative side, we therefore advise to take the theoretical threshold only as minimal lower bound and also to consider larger cut-off values.

CNR2 Receptor is Most-Connected Gene Because of the large number of nodes and edges it is difficult to visualize the resulting global network structure (however, see below for a discussion of a sub-network). However, the degree of connectivity of each gene is more easily amenable and also highly informative.

For example, in our inferred GGM network for the investigated breast cancer data set the cannabinoid receptor 2 gene (CNR2), also known as CB2 receptor, is the best-connected gene, as it contains significant correlations with 75 (!) other genes. The "peripheral" cannabinoid receptor CNR2 is mostly expressed in the immune system, and unlike the "central" CNR1 receptor it is unrelated to cannabinoid psychoactivity.

The existence of such "super hubs" in genetic networks is well known (e.g. Barabási, 2004). The interesting point about CNR2 is that it seems to be directly involved in controlling tumor growth. It has been characterized as putative oncogene for acute myeloid leukemia (Jorda et al., 2003).

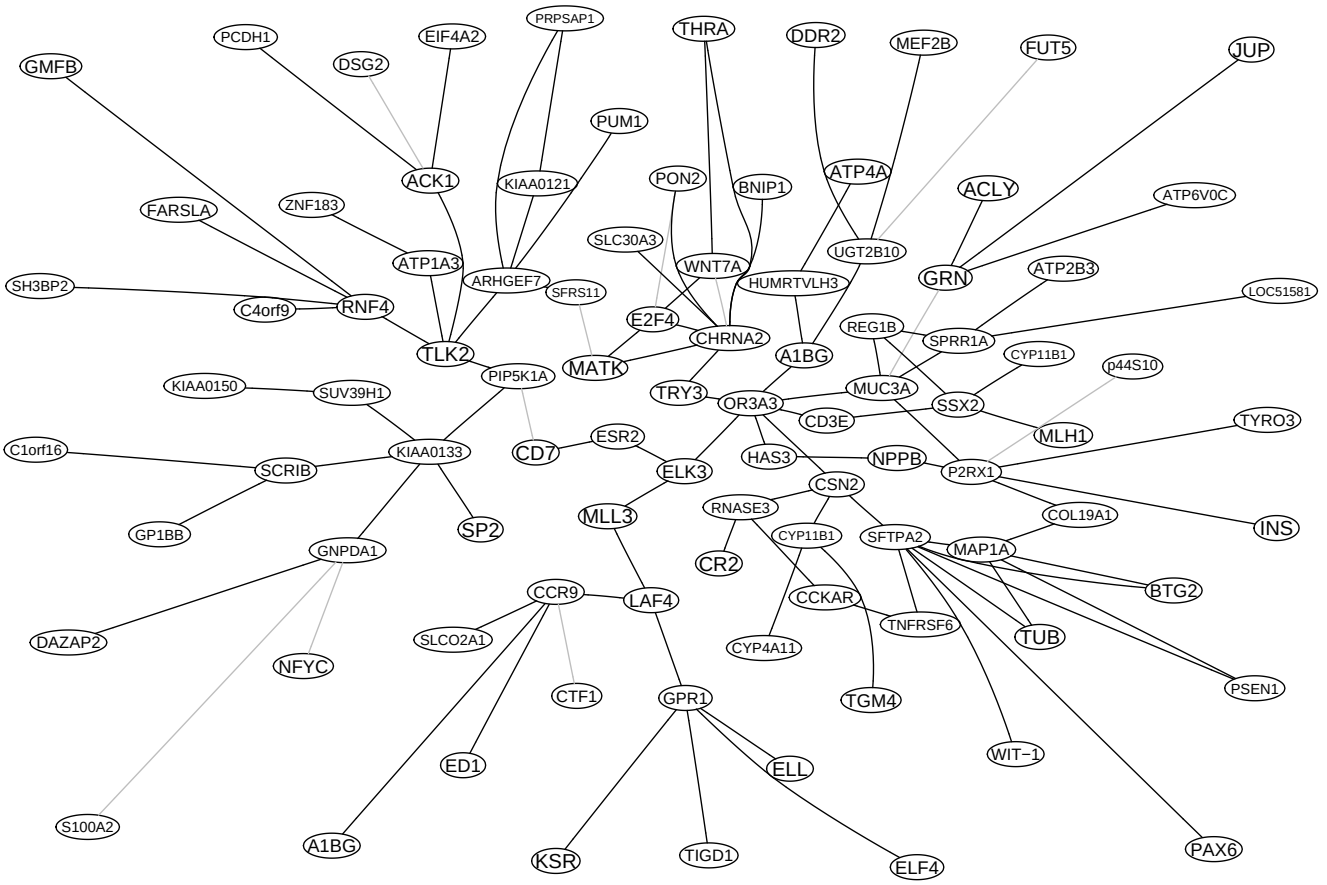

Fig. 7. Sub-network consisting of 96 genes centered around the ESR2 gene. This net was extracted from a global network with $G = 3,883$ genes reconstructed from the breast cancer data of West et al. (2001) using the small-sample estimator $\hat{\Pi}^2$. For a biological interpretation of selected genes neighboring ESR2 see the main text.

In addition, it has been shown that targeting CNR2 can lead to induction of apoptosis in malignant lymphoblastic disease (McKallip et al., 2002). Furthermore, stimulation of CNR2 leads to a regression of skin cancer tumors (Casanova et al., 2003).

Sub-Network of the ESR2 Gene For further illustration of the complexity of the inferred global network we now briefly describe the genes in the immediate surroundings of the ESR2 gene (the estrogen receptor 2). This gene was selected as "seed gene" for the sub-network because of its role in the pathobiology of breast cancer tumors (e.g. West et al., 2001). In Fig. 7 all 95 genes are shown that are correlated with ESR2 through at most five links. To reduce noise in this figure only edges with partial correlations with $\hat{\pi} > 0.13$ are shown. Interestingly, many close neighbors of ESR2 in this sub- network are known to be implicated in the development of malignant neuroplastic disease.

For example, ELK3 (also known as ERP, NET or SAP2) belongs to the Ets family of transcription factors. Ets proteins have been implicated in regulation of gene expression during a variety of biological processes, including growth control, transformation, and T-cell activation in many organisms. Loss of normal control is often associated with conversion to an oncoprotein (Wasylyk et al., 1993).

On the left to the ESR2 gene sits the human CD7 antigen (also known as gp40) which is a cell surface glycoprotein found on thymocytes and mature T-cells. CD7 is one of the earliest antigens to appear on cells of the T-lymphocyte lineage, and the most reliable clinical marker of T-cell acute lymphocytic leukemia (Aruffo and Seed, 1983).

The MLL3 gene, directly linked in our network with ELK3 and LADF4, is a member of the TRX/MLL gene family. It is associated with leukemia and developmental defects (Ruault et al., 2002).

Further down in the network one finds LAF4, a gene responsible for lymphocyte differentiation. Joint with MLL it is involved in lymphoblastic leukemia (von Bergh et al., 2002).

Many more genes depicted in Fig. 7 are related to the development of cancer (see, e.g., the CancerGene database at http://caroll.vjf.cnrs.fr/cancergene/). Hence, we are cautiously optimistic that the inferred correlation network may indeed be useful as a starting point from which to generate further medical and biochemical hypotheses.

DISCUSSION

Key Contributions and Novel Aspects

In this paper we have introduced a conceptually simple yet versatile and computationally fast framework for estimating large graphical Gaussian models from data sets of small sample size. The development of this approach was motivated by the challenge of inferring genetic networks from today's microarray data which typically contain only relatively few sample points compared to the number of investigated genes. This will continue to be an important issue also in the future: sample size is primarily restricted by the availability of tissue samples, and is not necessarily increased by improved technology.

Our framework relies on three key components:

1. Recognizing that small sample inference requires explicit regularization, we propose several new estimators of partial correlation. In particular, we employ a combination of singular value decomposition and bagging in order to compute improved coefficients (this corresponds to 0th and 1st order regularization, respectively).
2. We present an empirical Bayes approach to detect statistically significant edges. This allows to infer from the high-dimensional point estimate of partial correlations the exact null distribution needed for statistical testing, and also exploits the sparse degree of connectivity in real genetic networks. In microarray analysis a similar approach is already successfully being used to detect differential expression (Efron, 2003, 2004).
3. We suggest a heuristic to perform approximate model (network) selection using multiple testing using the false discovery rate method.

To our knowledge the present method is the first that uses an exact distribution (i.e. one that is valid for finite N) to test and infer GGMs on the gene-level from short microarray data. Thus our approach may be regarded as an extension of earlier work by Waddell and Kishino (2000), Toh and Horimoto (2002a,b), Bay et al. (2002), and Wu et al. (2003). Furthermore, we have conducted extensive simulations to investigate the performance of the proposed approach in dependence of sample size. These appear to be notably absent from many previous studies, as pointed out before by Husmeier (2003). In addition, we have verified our method by application to a realistic large-scale problem. We note that in contrast to a related MCMC approach by Dobra et al. (2004) our method can be run on low-cost PC hardware (no parallel cluster needed).

Review of GGM Model Assumptions

Our approach contains a number of implicit assumptions that need to be critically assessed.

First, GGMs are based on multivariate normality. Generally, this appears to be unproblematic given that calibration and normalization procedures are routinely used to preprocess gene expression measurements.

Second, more critical is the assumption of linear relationships among the investigated variables. While this may be a good approximation in many cases, we are well aware that a GGM has limited representational power if non-linear or combinatorial effects are present in the data. There are approaches that allow to test for deviations from linear models (Cox and Wermuth, 1994) but for small samples this may turn out to be very difficult.

Third, there may be (linear) higher-order interactions among more than two variables. GGMs in general model higher-order dependencies via the notion of cliques (i.e. fully connected groups of nodes). However, our heuristic model search using multiple testing is based on evaluating pairwise interaction only. Nevertheless, cliques can still occur in the inferred network, hence our approach will at least approximately detect higher-order effects.

Relation to Other Probabilistic Approaches for Modeling Genetic Networks

GGMs belong to the large class of linear graphical models (e.g. MacKay, 2003). Note that most other statistical methods for inferring genetic networks also fall into this group (e.g., D'haeseleer et al., 2000; Bay et al., 2002; De Hoon et al., 2003; Wu et al., 2003; Rangel et al., 2004; de la Fuente et al., 2004). Nevertheless, the important issue of regularization in the presence of small samples has only been discussed in a handful of papers (van Someren et al., 2001; Yeung et al., 2002; Liao et al., 2003; Dobra et al., 2004). One of the purposes of this paper is to further draw attention to this problem.

During the review process a referee has repeatedly pointed out that Bayesian networks are superior to GGMs as in theory the former allow to model non-linear relationships. If a lot of data are available, this is certainly true. In practice however, owing to the paucity of the data at hand, it is not generally possible to infer these non-linearities nor the global network structure (Husmeier,

2003; Friedman and Koller, 2003). Furthermore, the often exercised discretization causes information loss and might considerably influence the obtained results. Moreover, often Bayesian networks are in fact also linearized, which for time series data turns them into linear state-space models (Murphy, 2002).

Here, we simply argue that to model gene association and dependency on small-sample data sets it is prudent to choose a graphical model (such as a GGM) that requires very few assumptions and only a minimal number of parameters. Note that we do not endorse GGMs as the "true model" for genetic networks.

Challenges and Outlook

There are many directions that may be considered for further research. We believe that particularly three points are of prime importance.

First, the present approach needs to be properly adopted to time series data. While part of the longitudinal correlation will be accounted for by the empirical fit of the null-distribution, explicit dynamic and temporal elements in the model will be crucial for inferring directed relationships. GGMs have been generalized to time series models (e.g. Dahlhaus, 2000), and there are many other graphical models for time series data (e.g., Murphy, 2002; Rangel et al., 2004).

Second, for all of the above mentioned models it will be crucial to study more intensively appropriate regularization procedures. We are currently investigating a variety of methods that may lead to a better fit of the null distribution, and thus enhance statistical testing of edges.

Third, more research needs to be done in the field of model selection for gene association networks. In particular, the quality of search heuristics such as the one presented in this paper should be compared thoroughly with solutions obtained with exact approaches (only possible for small examples) and with those from the various proposed stochastic searches (e.g. Wong et al., 2003).

In conclusion, we find that the graphical modeling framework is a suitable statistical approach to modeling molecular genetic networks, but inference and appropriate model selection for small-sample data remains challenging. Our approach based on GGMs aims to be particularly simple and computationally efficient. We hope that it may serve as useful and practical exploratory tool and perhaps also as starting point for further development.

ACKNOWLEDGMENTS

This research was supported by an Emmy Noether research grant (STR 624/1-2,3) from the Deutsche Forschungsgemeinschaft (DFG). We thank Leonhard Held and Stefan Pilz for valuable discussion concerning the simulation of graphical Gaussian models and Jeff Gentry for help with his "Rgraphviz" library. We also thank the associate editor and the anonymous referees for many constructive comments that greatly helped to improve the manuscript.

REFERENCES

Aruffo, A. and B. Seed (1983). Molecular cloning of two CD7 (T-cell leukemia antigen) cDNAs by a COS cell expression system. *EMBO J. 6*, 3313–3316.

Barabási, A.-L. (2004). Network biology: understanding the cell's functional organization. *Nat. Rev. Genet. 5*, 101–113.

Bay, S. D., J. Shrager, A. Pohorille, and P. Langley (2002). Revising regulatory networks: from expression data to linear causal models. *J. Biomed. Informatics 35*, 298–297.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B 57*, 289–300.

Benjamini, Y. and Y. Hochberg (2000). The adaptive control of the false discovery rate in multiple hypotheses testing. *J. Behav. Educ. Statist. 25*, 60–83.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Casanova, M. L., C. Blázquez, J. Martínez-Palacio, C. Villanueva, M. J. Fernández-Acenero, J. W. Huffman, J. L. Jorcano, and M. Guzmán (2003). Inhibition of skin tumor growth and angiogenesis in vivo by activation of cannabinoid receptors. *J. Clin. Invest. 111*(1), 43–50.

Cox, D. R. and N. Wermuth (1994). Tests of linearity, multivariate normality and the adequacy of linear scores. *Applied Statistics 43*, 347–355.

Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika 51*, 157–172.

De Hoon, M. J. L., S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano (2003). Inferring gene regulatory networks from time-ordered gene expression data of bacillus subtilis using differential equations. *Pac. Symp. Biocomput. 8*, 17–28.

de la Fuente, A., N. Bing, I. Hoeschele, and P. Mendes (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* in press.

Dempster, A. P. (1972). Covariance selection. *Biometrics 28*, 157–175.

D'haeseleer, P., S. Liang, and R. Somogyi (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics 16*, 707–726.

Dobra, A., C. Hans, B. Jones, J. R. Nevins, and M. West (2004). Sparse graphical models for exploring gene expression data. *J. Multiv. Anal. 90*, 196–212.

Drton, M. and M. D. Perlman (2004). Model selection for Gaussian concentration graphs. *Biometrika 91*, 591–602.

Edwards, D. (1995). *Introduction to Graphical Modelling*. New York: Springer.

Efron, B. (2003). Robbins, empirical Bayes, and microarrays. *Annals of Statistics 31*, 366–378.

Efron, B. (2004). Large-scale simultaneous hypthesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc. 99*, 96–104.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer.*

Statist. Assoc. 96, 1151–1160.

Friedman, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc. 84*, 165–175.

Friedman, N. (2004). Inferring cellular networks using probabilistics graphical models. *Science 303*, 799–805.

Friedman, N. and D. Koller (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning 50*, 95–125.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er (2000). Using Bayesian networks to analyze gene expression data. *J. Comp. Biol. 7*, 601–620.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Hastie, T. and T. Tibshirani (2004). Efficient quadratic regularization for expression arrays. *Biostatistics 5*, 329–340.

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *J. R. Statist. Soc. B 15*, 193–232.

Husmeier, D. (2003). Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics 19*, 2271–2282.

Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res. 31*, e15.

Jorda, M. A., N. Rayman, P. Valk, E. De Wee, and R. Delwel (2003). Identification, characterization, and function of a novel oncogene: the peripheral cannabinoid receptor CB2. *Ann. N. Y. Acad. Sci. 996*, 10–16.

Kishino, H. and P. J. Waddell (2000). Correspondence analysis of genes and tissue types and finding genetic links from microarray data. *Genome Informatics 11*, 83–95.

Lauritzen, S. (1996). *Graphical Models*. Oxford: Oxford University Press.

Liao, J. C., R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury (2003). Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA 100*, 15522–15527.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

McKallip, R., C. Lombard, M. Fisher, B. R. Martin, S. Ryu, S. Grant, P. S. Nagarkatti, and M. Nagarkatti (2002). Targeting CB2 cannabinoid receptors as a novel therapy to treat malignant lymphoblastic disease. *Blood 100*(2), 627–634.

Murphy, K. P. (2002). *Dynamic Bayesian networks: Representation, Inference and Learning (PhD Thesis)*. Berkeley: Unversity of California, Computer Science Division.

Penrose, R. (1955). A generalized inverse for matrices. *Proc. Cambridge Phil. Soc. 51*, 406–413.

Rangel, C., J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani (2004). Modeling T-cell activation using gene expression profiling and state space modeling. *Bioinformatics 20*, 1361–1372.

Raudys, S. and R. P. W. Duin (1998). Expected classification error of the Fisher linear classifier with pseudoinverse covariance matrix. *Patt. Recogn. Lett. 19*, 385–392.

Ruault, M., M. E. Brun, M. Ventura, G. Roizes, and A. De Sario (2002). MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently deleted in myeloid leukemia. *Gene 284*, 73–81.

Sapir, M. and G. A. Churchill (2000). Estimating the posterior probability of differential gene expression from microarray data. Poster, Jackson Laboratory, Bar Harbor.

Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman (2003). Module networks: identifing regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics 34*, 166–176.

Skurichina, M. and R. P. W. Duin (2002). Bagging, boosting and the random subspace method for linear classifiers. *Patt. Analysis and Appl. 5*, 121–135.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B 64*, 479–498.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genome-wide experiments. *Proc. Natl. Acad. Sci. USA 100*, 9440–9445.

Toh, H. and K. Horimoto (2002a). Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics 18*, 287–297.

Toh, H. and K. Horimoto (2002b). System for automatically inferring a genetic network from expression profiles. *J. Biol. Physics 28*, 449–464.

van Someren, E. P., L. F. A. Wessels, M. J. T. Reinders, and E. Backer (2001, June). Robust genetic network modeling by adding noisy data. In *Proceeding of the Workshop on Nonlinear Signal and Image Processing (NSIP01)*. IEEE-EURASIP.

von Bergh, A. R., H. B. Beverlooand, P. Rombout, E. R. van Wering, M. H. van Weel, G. C. Beverstock, P. M. Kluin, R. M. Slater, and E. Schuuring (2002). LAF4, an AF4-related gene, is fused to MLL in infant acute lymphoblastic leukemia. *Genes Chromosomes Cancer 35*, 92–96.

Waddell, P. J. and H. Kishino (2000). Cluster inferences methods and graphical models evaluated on NCI60 microarray gene expression data. *Genome Informatics 11*, 129–140.

Wang, J., O. Myklebost, and E. Hovig (2003). MGraph: graphical model for microarray data analysis. *Bioinformatics 19*, 2210–2211.

Wasylyk, B., S. L. Hahn, and A. Giovane (1993). The Ets family of transcription factors. *Eur. J. Biochem. 211*, 7–18.

West, M., C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA 98*, 11462–11467.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

Wong, F., C. K. Carter, and R. Kohn (2003). Efficient estimation of covariance selection models. *Biometrika 90*, 809–830.

Wu, X., Y. Ye, and K. R. Subramanian (2003). Interactive analysis of gene interactions using graphical Gaussian model. *ACM SIGKDD Workshop on Data Mining in Bioinformatics 3*, 63–69.

Yeung, M. K. S., J. Tegnér, and J. J. Collins (2002). Reverse engineering gene networks using singular value decomposition and robust regression. *Proc. Natl. Acad. Sci. USA 99*, 6163–6168.