

Efficient estimation of covariances and dependencies in high-dimensional gene expression data

Rainer Opgen-Rhein¹

University of Munich, e-mail: opgen-rhein@lrz.uni-muenchen.de

Abstract: The learning of dependencies in microarray data is challenging. Here, we will give a review of estimation methods based on Stein-type shrinkage. At their core lies a regularized estimation of the covariance matrix of the data. Subsequently, genetic networks from both, static and time-series data, can be inferred. The algorithms described exhibit a high accuracy, are computationally efficient, and are applicable for large-scale data sets.

Keywords: Genetic networks, Shrinkage estimation, Gaussian Graphical models, Directed networks, VAR process

1. Introduction

The identification of networked genetic interdependencies that form the basis of cellular regulation is one of the key issues in systems biology. With the appearance of new high throughput technologies, e.g., microarray technology, much progress has been made in the measurement of the components of the genetic networks, so that hundreds or even thousands of variables can be observed at the same time. These developments are challenging for statistical inference, as in this “small n , large p ” paradigm standard estimation procedures tend to be unreliable or to be even inapplicable (West *et al.*, 2000). In this article we will give a review of regularization techniques based on James-Stein estimation (James and Stein, 1961) to learn genetic networks from microarray data.

2. Stein-type shrinkage estimation of the covariance matrix

To estimate the covariances between gene expression profiles we use Stein-type shrinkage estimation. Its basic idea is that all separate estimators of the different parameters are shrunk towards a common target (see Efron and Morris, 1975).

Here, we apply Stein-type inference for the estimation of the covariance structure of the variables. The reason for the central role of the covariance matrix is the possibility to infer all linear relationships among the variables *directly* from the covariance matrix, without further estimation.

As the covariance matrix can be separated into two natural groups – the variances of the variables and the correlation between the variables – the shrinkage intensities are estimated separately. The estimator of the covariance matrix developed in Schäfer and Strimmer (2005b) and in Opgen-Rhein and Strimmer (2006a) forms the basis of the methods presented here.

¹Address of correspondence: Ludwigstraße 28 Rgb., Munich, Germany

3. Network construction

Using networks to visualize the connection between parameters is a common and popular method in many areas of research. The nodes represent the variables, e.g., genes in systems biology, and the links or edges a connection between them. An arrow from a node A to a node B constitutes directed networks and imply that node A influences node B .

Graphical models A graphical model is a representation of stochastic conditional dependencies between the investigated variables. That means that only direct dependencies between variables are taken into account. Especially in systems biology, many authors have followed this statistical approach to estimate genetic networks from high-throughput data (for a discussion, see Schäfer and Strimmer, 2005a).

Among the simplest graphical models is the class of graphical Gaussian models (GGMs) – see, e.g., Whittaker (1990). These are based on the partial correlation matrix: the strength of these coefficients indicates the presence or absence of a direct association between each pair of genes. The partial correlation $\tilde{\rho}_{kl}$ and also the partial variance $\tilde{\sigma}_{kk}$ are defined via the inverse of the covariance matrix and are based on the same idea of conditioning on other variables.

Model selection using false discovery rate (fdr) For the identification of large scale networks we need a test for the significance of the partial correlations applicable to large scale testing problems. We find such a test in the local fdr algorithm (e.g., Efron, 2008), which is based on the idea of separating the distribution of the estimations of the partial correlations resulting from variables without any connections (the null distribution) from the distribution of the estimations from variables that have a non-zero partial correlation. The null distribution, the density under null hypothesis, depends on the test problem. For identifying significant edges of the GGM, we have to focus on the distribution of the partial correlation coefficients with the null hypothesis $\tilde{\rho} = 0$.

An algorithm to infer a graphical Gaussian model We now summarize the procedure to infer a graphical Gaussian network using shrinkage estimation of the correlation matrix and local fdr model selection.

1. The correlation matrix of the data is estimated by the Stein-type shrinkage estimator.
2. The estimated correlations allow calculating the *partial* correlation matrix.
3. The local fdr algorithm provides *significant* partial correlations.
4. Significant partial correlations are included as edges in the network, which connect the variables (the nodes).

Inferring directed networks The partial correlation graph allows another interpretation: the partial correlation is the geometric mean of β_k^y and the corresponding reciprocal coefficient β_y^k , i.e. $\sqrt{\beta_y^k \beta_k^y} = |\tilde{\rho}_{yk}|$. In this light, the GGM represents a system of linear regression equations, where each node is in turn taken as a response variable and regressed against the other remaining nodes. Hence, an undirected edge between node A and B in a GGM is in fact rather a bi-directed edge, in the sense, that A influences B and vice versa in the underlying system of regression. It is possible to induce directionality by testing

the ratio of the standardized partial variances $\mathcal{B} = 1$. We suggest a simple yet versatile heuristic for screening large-scale data set for causal structure.

Our procedure is based on projecting an estimated (partial) ordering of nodes onto a partial correlation graph (Opge-Rhein and Strimmer, 2007a).

4. Time series data

Dynamic correlation To account for time series data of complex systems, we now investigate GGM network inference from the perspective of functional data analysis. This approach is based on the notion of dynamical correlation and covariance (Opge-Rhein and Strimmer, 2006b,c). It provides a similarity score for pairs of groups of randomly sampled curves. Shrinkage allows to improve the precision of the estimation and to extend the method to high dimensional data. Once the dynamical correlation matrix is obtained, partial dynamical correlations can be computed and the identification of the associated network structure be inferred according to the inference of GGMs from the correlation structure as described above.

While this method covers the dynamical correlation through time it is not able to account, e.g., for a time shift between any two variables. These dependencies and the associated time shifts can in the discrete case be accounted for by using an autoregressive process. This will be described next.

Genetic networks based on a vector autoregressive (VAR) process For a vector autoregressive (VAR) model to estimate the interdependencies of microarray data, we need a regularized version of linear regression. Given response variables Y which are regressed on predictor variables X , we can define a matrix of the combined response and predictor variables Φ . The covariance matrix allows to derive shrinkage regression. Note that the empirical estimator for the covariance S of the combined response and predictor variables Φ contains two submatrices $S_1 = X^T X$ and $S_2 = X^T Y$. The OLS estimate of the regression coefficients can then be written as $\hat{B}^{\text{OLS}} = (S_1)^{-1} S_2$. Replacing S by the shrinkage estimator S^* leads to the shrinkage estimation of the regression coefficients $\hat{B}^{\text{Shrink}} = (S_1^*)^{-1} S_2^*$ and hence to the shrinkage regression.

To introduce the vector autoregressive model, we consider vector-valued time series data $\mathbf{x}(t) = (x_1(t), \dots, x_p(t))$. Each component of this row vector corresponds to a variable of interest, e.g., the expression level of a specific gene. We denote the centered *matrices of observations* corresponding to $\mathbf{x}(t+1)$ and $\mathbf{x}(t)$ by X_f (“future”) and X_p (“past”). The autoregressive model has the form of a standard regression problem. Therefore, estimation of the matrices B_i is straightforward.

Small sample shrinkage estimates of VAR regression coefficients may be obtained by applying the shrinkage regression described above appropriately to the VAR process (Opge-Rhein and Strimmer, 2007b):

1. Combine the centered observations X_p and X_f into $\Phi = [X_p X_f]$.
2. The $(n-1)$ multiple of the *empirical* covariance matrix, $S = \Phi^T \Phi$, contains $S_1 = X_p^T X_p$ and $S_2 = X_p^T X_f$.
3. Calculate the shrinkage estimate S^* for the covariance matrix of Φ .
4. Determine the submatrices S_1^* and S_2^* and compute $\hat{B}^{\text{Shrink}} = (S_1^*)^{-1} S_2^*$.

The network representing potentially directed causal influences is given by the non-zero entries in the matrix of VAR coefficients.

The network can be inferred in the same way as the GGM described above, nevertheless, unlike in a graphical Gaussian model, the edges in a VAR network are by design directed.

5. Summary

Learning of genetic interdependencies from microarray data is a challenge for statistical analysis. Here, we presented methods that use a Stein-type shrinkage approach to extend the area of application of traditional methods to high-dimensional data sets. The combination of frequentist and Bayesian ideas Stein-type shrinkage builds upon seems very promising in solving the problems that learning in complex systems implicates (Efron, 2008).

References

- Efron B. (2008) Microarrays, empirical bayes and the two-groups model, *Statistical Science*, 23, 1, 1–22.
- Efron B. and Morris C.N. (1975) Data analysis using Stein’s estimator and its generalizations, *J. Amer. Statist. Assoc.*, 70, 311–319.
- James W. and Stein C. (1961) Estimation with quadratic loss, in: *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, Neyman J., ed., Univ. California Press, Berkeley, volume 1, 361–379.
- Opge-Rhein R. and Strimmer K. (2006a) Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach, *Stat. Appl. Genet. Mol. Biol.*, 6, 1, 9.
- Opge-Rhein R. and Strimmer K. (2006b) Inferring gene dependency networks from genomic longitudinal data: a functional data approach, *REVSTAT*, 4, 53–65.
- Opge-Rhein R. and Strimmer K. (2006c) Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data, in: *Proc. of WCSB 2006*, Tampere, volume 4, 73–76.
- Opge-Rhein R. and Strimmer K. (2007a) From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data, *BMC Systems Biology*, 1, 37.
- Opge-Rhein R. and Strimmer K. (2007b) Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process, *BMC Bioinformatics*, 8 (Suppl. 2), S3.
- Schäfer J. and Strimmer K. (2005a) An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics*, 21, 6, 754–764.
- Schäfer J. and Strimmer K. (2005b) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statistical Applications in Genetics and Molecular Biology*, 4, 1, 32.
- West M., Nevins J.R., Marks J.R., Spang R. and Zuzan H. (2000) Bayesian regression analysis in the “large p , small n ” paradigm with application in DNA microarray studies, *Technical Report*.
- Whittaker J. (1990) *Graphical Models in Applied Multivariate Statistics*, Wiley, New York.