USING REGULARIZED DYNAMIC CORRELATION TO INFER GENE DEPENDENCY NETWORKS FROM TIME-SERIES MICROARRAY DATA

Rainer Opgen-Rhein and Korbinian Strimmer

Department of Statistics, University of Munich,
Ludwigstrasse 33, D-80539 Munich, Germany
opgen-rhein@stat.uni-muenchen.de, korbinian.strimmer@lmu.de

ABSTRACT

Graphical models allow to understand regulatory interactions among genes and gene products in a cell, and hence contribute to an enhanced understanding of systems biology. Here we investigate a graphical model that treats the observed gene expression over time as realizations of random curves. This approach is centered around a regularized estimator of dynamical pairwise correlation that takes account of the functional nature of the observed data. The new method is illustrated by analyzing highly replicated gene expression time series data.

1. INTRODUCTION

The identification of networked genetic interdependencies that form the basis of cellular regulation is one of the key issues in systems biology. Consequently, many authors have investigated statistical approaches such as graphical models to estimate genetic networks from high-throughput data [e.g., 1]. Among the simplest graphical models is the class of graphical Gaussian models (GGMs) – see, e.g., Whittaker, 1990 [2].

A drawback shared by the GGM approach and other graphical models such as Bayesian networks is that these methods rely on the assumption of i.i.d. data. However, an increasing proportion of microarray expression experiments are concerned with *longitudinal* measurements of mRNA and protein concentrations.

In order to account for this we investigate GGM network inference from the perspective of functional data analysis (FDA) [3, 4]. Specifically, we describe a graphical model that treats the observed gene expression over time as realizations of random curves, rather than to describe the individual time points separately.

The remainder of the paper is organized as follows. In the next section following [4] we summarize the basic notation for functional data analysis and also introduce the functional inner product. Next, we discuss the concept of dynamical correlation and introduce a regularization technique for the “small n, large p” domain [5]. Subsequently, the dynamical correlation is employed for GGM network selection. Finally, we analyze data from a human T-cell experiment [6].

2. METHODS

2.1. Setup and Notation

We consider data from a typical gene expression time course experiment. For p genes (variables) and n subjects (replications) mRNA concentrations are measured over a time interval $[A, B]$. This results in functional observations $f_{ik}(t)$ where $1 \leq i \leq n$ and $1 \leq k, l \leq p$. We assume all functions $f_{ik}(t)$ to be square-integrable so that the functional inner product

$$\langle g(t), h(t) \rangle = \frac{1}{B-A} \int_A^B g(t)h(t)dt \qquad (1)$$

exists, where $g(t)$ and $h(t)$ are any of the observed functions. The time average of $f_{ik}(t)$ may then be conveniently expressed by $\langle f_{ik}(t), 1 \rangle$. The average over the n replicates gives the empirical mean function $\bar{f}_k(t) = \frac{1}{n}\sum_{i=1}^n f_{ik}(t)$.

In practice, however, the functions $f_{ik}(t)$ are not continuously measured but rather obtained by experiments at discrete time points t_j, with $1 \leq j \leq m$ and $A = t_1 < t_2 < \ldots < t_{m-1} < t_m = B$. Note that the time points need not be equidistant. If one assumes a linear approximation of $g(t)$ and $h(t)$ the inner product of Eq. 1 turns into the weighted sum

$$\langle g(t), h(t) \rangle \approx \sum_{j=1}^m g(t_j)h(t_j)\,\frac{\delta_j + \delta_{j+1}}{2(B-A)} \qquad (2)$$

where the $\delta_j = t_j - t_{j-1}$ are the time differences between subsequent measurements (with $\delta_1 = \delta_{m+1} = 0$).

In the random effects representation of Dubin and Müller, 2005 [7] each observed $f_{ik}(t)$ is a realization of the random function

$$f_k(t) = \mu_k(t) + \mu_{0k} + \epsilon_{0k} + \sum_{u=1}^{\infty} \epsilon_{uk}\eta_u(t), \qquad (3)$$

where ϵ_{0k} and ϵ_{uk} are random variables with $E(\epsilon_{0k}) = 0$ and $E(\epsilon_{uk}) = 0$, $\mu_k(t)$ is the fixed time dependent mean function with zero time average $\langle \mu_k(t), 1 \rangle = 0$, $\mu_{0k} + \epsilon_{0k}$ represents the static random part and the remaining terms describe the dynamic random part. In Eq. 3 the $\eta_u(t)$

are orthonormal basis functions with zero time average $\langle \eta_u(t), 1\rangle = 0$.

In this notation the empirical mean function $\bar{f}_k(t)$ is an estimate of $E(f_k(t)) = \mu_k(t) + \mu_{0k}$. As $\mu_k(t)$ has time average zero we are also able to identify the two components of $E(f_k(t))$ by using $\hat{\mu}_{0k} = \langle \bar{f}_k(t), 1\rangle$ and $\hat{\mu}_k(t) = \bar{f}_k(t) - \hat{\mu}_{0k}$.

2.2. Dynamical Correlation

2.2.1. Measuring similarity between two exactly known curves

Suppose for a moment that we have sufficient data to estimate the expression levels through time of two genes k and l *exactly*, i.e. that we know the mean functions $E(f_k(t))$ and $E(f_l(t))$. In order to understand the functional connection between these two variables a measure of similarity between the two curves is required. Dubin and Müller, 2005 [7] suggest to introduce the notion of *dynamical correlation* with the informal proposition that "if both trajectories tend to be mostly on the same side of their time average (a constant) then the dynamical correlation is positive; if the opposite occurs, then dynamical correlation is negative".

This immediately leads to the following straightforward definition of dynamical correlation between two curves $g(t)$ and $h(t)$. First, calculate the time-centered functions $g^C(t) = g(t) - \langle g(t), 1\rangle$ and $h^C(t) = h(t) - \langle h(t), 1\rangle$. Then define the variances as

$$\text{Var}(g(t)) = \langle g^C(t), g^C(t)\rangle$$

and

$$\text{Var}(h(t)) = \langle h^C(t), h^C(t)\rangle.$$

Finally, compute the standardized functions $g^S(t) = g^C(t)/\sqrt{\text{Var}(g(t))}$ and $h^S(t) = h^C(t)/\sqrt{\text{Var}(h(t))}$, and obtain the correlation by

$$\text{Cor}(g(t), h(t)) = \langle g^S(t), h^S(t)\rangle.$$

2.2.2. The general case including sampling error

The above definition of dynamical correlation for a single curve extends in a straightforward fashion to the case where each observed time course f_{ik} represents a noisy realization of the mean function $E(f_k)$.

In order to estimate the correlation between two variables k and l we first define the simultaneously time- *and* space-centered functions according to $f^C_{ik}(t) = f_{ik}(t) - \langle \bar{f}_k(t), 1\rangle$. Note that here the inner product is computed over the mean function $\bar{f}_k(t)$. Based on the $f^C_{ik}(t)$ the empirical estimate of the variance of variable k is then given by

$$\widehat{\text{Var}}_k = \hat{\sigma}_{kk} = s_{kk} = \frac{1}{n}\sum_{i=1}^{n}\langle f^C_{ik}(t), f^C_{ik}(t)\rangle. \qquad (4)$$

This allows to compute standardized residual functions $f^S_{ik}(t) = f^C_{ik}/\sqrt{s_{kk}}$ that form the basis for the estimate of dynamical correlation

$$\widehat{\text{Cor}}_{kl} = \hat{\rho}_{kl} = r_{kl} = \frac{1}{n}\sum_{i=1}^{n}\langle f^S_{ik}(t), f^S_{il}(t)\rangle. \qquad (5)$$

Correspondingly, the estimated dynamical covariance between variables k and l is simply

$$\widehat{\text{Cov}}_{kl} = \hat{\sigma}_{kl} = s_{kl} = r_{kl}\sqrt{s_{kk}s_{ll}}. \qquad (6)$$

This simple estimator of dynamical correlation exhibits several attractive properties. In particular, it is a generalization of the standard correlation for cross-sectional data. Specifically, if $m = 1$ and $n > 1$ then it reduces to the usual maximum-likelihood estimator of correlation. Furthermore, it is also applicable if there is only a single realization of each time series available ($n = 1$, $m > 1$).

2.2.3. Regularization

The above definition allows the inference of correlations between sets of curves. However, if there are only a few observations for a large number of variables ("small n, large p"-problem), the unbiased empirical estimator is suboptimal in the sense that other, biased estimators may be constructed that are more efficient and exhibit higher accuracy in terms of MSE [8]. The pivotal element in successful learning of complex models from sparse data is regularization. It is possible to achieve a better estimation of dynamic correlation by means of *shrinkage*.

In the present case, we can construct a shrinkage estimate S^* of the dynamic covariance matrix by the convex combination $S^* = \lambda S^{\text{Target}} + (1-\lambda)S$ of the unregularized estimator S and a suitable target S^{Target}. The selection of the shrinkage parameter λ will have to take place in a data-driven fashion and has to meet some requirements. For instance, given a large sample size the shrinkage intensity λ must vanish. A simple rule to estimate the optimal shrinkage intensity can be found by minimizing the MSE risk function

$$R(\lambda) = E\left(\sum_{k=1}^{p}\sum_{l=1}^{p}(s^*_{kl} - s_{kl})^2\right). \qquad (7)$$

It can be shown [9] that the minimum mean squared error $R(\lambda^*)$ is achieved *exactly* and uniquely for the choice

$$\begin{aligned} \lambda^* \;&=\; \frac{1}{\sum_{k=1}^{p}\sum_{l=1}^{p} E\left[(s_{kl} - s^{\text{Target}}_{kl})^2\right]} \\ &\cdot\;\; [\sum_{k=1}^{p}\sum_{l=1}^{p}\text{Var}(s_{kl}) - \text{Cov}(s_{kl}, s^{\text{Target}}_{kl}) + \\ &+\;\; \text{Bias}(s_{kl})E(s_{kl} - s^{\text{Target}}_{kl})]. \end{aligned} \qquad (8)$$

Here we choose S^{Target} to be the diagonal matrix with the variances s_{kk} on the diagonal. Defining

$$\overline{f_{kl}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\underbrace{f^C_{ik}(t_j)f^C_{il}(t_j)}_{f_{ijkl}}\underbrace{\frac{\delta_j + \delta_{j+1}}{2(B-A)n}}_{w_{ij}} \qquad (9)$$

and the sum of squared weights

$$\tau = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}^2 = \frac{1}{n}\sum_{j=1}^{m}\left(\frac{\delta_j + \delta_{j+1}}{2(B-A)}\right)^2 , \qquad (10)$$

the *unbiased* empirical covariance equals

$$\widehat{\text{Cov}}(g(t), h(t)) = s_{kl} = \frac{1}{1-\tau}\overline{f_{kl}} \qquad (11)$$

and find after some calculation the individual entries for

$$\widehat{\text{Var}}(s_{kl}) = \frac{\tau}{(1-\tau)^3}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}\left(f_{ijkl} - \overline{f_{kl}}\right)^2 . \quad (12)$$

For scaling reasons [9] we apply shrinkage to the correlation matrix. The variances $\text{Var}(r_{kl})$ of the empirical correlation coefficients can be estimated by applying the above formulae to the *standardized* data (f_{ik}^S). This leads to the sample approximation of the shrinkage intensity

$$\hat{\lambda}^* = \frac{\sum_{k\neq l}\widehat{\text{Var}}(r_{kl})}{\sum_{k\neq l} r_{kl}^2}, \qquad (13)$$

which in turns allows to calculate the matrix of shrunken dynamical correlation coefficients.

2.3. Estimating Gene Association Networks Using Dynamical Correlation

The basic concept behind inferring a gene dependency network from the pairwise dynamical correlation is to investigate the correlation structure. However, we cannot simply use the correlations directly, because these represent only marginal dependencies and also include indirect interactions between two variables. Instead, we need to rely on the concept of *partial* correlation which describe the correlation between any two variables i and j conditioned on all the other variables. It is straightforward to compute the matrix of partial dynamical correlations $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl})$ from the correlation coefficients $\boldsymbol{P} = (\rho_{kl})$ via the inverse relationship

$$\begin{aligned} \boldsymbol{\Omega} &= \boldsymbol{P}^{-1} = (\omega_{kl}) & (14)\\ \tilde{\rho}_{kl} &= -\frac{\omega_{kl}}{\sqrt{\omega_{kk}\omega_{ll}}} & (15)\end{aligned}$$

[10]. Applying these equations to estimates $\boldsymbol{R} = (r_{kl})$ of (dynamical) correlations allows to obtain estimates $\tilde{\boldsymbol{R}} = (\tilde{r}_{kl})$ of the associated partial (dynamical) correlations.

In order to test the significance of the correlations and to decide which of the possible edges to include in the resulting gene association network statistical tests are needed. In this paper we employ the “local fdr” network search [1, 9]. The false discovery rate (fdr) is the expected proportion of false positives among the proposed edges. The local fdr is an empirical Bayes estimator of the false discovery rate [11]. In the network search the local fdr is utilized to compute the posterior probability for an edge to be present or absent, and takes account of the multiplicity in the simultaneous testing of edges. The final network is obtained by visualizing all significant edges in an undirected graph.

3. RESULTS

We now employ shrinkage estimation of the (partial) dynamical correlation to a real world example and compare it with the results of the traditional GGM method. Specifically, we reanalyzed a microarray time series data set [6]. These data characterize the response of a human T-cell line (Jirkat) to a treatment with PMA and ioconomin, and consist of 10 time points with 44 replications each.

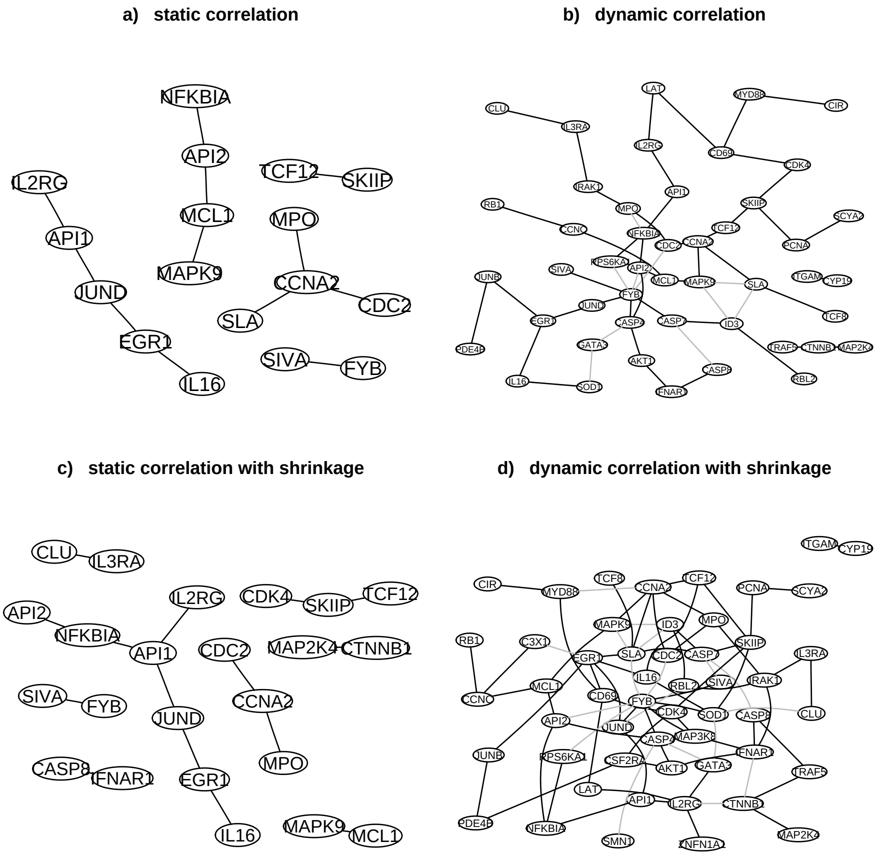
As approximation of the temporal expression of the 58 genes we used a linear spline and employed Eq. 2 for the functional inner product. After estimation of the dynamical correlations with Eq. 5 and regularization (section 2.2.3) we computed the associated partial correlation coefficients employing Eq. 14 and Eq. 15. Using the locfdr algorithm [11] we then identified significant edges. The resulting network is displayed in Fig. 1d.

For comparison we also compute the network as obtained by the classic GGM approach. For this analysis we ignored the dynamic aspects of the data and assumed that all measurements were taken at the same time point. Furthermore, we examine the influence of shrinking. This leads to the four networks displayed in Fig. 1.

Ignoring the time series aspects and using static correlation leads to less well-connected networks compared with the ones calculated by dynamical correlation. This indicates that our dynamical FDA-based estimator is able to extract additional information about the interaction among the investigated genes. Furthermore, shrinkage also improves the power of the network reconstruction. Hence, we conclude that the best of the four investigated methods to infer gene association networks is the one relying on regularized dynamical correlation.

4. CONCLUSION

A growing interest in genetics lies in observing and inferring the gene interactions over time. Here, we introduced a method to infer a *regularized* gene dependency network from functional data. It generalizes the static regularized GGM approach [1] and is able to unravel the dependency structure of longitudinal data across the whole time series. Furthermore, unlike many other time series methods FDA does not require equally spaced measurements. Note that in FDA unequal time points are accounted for by the weights employed in the functional inner product. Furthermore, our algorithm is easily implemented and computationally inexpensive. Shrinkage allows to improve the precision of the estimation and to extend the method to high dimensional data. In order to further develop our approach many extensions are conceivable. An important topic is the inclusion of auto-regressive aspects. While our 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 could be accounted for by modeling the temporal mean via a system of differential equation.



a) static correlation

b) dynamic correlation

c) static correlation with shrinkage

d) dynamic correlation with shrinkage

Figure 1. Gene dependency networks inferred from human T-cell data [6].

5. ACKNOWLEDGMENTS

This work was supported by Deutsche Forschungsgemeinschaft (DFG) Emmy-Noether research award to K.S.

References

[1] J. Schäfer and K. Strimmer, “An empirical Bayes approach to inferring large-scale gene association networks,” *Bioinformatics*, vol. 21, pp. 754–764, 2005.

[2] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*, Wiley, New York, 1990.

[3] J. O. Ramsay and B. W. Silverman, *Functional Data Analysis. 2nd Edition.*, Springer Verlag, New York, 2005.

[4] R. Opgen-Rhein and K. Strimmer, “Inferring gene dependency networks from genomic longitudinal data: a functional data approach,” *RevStat, to appear*, 2006.

[5] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, “Sparse graphical models for exploring gene expression data,” *J. Multiv. Anal.*, vol. 90, pp. 196–212, 2004.

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

[7] J. A. Dubin and H.-G. Müller, “Dynamical correlation for multivariate longitudinal data,” *J. Amer. Statist. Assoc.*, vol. 100, pp. 872–881, 2005.

[8] C. Stein, “Inadmissibility of the usual estimator for the mean of a multivariate normal distribution,” 1956, pp. 197–206.

[9] J. Schäfer and K. Strimmer, “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics,” *Statist. Appl. Genet. Mol. Biol.*, vol. 4, pp. 32, 2005.

[10] D. Edwards, *Introduction to Graphical Modelling*, Springer, New York, 1995.

[11] B. Efron, “Large-scale simultaneous hypothesis testing: the choice of a null hypothesis,” *J. Amer. Statist. Assoc.*, vol. 99, pp. 96–104, 2004.