

Learning Large-Scale Graphical Gaussian Models from Genomic Data

Juliane Schäfer* and Korbinian Strimmer[†]

*Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany,
email: juliane.schaefer@stat.uni-muenchen.de

[†]Department of Statistics, University of Munich, Ludwigstrasse 33, D-80539 Munich, Germany,
email: korbinian.strimmer@lmu.de

Abstract. The inference and modeling of network-like structures in genomic data is of prime importance in systems biology. Complex stochastic associations and interdependencies can very generally be described as a graphical model. However, the paucity of available samples in current high-throughput experiments renders learning graphical models from genome data, such as microarray expression profiles, a challenging and very hard problem. Here we review several recently developed approaches to small-sample inference of graphical Gaussian modeling and discuss strategies to cope with the high dimensionality of functional genomics data. Particular emphasis is put on regularization methods and an empirical Bayes network inference procedure.

Keywords: Gene association network, systems biology, graphical Gaussian model, microarray, regularization, empirical Bayes, small-sample inference.

PACS: 02.50.Sk; 87.16.Yc; 89.75.Hc

INTRODUCTION

Graphical models [1, 2] are promising tools for the analysis of gene interaction because they allow the stochastic description of net-like association and dependency structures in complex highly structured data. At the same time, graphical models offer an advanced statistical framework for inference. In theory, this makes them perfectly suited for modeling biological processes in the cell such as biochemical interactions and regulatory activities.

However, the practical application of graphical models in systems biology is strongly limited by the amount of available experimental data. This apparent paradox arises as today's high-throughput facilities allow to investigate experimentally a greatly increased number of features while the number of samples has not, and cannot, similarly be increased. For instance, in a typical microarray data set the number of genes p will exceed by far the number of sample points n . This poses a serious challenge to any statistical inference procedure, and also renders estimation of gene regulatory networks an extremely hard problem. This is corroborated by a recent study on the popular Bayesian network method where Husmeier [3] demonstrated that this approach tends to perform poorly on sparse microarray data.

In this paper we focus on recent developments with respect to the use of graphical Gaussian models (GGMs) for modeling genome data. These are similar to the in the bioinformatics community widely applied "relevance networks" in that edges indicate some degree of correlation between two genes. However, in contrast to correlation

networks GGMs allow to distinguish direct from indirect interactions, i.e. whether gene A acts on gene B directly or via mediation through a third gene C . More precisely, GGMs are based on the concept of conditional independence. In this respect GGMs behave similarly as Bayesian networks. However, unlike the latter GGMs contain only undirected edges, hence they do not suffer from a restriction inherent in Bayesian networks, namely that they can only be applied to network graphs without feedback loops, i.e. directed cycles.

The outline of the paper is as follows. We will first give a brief overview over the concept and technical details of graphical Gaussian models and define the challenge of their application to genomic data. Currently known strategies to cope with dimensionality problems are reviewed. In this paper we highlight regularization methods and specifically present a procedure to inferring GGM gene association networks based on improved estimation of partial correlation and empirical Bayes multiple testing. The statistical properties of this approach are investigated in computer simulations and application to real data. Finally, we discuss some directions of further research.

GRAPHICAL GAUSSIAN MODELS

General Concept

In order to elucidate functional interaction, and as a basis for subsequent clustering and network inference, a popular strategy in bioinformatics is to compute the *standard Pearson correlation* between any two genes. If the correlation coefficient exceeds a certain a priori specified threshold then an edge is drawn between the appropriate genes. The resulting graph is called a “relevance network” where missing edges denote *marginal independence*. In statistical terminology this type of network model is also known as “covariance graph model”.

However, for understanding gene interaction this approach is only of limited use. For instance, a high correlation coefficient between two genes may be indicative of either (i) direct interaction, (ii) indirect interaction, or (iii) regulation by a common gene. In learning a genetic network from data we need to be able to distinguish among these three alternatives.

Therefore, for constructing a “gene association network” where only direct interactions among genes are depicted by edges, another framework is needed: “graphical Gaussian models” (GGMs). The key idea behind GGMs is to use *partial correlations* as a *measure of conditional independence* between any two genes. This overcomes the edge identifiability problems of standard correlation networks. Consequently, GGMs (also known as “covariance selection” or “concentration graph” models) have recently become a popular tool to study gene association networks.

Note that GGMs and the covariance graph models are only superficially similar approaches. However, both conceptually as well as practically they constitute completely different theories.

Classical Theory

Graphical Gaussian models are undirected graphical models [1, 4, 5]. Under this approach the observed data matrix X with n rows, corresponding to the samples, and p columns, corresponding to the genes, is considered to be drawn from a p -variate normal distribution $N_p(\mu, \Sigma)$ with some mean vector $\mu = (\mu_1, \dots, \mu_p)^T$ and positive definite covariance matrix $\Sigma = (\sigma_{ij})$, where $1 \leq i, j \leq p$. 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 $P = (\rho_{ij})$.

In the GGM framework the strength of direct pairwise correlation is characterized by the partial correlation matrix $Z = (\zeta_{ij})$. 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 ζ_{12} between genes 1 and 2 is simply the correlation $\text{cor}(\varepsilon_1, \varepsilon_2)$ of the residuals ε_1 and ε_2 resulting from linearly regressing gene 1 and gene 2 against genes 3 to p , respectively. Standard graphical modeling theory [e.g. 5] shows that the matrix Z is related to the inverse of the standard correlation coefficients P . This leads to a straightforward procedure to compute Z via the relations

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

and

$$\zeta_{ij} = -\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}}. \quad (2)$$

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 ζ_{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 [1].

In order to reconstruct a GGM network from a given data set one typically employs the following procedure. 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{Z} are significantly different from zero. Finally, the inferred correlation structure is visualized by a graph, with edges corresponding to non-zero partial correlation coefficients.

Problems with High-Dimensional Data

Unfortunately, a number of difficulties arise when the standard graphical Gaussian modeling concept is applied for the analysis of high-dimensional data such as from

a microarray experiment. First, classical GGM theory [1] may only be applied when $n > p$, because otherwise the sample covariance and correlation matrices are not well conditioned, which in turn prevents the computation of partial correlations. Moreover, often there are 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.

Note that the small n large p problem affects both GGMs and relevance networks. In particular, the standard correlation estimates are not valid for small sample size n , a fact that appears to have gone largely unnoticed in the bioinformatics community.

STRATEGIES FOR INFERRING LARGE-SCALE GRAPHICAL GAUSSIAN MODELS FROM GENOMIC DATA

As outlined above, application of GGMs to high-dimensional genomic data is quite challenging, as the number of genes p is usually much larger than the number of available samples n , and standard graphical modeling is not valid in a small-sample setting.

Thus, sound methodologies to circumvent these dimensionality problems need to be devised. Generally speaking, three different paths have been described in the literature that we summarize in the following.

Dimension reduction

The most obvious and simplest approach to avoid problems with high dimensionality is to assess network-like relationships among either only a rather small subset of genes [6, 7, 8, 9] or among a small number of clusters of genes [10, 11, 12]. The number p of selected genes or clusters has to be chosen such that it does not exceed the sample size n .

However, this strategy is unsatisfying for a variety of reasons. Primarily, it is a matter of on-going debate to choose reasonable (meta)-genes for inclusion in the reduced data set. The restriction to a limited number of genes risks that the estimated network topology is seriously distorted because important genes may have been excluded from the analysis. Furthermore, the partial correlation coefficients for gene clusters are hard to understand. For instance, typically, not all the genes of one cluster will interact with all the genes of another cluster, which renders conditional dependence properties among clusters meaningless. In addition, information regarding quality and strength of the association on the gene level is lost when only clusters of genes are considered.

Limited Order Partial Correlations

Another possibility to tackle the small n large p problem is to compute partial correlation coefficients of limited order. For instance, de la Fuente et al. [13] propose to calculate partial correlation coefficients up to second-order only, i.e. to condition the partial correlations not on all other $p - 2$ genes as in a full GGM but only on two genes at most. Similar strategies, based on first-order conditional independence, are also employed by Wille et al. [14] and Magwene and Kim [15].

From a statistical point of view the resulting gene network constitutes something inbetween a full GGM and a relevance network model based on standard correlations. It therefore remains unclear whether missing edges indicate conditional or marginal independence. However, as interactions are likely to be short range, we believe that the above methods may nevertheless provide a good approximation.

Regularized GGMs

In our opinion the statistically and also biologically most sound way to marry GGMs with small-sample modeling is to introduce regularization and moderation. In the first instance, this boils down to finding suitable estimates for the covariance matrix and its inverse when $n < p$. In a statistical context regularization is best done either in a full Bayesian manner or in an empirical Bayes way (a further possibility constitutes the explicit frequentist penalization of the number of free parameters in Z). Once regularized estimates of partial correlation are available heuristic or stochastic model searches need to be employed in order to find an optimal network (or set of networks).

Outside a genomic context using regularized GGMs was first proposed by Wong et al. [16]. For gene expression data this strategy is pursued in the following papers:

1. Dobra et al. [17] describe a variant of Bayesian covariance selection and the associated HdBCS search algorithm. Further related preprints on this method are available on the web page of Prof. M. West (<http://www.isds.duke.edu/~mw/>).
2. In the approach of Schäfer and Strimmer [18] the (inverse) correlation matrix is regularized by bootstrap variance reduction. Subsequently, the final GGM is selected via an heuristic based on empirical Bayes multiple testing.

Full Bayesian MCMC methods such as [17] are computationally very expensive. This in great contrast to an empirical Bayes approach which we describe in the next section in more detail.

EMPIRICAL BAYES APPROACH TO ESTIMATING GGM NETWORKS

This method was first described in Schäfer and Strimmer [18] and consists of two steps. First, an improved small-sample estimator is obtained using variance reduction. Second,

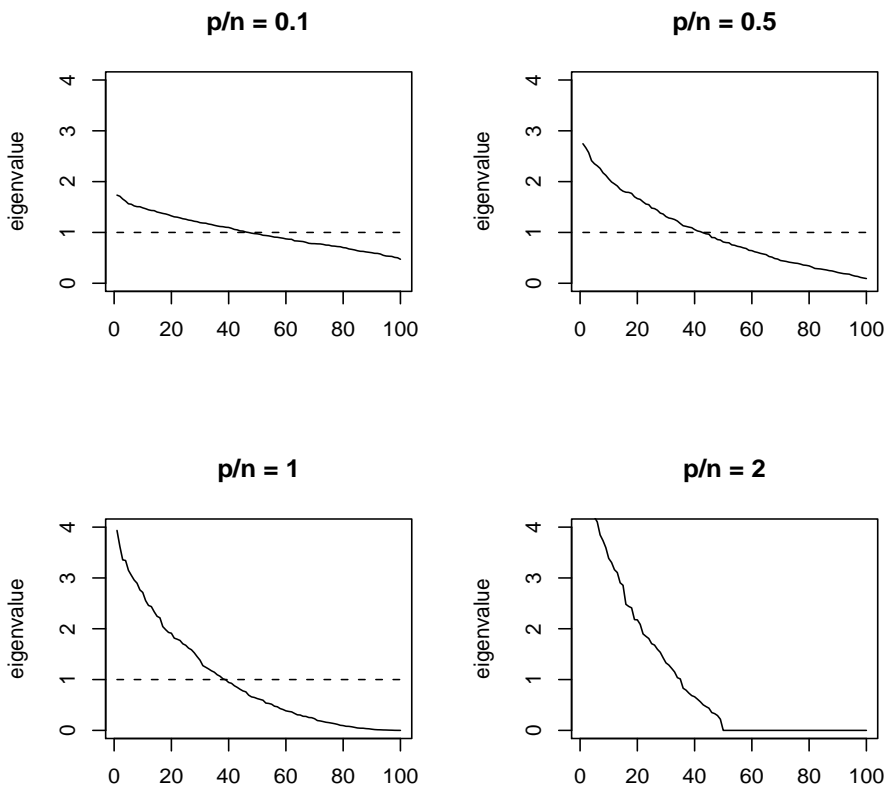


FIGURE 1. Ordered eigenvalues of the sample covariance matrix, calculated from simulated data with underlying p -dimensional standard normal distribution, for $p = 100$ and various ratios p/n .

small-sample network selection is done via an empirical Bayes approach applied to large-scale multiple testing.

Improved Partial Correlation Estimator

The standard sample covariance and correlation estimators have the defect that they are not well conditioned for $n < p$. In this case, the eigenvalues are distorted, in particular many of them are zero (cf. Fig. 1).

This has several negative consequences. First, both the sample covariance and correlation matrices cannot easily be inverted to obtain partial correlation coefficients. Second,

and less obvious, the unbiased sample covariance S is only a very poor approximation of the true covariance Σ with large mean squared error (MSE). It is well known from Stein [19] that one can construct better estimators of Σ with smaller MSE if one drops the requirement of unbiasedness. The MSE equals variance plus squared bias, hence a variance-reduced biased covariance estimator may outperform S in small samples.

Consequently, Schäfer and Strimmer [18] propose as a simple non-parametric approach to employ bootstrap aggregation [20] as a mechanism of variance reduction in the estimation of the covariance matrix and its inverse. This algorithm proceeds as follows for an estimator $\hat{\theta}(y)$ on data y :

1. Generate a bootstrap sample y^{*b} with replacement from the original data. Repeat this process for each $b = 1, \dots, B$ 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.

Another interpretation of the bagged estimate is as an approximate Bayesian posterior mean estimate [21].

In order to invert the covariance matrix Schäfer and Strimmer [18] further suggest to use the Moore-Penrose pseudo-inverse [22], 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 $P = UDV^T$ where D is a square diagonal matrix of rank $m \leq \min(n, p)$ containing all non-vanishing singular values. The pseudo-inverse P^+ is then defined as $P^+ = VD^{-1}U^T$ 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.

Both these techniques combined allow to construct a small-sample estimator of the partial correlation matrix $Z = (\zeta_{ij})$. In particular, Schäfer and Strimmer [18] consider the following three possibilities:

- \hat{Z}^1 : Use the pseudo-inverse for inverting the sample correlation matrix \hat{P} in order to obtain an estimate of Z , without performing any form of bagging (= "observed partial correlation").
- \hat{Z}^2 : Use bagging to estimate the correlation matrix P , then invert the bagged correlation matrix with the pseudo-inverse to obtain an estimate of Z (= "partial bagged correlation").
- \hat{Z}^3 : Apply bagging to the estimator \hat{Z}^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 - see below in the section on computer simulations.

Small-Sample Model Selection

Model selection for a network graph is equivalent to determining which edges are assumed to be present. For GGMs this amounts to deciding which partial correlations differ significantly from zero.

Use of the massively parallel structure of testing $p(p-1)/2$ edges simultaneously allows for the empirical Bayes estimating and inference approach. In particular, one may assume that the estimated partial correlation coefficients z across all edges in the network follow a mixture density

$$f(z) = \eta_0 f_0(z; \kappa) + \eta_A f_A(z), \quad (3)$$

where η_0 and η_A are the priors for the null and alternative density, f_0 and f_A , respectively, with $\eta_0 + \eta_A = 1$ and $\eta_0 \gg \eta_A$. In other words, only a very small fraction η_A of pairwise partial correlations will correspond to non-zero edges, whereas for the remaining majority the corresponding partial correlations will vanish.

The density of an estimated partial correlation coefficient z under the null hypothesis of vanishing true partial correlation ζ , f_0 , is given by Hotelling [23] as

$$f_0(z; \kappa) = (1 - z^2)^{(\kappa-3)/2} \frac{\Gamma(\frac{\kappa}{2})}{\pi^{\frac{1}{2}} \Gamma(\frac{\kappa-1}{2})}, \quad (4)$$

where κ is the degree of freedom. In high-dimensional experimental settings it cannot be computed as a simple function of sample size n and the number of features p but rather has to be estimated itself. Note that positive values of κ are only guaranteed for $n > p$. For $\zeta = 0$ the variance of z also equals the inverse of κ , i.e. $\text{Var}(z) = \frac{1}{\kappa}$.

Fitting this mixture density to the estimated partial correlation coefficients (via optimizing the corresponding likelihood function or an EM-type algorithm) allows for estimating the parameters $\hat{\eta}_0$ and $\hat{\kappa}$. In doing so one carries out the type of empirical Bayes analysis proposed by Robbins [24] and Efron [25].

It is then straightforward to compute two-sided p -values for each possible edge in the network using the null distribution f_0 with $\hat{\kappa}$ as plug-in estimate. Alternatively, one may also compute

$$\text{Prob}(\text{non-zero edge}|z) = 1 - \frac{\hat{\eta}_0 f_0(z; \hat{\kappa})}{f(z)}, \quad (5)$$

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 [25, 26, 27]. There, the mixture density represents the assumption of two classes of genes, differentially and not differentially expressed genes, with the majority of investigated genes coming from the latter class.

There is a close connection with false discovery rate (FDR) multiple testing. Instead of controlling the empirical posterior probability we may also compute an ordered list of p -values for all possible edges, and then apply the algorithm by Benjamini and Hochberg [28] to control the expected proportion of false positives out of the total number of rejections. We refer to [28] and [18] for details.

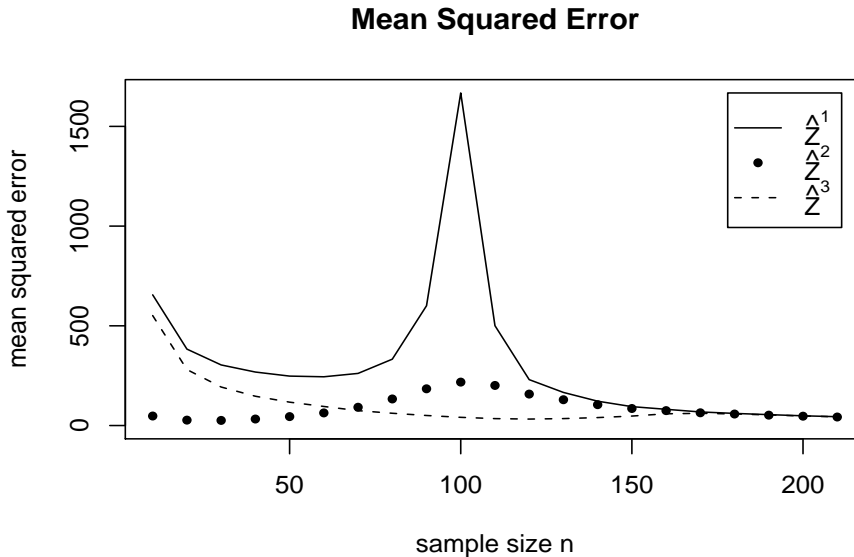


FIGURE 2. Mean squared error of the three small-sample estimators \hat{Z}^1 , \hat{Z}^2 , and \hat{Z}^3 in dependence of sample size n for $p = 100$ genes.

EXAMPLES

Computer Simulations

The computational efficiency of the empirical Bayes network estimator enables the study of accuracy and power of the network reconstruction in computer simulations.

For this purpose we generated random networks and corresponding data (see [18] for the detailed algorithm). This allows to control parameters of interest such as the number of variables p , the fraction of non-zero edges η_A , and the sample size n of simulated data. From the simulated data point estimates \hat{Z}^1 , \hat{Z}^2 , and \hat{Z}^3 are calculated. We compare these to the known underlying partial correlation matrix Z . As a measure of the accuracy of the estimates we employ the squared error loss $L(\hat{Z}^k, Z) = \|\hat{Z}^k - Z\|_F^2 = \sum_{ij} (\hat{\zeta}_{ij}^k - \zeta_{ij})^2$. The expected loss (risk), or mean squared error (MSE), is estimated by averaging $L(\hat{Z}^k, Z)$ over multiple simulation runs.

The results are shown in Fig. 2 for simulations conducted with $p = 100$, $\eta_A = 0.02$, and $n = 10, 20, \dots, 210$. The most striking finding is the existence of three different regions, where all three estimators exhibit clearly different properties. For large samples with $n \gg p$ the estimators \hat{Z}^1 , \hat{Z}^2 , and \hat{Z}^3 mainly agree with each other, with the same low error. Note that this is the only region where classical graphical Gaussian modeling

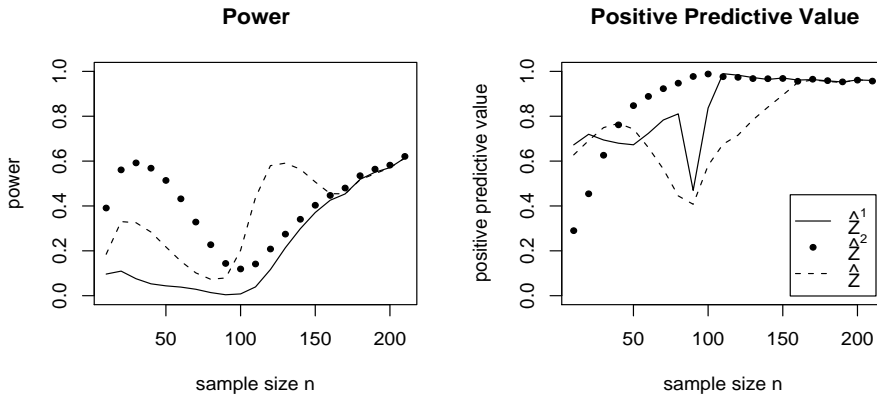


FIGURE 3. Power and positive predictive value for recovering the true GGM network. See the main text for definitions of the investigated quantities and the simulation setup with $p = 100$ genes.

theory is valid. On the other end, for very small samples of size $n \ll p$ the best estimate in terms of mean squared error is \hat{Z}^2 . In the “critical n ” zone with $n \approx p$ an eye-catching dimensionality resonance effect is observed. The MSE of \hat{Z}^1 increases dramatically around this region, 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 [29, 30].

Complementary to the study of accuracy for the diverse partial correlation estimates we also investigated the power of the suggested network selection procedure via empirical Bayes multiple testing.

Again, simulations with sample size n ranging from 10 to 210 in steps of 10, $p = 100$ and $\eta_A = 0.02$ were conducted. 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 $p(p-1)/2 = 4950$ edges with the desired FDR level fixed at $Q = 0.05$.

For each inferred network we counted the number of true positive features as well as the number of true negatives. From these raw statistics, and repeated simulations of networks and data, we obtained estimates of false positive rate, power, and positive predictive value for \hat{Z}^1 , \hat{Z}^2 , and \hat{Z}^3 at a given sample size n . The positive predictive value (PPV) is defined as the expected proportion of true positives among all significant findings. Figure 3 visualizes our results.

All three small-sample estimators, \hat{Z}^1 , \hat{Z}^2 , and \hat{Z}^3 , exhibit the same low empirical false positive rate regardless of n (data not shown). For large $n > 170$ they also all agree in power and in positive predictive value. However, they differ drastically in the small-sample case $n < p$ and for $n \approx p$. In terms of power the bagged estimators both \hat{Z}^2 and \hat{Z}^3 consistently outperform the simple estimator \hat{Z}^1 that fares rather poorly particularly

for $n < p$. In the latter region \hat{Z}^2 exhibits the overall highest power, whereas for $n \approx p$ and sample sizes slightly above p the estimator \hat{Z}^3 performs best.

The largest PPV is generally obtained by using the estimator \hat{Z}^2 . However, for very small samples the PPV of \hat{Z}^2 drops sharply. This is likely due to its imperfect goodness of fit with the theoretical null distribution.

A further noteworthy result from all our simulations is that in the region of $p \approx n$ 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.

For further conclusions we refer to [18].

Choice of Small-Sample Estimator

The estimators \hat{Z}^1 , \hat{Z}^2 , and \hat{Z}^3 perform very differently, as emanates from Figs. 2 and 3. For choosing a suitable estimator we suggest the following guidelines:

\hat{Z}^1 : Should only be used for $n \gg p$, 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{Z}^2 : This estimator is best used for small-sample applications with $n < p$. Here the main advantage of \hat{Z}^2 is its high accuracy as a point estimate. It exhibits the overall best power for sample sizes n that are small compared to the number of features p . Furthermore, it is computationally less expensive than \hat{Z}^3 . However, note its low PPV for very small n .

\hat{Z}^3 : In the “critical n ” zone with $n \approx p$ \hat{Z}^3 offers small error and thus large effective sample size. For n slightly larger than p this estimator also provides the overall best power, though in terms of PPV this estimator performs less well than \hat{Z}^2 .

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

Microarray Data

For illustration purposes we also applied the suggested empirical Bayes framework of inferring GGM networks from small samples to a large-scale biological data set.

We reanalyzed data from West et al. [31]. After preprocessing and calibration gene expression data for 3,883 genes across 49 samples remained for further analysis. In order to infer the global association structure and the corresponding GGM network for all 3,883 genes the small-sample estimator \hat{Z}^2 was employed with $B = 10,000$ bootstrap replications.

The empirical Bayes version of fitting the mixture density (Eq. 3) resulted in an estimated degree of freedom $\hat{\kappa} = 4601.98$ with $\hat{\eta}_0 = 0.9924$. Using FDR multiple testing with a desired level $Q = 0.05$ we determined 88,822 significantly non-zero coefficients, corresponding to a raw p -value cutoff of 0.0006 and a threshold of partial correlation $\hat{\zeta} > 0.051$. Note that for this network size most of the coefficients are very close to zero,

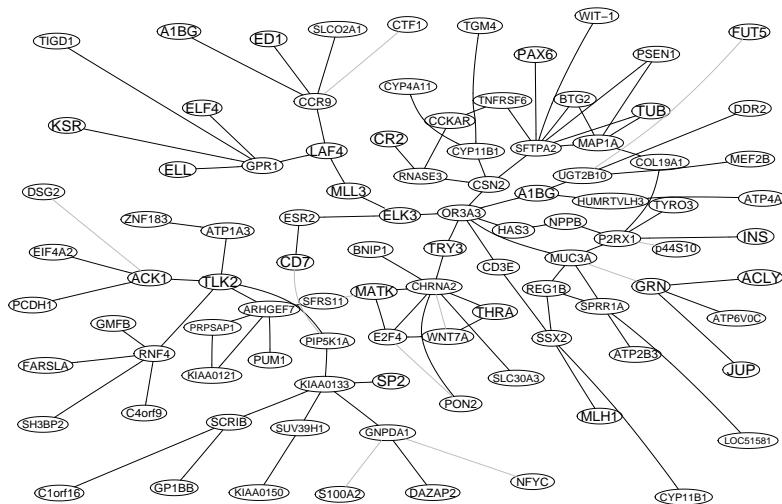


FIGURE 4. Example gene association network inferred from breast cancer data.

so that even small values are statistically significant. This is also reflected in the large value of $\hat{\kappa}$.

As an example a subnetwork consisting of 96 genes centered around the *ESR2* gene is shown in Fig. 4. Note that this network does not represent a picture of mechanistic interactions among the depicted genes. Instead, it shows interactions on a phenomenological level, similarly as in clustering approaches. The depicted network includes many genes related to the development and regulation of cancer (cf. [18]). Hence, we are cautiously optimistic that the inferred gene association network may indeed be useful as a starting point from which to generate further medical and biochemical hypotheses.

DISCUSSION

Among the methods for inferring networked interactions among genes graphical Gaussian models are becoming increasingly popular [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18]. Their advantage over simple correlation networks, namely the ability to distinguish direct from mediated interactions, is apparent. However, their application to genome data is hampered by the small n large p problem which renders both estimation and inference difficult.

In this paper we have reviewed the three main strategies used in the literature to circumvent these dimensionality problems: dimension reduction prior to the analysis, computation of low order partial correlation coefficients, and regularized variants of graphical Gaussian modeling.

We believe that the latter approach is most promising. If one can afford the large computation time, we recommend the full Bayesian approach by Dobra et al. [17]. Alternatively, the much more computationally efficient empirical Bayes approach developed in [18] is advised. Note that the latter method can be evaluated in extensive simulations with respect to its performance in dependence of the sample size. We recommend this type of power analysis to be done also for other network inference approaches (where studies of this kind appear to be notably absent as pointed out before by Husmeier [3]).

The main conclusion of the work review here is that small-sample inference of graphical models, specifically GGMs, is possible even for small samples. However, the assumption of linear relationships as measured by partial correlations is limiting. Non-linear interactions as well as combinatorial effects will most likely better characterize biomolecular networks. Owing to the sparsity of genomic data it is yet prudent to choose a simple model that requires few parameters and to act on the assumption of approximate validity. This is corroborated by several examples of successful application of graphical modeling to gene expression data [e.g. 14, 15, 17].

ACKNOWLEDGMENTS

This research was supported by an Emmy Noether research grant (STR 624/1-2,3) from the Deutsche Forschungsgemeinschaft (DFG).

APPENDIX: SOFTWARE

The empirical Bayes approach to infer GGMs is implemented 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>).

REFERENCES

1. J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley, New York, 1990.
2. S. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, 1996.
3. D. Husmeier, *Bioinformatics*, **19**, 2271–2282 (2003).
4. A. P. Dempster, *Biometrics*, **28**, 157–175 (1972).
5. D. Edwards, *Introduction to Graphical Modelling*, Springer, New York, 1995.
6. H. Kishino, and P. J. Waddell, *Genome Informatics*, **11**, 83–95 (2000).
7. P. J. Waddell, and H. Kishino, *Genome Informatics*, **11**, 129–140 (2000).
8. S. D. Bay, J. Shrager, A. Pohorille, and P. Langley, *J. Biomed. Informatics*, **35**, 298–297 (2002).
9. J. Wang, O. Myklebost, and E. Hovig, *Bioinformatics*, **19**, 2210–2211 (2003).

10. H. Toh, and K. Horimoto, *Bioinformatics*, **18**, 287–297 (2002).
11. H. Toh, and K. Horimoto, *J. Biol. Physics*, **28**, 449–464 (2002).
12. X. Wu, Y. Ye, and K. R. Subramanian, *ACM SIGKDD Workshop on Data Mining in Bioinformatics*, **3**, 63–69 (2003).
13. A. de la Fuente, N. Bing, I. Hoeschele, and P. Mendes, *Bioinformatics*, **20**, 3565–3574 (2004).
14. A. Wille, P. Zimmermann, E. Vranová, A. Fürholz, O. Laule, S. Bleuler, L. Hennig, A. Prelić, P. von Rohr, L. Thiele, E. Zitzler, W. Gruissem, and P. Bühlmann, *Genome Biology*, **5**, R92 (2004).
15. P. M. Magwene, and J. Kim, *Genome Biology*, **5**, R100 (2004).
16. F. Wong, C. K. Carter, and R. Kohn, *Biometrika*, **90**, 809–830 (2003).
17. A. Dobra, C. Hans, B. Jones, J. R. Nevins, and M. West, *J. Multiv. Anal.*, **90**, 196–212 (2004).
18. J. Schäfer, and K. Strimmer, *Bioinformatics*, **21**, 754–764 (2005).
19. C. Stein, “Inadmissibility of the usual estimator for the mean of a multivariate distribution,” in [32], pp. 197–206.
20. L. Breiman, *Machine Learning*, **24**, 123–140 (1996).
21. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2001.
22. R. Penrose, *Proc. Cambridge Phil. Soc.*, **51**, 406–413 (1955).
23. H. Hotelling, *J. R. Statist. Soc. B*, **15**, 193–232 (1953).
24. H. Robbins, “An empirical Bayes approach to statistics,” in [32], pp. 157–163.
25. B. Efron, *Annals of Statistics*, **31**, 366–378 (2003).
26. M. Sapir, and G. A. Churchill, Estimating the posterior probability of differential gene expression from microarray data, Poster, Jackson Laboratory, Bar Harbor (2000).
27. B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher, *J. Amer. Statist. Assoc.*, **96**, 1151–1160 (2001).
28. Y. Benjamini, and Y. Hochberg, *J. R. Statist. Soc. B*, **57**, 289–300 (1995).
29. S. Raudys, and R. P. W. Duin, *Patt. Recogn. Lett.*, **19**, 385–392 (1998).
30. M. Skurichina, and R. P. W. Duin, *Patt. Analysis and Appl.*, **5**, 121–135 (2002).
31. M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, *Proc. Natl. Acad. Sci. USA*, **98**, 11462–11467 (2001).
32. J. Neyman, editor, *Proc. Third Berkeley Symp. Math. Statist. Probab.*, vol. 1, Univ. California Press, Berkeley, 1956.