INFERRING GENE DEPENDENCY NETWORKS FROM GENOMIC LONGITUDINAL DATA: A FUNCTIONAL DATA APPROACH

Authors: RAINER OPGEN-RHEIN
–   Department of Statistics, University of Munich,
    Ludwigstrasse 33, D-80539 Munich, Germany
    opgen-rhein@stat.uni-muenchen.de

KORBINIAN STRIMMER
–   Department of Statistics, University of Munich,
    Ludwigstrasse 33, D-80539 Munich, Germany
    korbinian.strimmer@lmu.de

Abstract:

• A key aim of systems biology is to unravel the regulatory interactions among genes and gene products in a cell. Here we investigate a graphical model that treats the observed gene expression over time as realizations of random curves. This approach is centered around an estimator of dynamical pairwise correlation that takes account of the functional nature of the observed data. This allows to extend the graphical Gaussian modeling framework from i.i.d. data to analyze longitudinal genomic data. The new method is illustrated by analyzing highly replicated data from a genome experiment concerning the expression response of human T-cells to PMA and ionomicin treatment.

Key-Words:

• *graphical model; longitudinal data; dynamical correlation; gene dependency networks.*

AMS Subject Classification:

• 37N25, 62M10, 92B15, 92D10.

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., 8, 7, 11].

A graphical model is a representation of stochastic conditional dependencies between the investigated variables. Among the simplest graphical models is the class of graphical Gaussian models (GGMs) — see, e.g., Whittaker [13]. In this framework gene network may be constructed as follows. First, a positive definite and well-conditioned estimate $\boldsymbol{R} = (r_{kl})$ of the linear correlation matrix $\boldsymbol{P} = (\rho_{kl})$ is inferred from the data. Second, the standardized inverse of this matrix gives an estimate $\tilde{\boldsymbol{R}} = (\tilde{r}_{kl})$ of the *partial* correlations $\tilde{\boldsymbol{P}} = (\tilde{\rho}_{kl})$. The strength of these coefficients indicate the presence or absence of a direct association between each pair of genes. For large sample size computation of covariances and GGM selection can be conducted using classical estimation and testing theory as outlined in Whittaker [13]. However, the small sample size relative to the large number of genes typically considered in genome experiments requires the additional application of shrinkage and other regularization techniques [2, 12].

A drawback shared by the GGM approach and other graphical models such as Bayesian networks is that these methods rely on the assumption of identically and independently distributed (i.i.d.) data. However, an increasing proportion of microarray expression experiments are concerned with *longitudinal* measurements of mRNA and protein concentrations. For instance, stress response and cell cycle experiments by design produce time course data. A further characteristic of these data is that the time points at which the experiments are conducted are almost always not equidistant but irregularly spaced.

In order to avoid these issues, in this paper we investigate GGM network inference from the perspective of functional data analysis [9]. 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. This approach is based on the notion of *dynamical correlation* which provides a similarity score for pairs of groups of randomly sampled curves. Subsequently, it allows computation of partial dynamical correlations and the identification of the associated network structure.

The remainder of the paper is organized as follows. In the next section we summarize the basic notation for functional data analysis and also introduce the functional inner product. Next, we discuss the concept of dynamical correlation of which we describe two different variants, one introduced in this paper and one by Dubin and Müller [3]. Subsequently, the dynamical correlation is employed for GGM network selection. Finally, in order to compare the traditional GGM method with the present approach we reanalyze data from a human T-cell experiment with 58 genes, 10 time points, and 44 replications [10], and compare the networks resulting from dynamical correlation with those from static correlation.

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

$$(2.1) \qquad \langle g(t), h(t) \rangle = \frac{1}{B-A} \int_A^B g(t)\, h(t)\, dt$$

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. 2.1 turns into the weighted sum

$$(2.2) \qquad \langle g(t), h(t) \rangle \approx \sum_{j=1}^m g(t_j)\, h(t_j)\, \frac{\delta_j + \delta_{j+1}}{2(B-A)}$$

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 [3] each observed $f_{ik}(t)$ is a realization of the random function

$$(2.3) \qquad f_k(t) = \mu_k(t) + \mu_{0k} + \epsilon_{0k} + \sum_{u=1}^\infty \epsilon_{uk}\, \eta_u(t)\ ,$$

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. 2.3 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 [3] 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, compute 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

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

and

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

Finally, compute the 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

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

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 $f_{ik}^C(t) = f_{ik}(t) - \langle \bar{f}_k(t), 1 \rangle$. Note that here the inner product is computed over the mean function $\bar{f}_k(t)$. Based on the $f_{ik}^C(t)$ an estimate of the variance of variable k is then given by

$$(2.4) \qquad \widehat{\text{Var}}_k = \hat{\sigma}_{kk} = s_{kk} = \frac{1}{n-1} \sum_{i=1}^{n} \langle f_{ik}^C(t), f_{ik}^C(t) \rangle \ .$$

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

$$(2.5) \qquad \widehat{\mathrm{Cor}}_{kl} \;=\; \hat{\rho}_{kl} \;=\; r_{kl} \;=\; \frac{1}{n-1} \sum_{i=1}^{n} \left\langle f_{ik}^S(t), f_{il}^S(t) \right\rangle .$$

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

$$(2.6) \qquad \widehat{\mathrm{Cov}}_{kl} \;=\; \hat{\sigma}_{kl} \;=\; s_{kl} \;=\; r_{kl} \sqrt{s_{kk} s_{ll}} .$$

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. The Dubin–Müller definition of dynamical correlation

Another related but different definition of dynamical correlation is given by Dubin and Müller [3]. They propose to compute the standardized residual functions according to

$$(2.7) \qquad f_{ik}^S(t) \;=\; q_{ik}(t)/\sqrt{\left\langle q_{ik}(t), q_{ik}(t) \right\rangle}$$

using

$$(2.8) \qquad q_{ik}(t) \;=\; f_{ik}(t) - \bar{f}_{ik}(t) - \left\langle f_{ik}(t), 1 \right\rangle + \left\langle \bar{f}_{ik}(t), 1 \right\rangle .$$

This definition has the drawback that it is only defined if both $m > 1$ and $n > 1$. As we will exemplify below, it also produces counter-intuitive correlations.

2.3. Estimating gene association networks using dynamical correlation

The basic idea to infer a network from the pairwise dynamical correlation is to refer to the genes as the nodes and to the correlations as the connectivity strengths assigned to the edges of the network. However, we cannot use the correlations directly, because they 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

$$(2.9) \qquad\qquad \boldsymbol{\Omega} \;=\; \boldsymbol{P}^{-1} \;=\; (\omega_{ij})$$

$$(2.10) \qquad\qquad \tilde{\rho}_{kl} \;=\; -\frac{\omega_{kl}}{\sqrt{\omega_{kk}\,\omega_{ll}}}$$

[4]. 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 as proposed by Schäfer and Strimmer [11, 12]. 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 proposed by Efron [5, 6]. This method computes 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

In the following section we first apply our method of computing dynamical correlation to example data to clarify our definition and to compare it with the related concept of Dubin and Müller [3]. Subsequently, we infer the gene association network for a longitudinal gene expression data set described in Rangel et al. [10].

3.1. Illustrative example

In order to understand the concept of dynamical correlation and to illustrate the difference between our definition (Eq. 2.5) and that of Dubin and Müller [3] we first consider a set of artificial examples. These are shown in Fig. 1 where two negatively dependent variables are depicted. For instance, this may represent the case where one gene is up-regulated and the other is correspondingly down-regulated. For each gene there are two measured curves, and there are three slightly different ways in which the sampled curves relate to each other (Fig. 1a, b, and c). The exact definition of the curves can be found in Tab. 1. Note that the two realizations are paired, i.e. the upper lines belong to individual 1 and the lower ones to individual 2.

a)

b)   **c)**
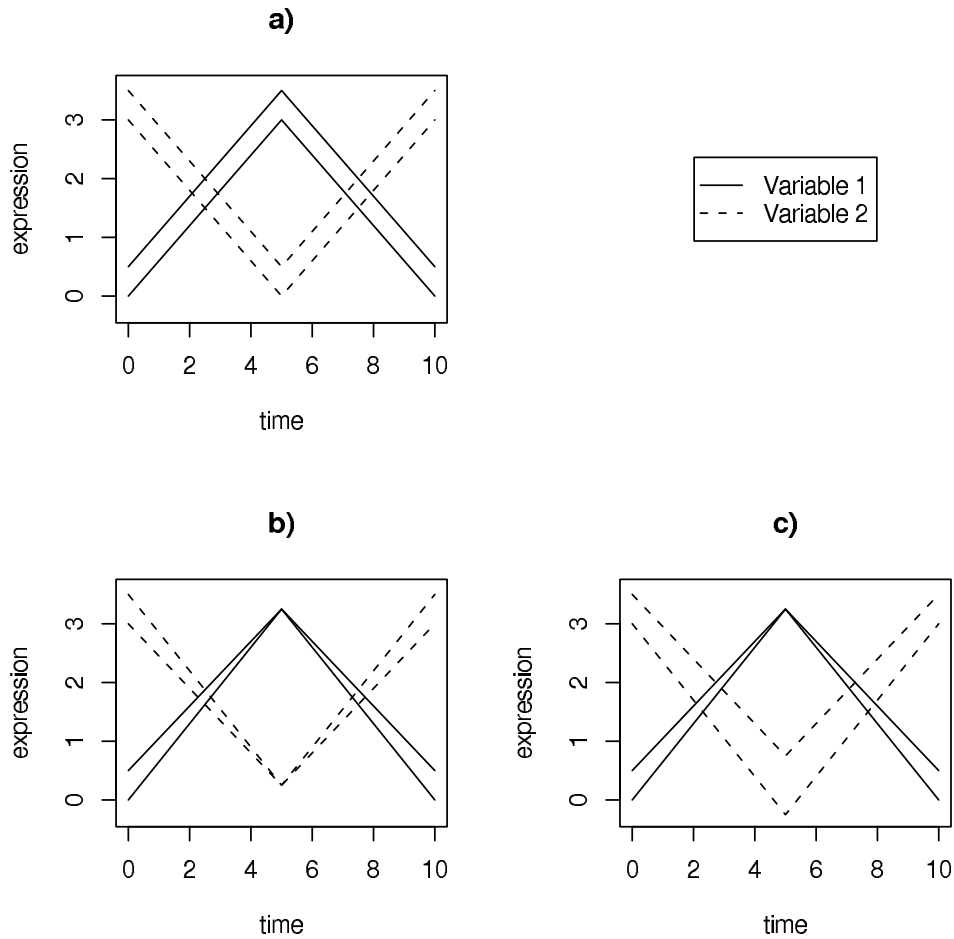
Figure 1:   Toy example to illustrate the concept of dynamical correlation
between two variables ("genes"). In all three cases a), b) and c)
there are two realizations ("individuals"). See main text for details,
and Tab. 1 for the underlying data.

Table 1:   Data points of the toy examples in Fig. 1.

Data		**Variable 1**			**Variable 2**		
Time points		*0*	*5*	*10*	*0*	*5*	*10*
Fig. 1a	*Realization 1*	0	3	0	3	0	3
	Realization 2	0.5	3.5	0.5	3.5	0.5	3.5
Fig. 1b	*Realization 1*	0	3.25	0	3	0.25	3
	Realization 2	0.5	3.25	0.5	3.5	0.25	3.5
Fig. 1c	*Realization 1*	0	3.25	0	3	-0.25	3
	Realization 2	0.5	3.25	0.5	3.5	0.75	3.5

Intuitively, one would expect that the dynamical correlation between the two variables is strongly negative in all three cases. For our definition of dynamical correlation according to Eq. 2.5 this is indeed the case: the correlations for the three examples cases Fig. 1a, b, and c are -0.946, -0.982, and -0.947, respectively. In contrast, the dynamical correlation of Dubin and Müller [3] behaves in a completely different fashion. For Fig. 1a it is not defined, for case b) it is equal to $+1$ and for case c) it is equal to -1.

Therefore, it is easy to see that the Dubin and Müller [3] estimator is *not* suited for detecting functional dependencies in genomic longitudinal data. This is because that estimator is geared towards detecting changes in the relative trends of the individual realizations, rather than between the common trend. However, note that this is generally not the effect one wants to identify when looking for gene interaction. In addition, the Dubin and Müller [3] definition of dynamical correlation has the additional disadvantage over that of Eq. 2.5 that it is not defined if there is only a single time course per gene available. In contrast, the above toy examples show that our definition of dynamical correlation is able to detect the main trend of positive or negative dependency between two variable, and is not susceptible to the small changes in the sampled curves.

3.2. Gene expression time course data

We now employ our method of 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 described in detail in Rangel et al. [10]. These data characterize the response of a human T-cell line (Jirkat) to a treatment with PMA and ioconomin. After preprocessing the time course data consist of 58 genes measured across 10 time points with 44 replications. Rangel et al. [10] used a state space model to estimate the influence between genes and measured a genetic network by combining direct effects and indirect effects via hidden states. This approach is generally very time-consuming due to the necessity of using of the EM algorithm for optimization. A peculiarity of the Rangel et al. [10] data is also that the measurements in the experiment were taken at unequally spaced time points, i.e. after 0, 2, 4, 6, 8, 18, 24, 32, 48, and 72 hours after treatment. This was neglected in the original state-space analysis which assumed equally spaced data. In contrast, note that the present functional data approach allows the incorporation of arbitrary time distances between subsequent measurements.

As approximation of the temporal expression of the 58 genes we used a linear spline and employed Eq. 2.2 for the functional inner product. After estimating the dynamical correlations with Eq. 2.5 we computed the associated partial correlation coefficients employing Eq. 2.9 and Eq. 2.10. Fig. 2 shows the histogram of the estimated partial correlation coefficients after Fisher's normaliz-

ing z-transformation. Also depicted in this plot are the fitted overall distribution (fat line) and the null (dashed line) and alternative distribution (filled histogram) as estimated by the locfdr algorithm [5, 6]. The 0.2 local fdr cut-off values for the partial correlations are indicated by the black triangles. As expected, the distribution of the partial correlations is centered around zero and most of the coefficients are not significant. Consequently, the resulting network is sparse and there are only 54 significant edges. The network itself is displayed in Fig. 3b.

delhat= 0.007 sighat= 0.082 p0= 0.871
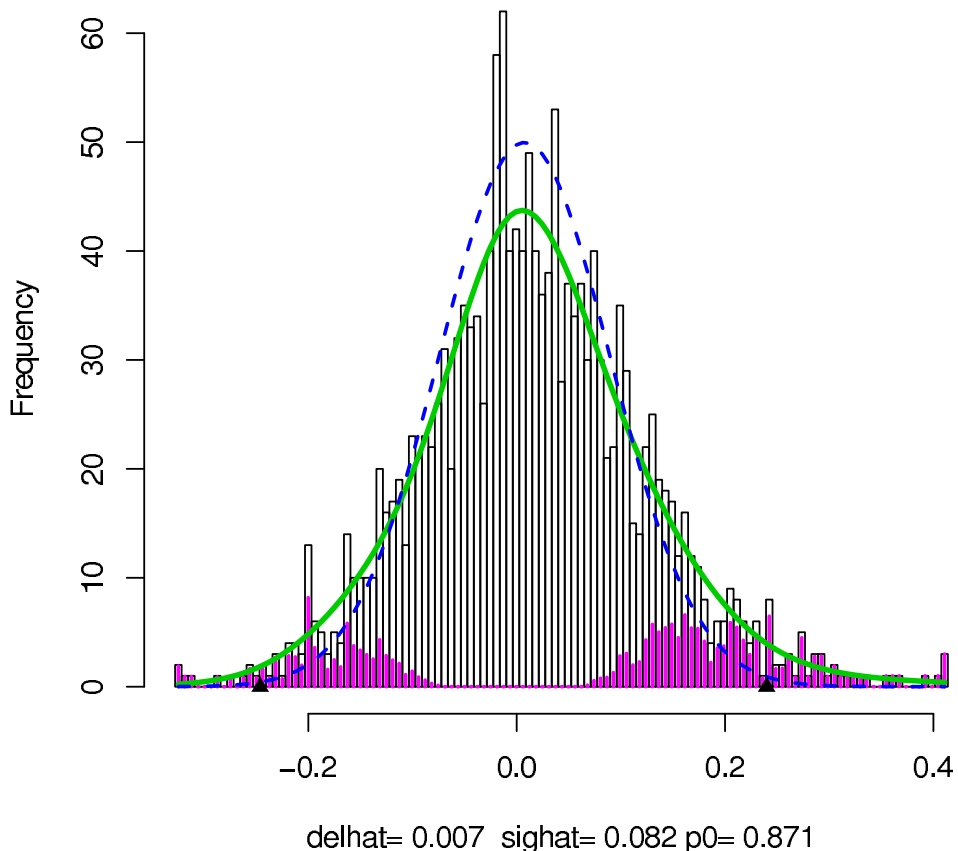
Figure 2:   Histogramm of the Fisher z-transformed estimated partial dynamical correlations. Values left and right the two black triangles are considered significantly different from zero, and thus correspond to edges in a gene dependency network.

It is instructive to compare the genetic network inferred with dynamical correlation to the gene association network 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, which leads to 440 observations (44 replications times 10 time points) for each of the 58 genes. As this number of observations is not small in comparison to the number of the genes no regularization is needed (cf. Schäfer and Strimmer [12] for the opposite case).

From the empirical correlation matrix we proceeded as above, obtaining estimates of partial correlation and a static GGM network. This is displayed on the left side of figure 3. For comparison, the network estimated with dynamical correlation in shown on its right side. For clarity only the nodes which have at least one connection are displayed.
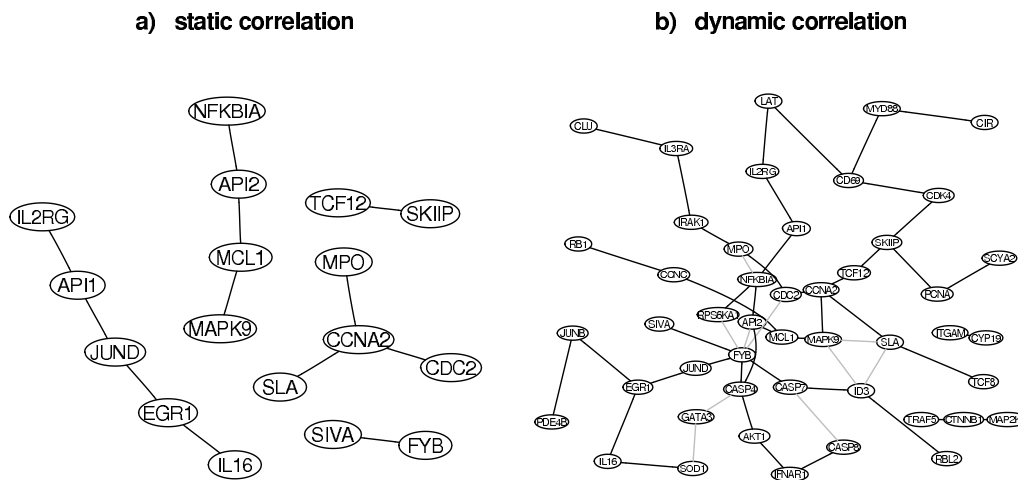
Figure 3: Gene dependency networks inferred from human T-cell data [10] using (a) static correlation and (b) dynamical correlation.

The network calculated with static correlation consists of 17 nodes with 12 edges, a smaller network than the one based on dynamical correlation. This indicates that our dynamical estimator is able to identify additional time-varying components of the interaction between the investigated genes.

4. DISCUSSION

A growing interest in genetics lies in observing and inferring the gene interactions over time. Here, we introduced a method to infer a gene dependency network from functional data. In this approach time course experiments are seen as a realization of random curves. The method described generalizes the widely used static GGM approach (see the corresponding references in [11]) and is able to unravel the dependency structure of longitudinal data across the whole time series rather than at single time points. Furthermore, unlike many other time series method the functional approach does not require equally spaced measurements. In addition, our algorithm is easily implemented and computationally inexpensive (the calculation of the above gene dependency network takes only a fraction of a second).

   In order to further develop our approach many extensions are conceivable. For instance, in the above analysis of human T-cells the data was highly replicated. In genomics, however, it is more typical that the sample size is very small compared to the number of genes (this is the so-called "small n, large p" paradigm). In this case, the empirical covariance is a highly inefficient estimator, and needs to be regularized [12]. For small n this will also be the case with our estimate of dynamical correlation (Eq. 2.5). Thus, shrinkage techniques similar to those of Schäfer and Strimmer [12] are needed.

   A further important extension is the inclusion of autoregressive aspects [1]. 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. This is illustrated in Fig. 4 which is a variation of the toy examples presented in section 3. For this data the Dubin and Müller [3] estimate is (again) not defined and our suggested dynamical estimator results in very small correlation close to zero, even though it is clear by inspection that the two depicted variables are strongly connected. These dependencies and the associated time shifts could be accounted for by modeling the temporal mean via a system of differential equation (or in the discrete case by some autoregressive process). We also note that for this reason we have also refrained here from a comparison of the gene association network inferred from dynamical correlation (Fig. 3b) with the state space network presented by Rangel et al. [10]. Future work should regard for these aspects.
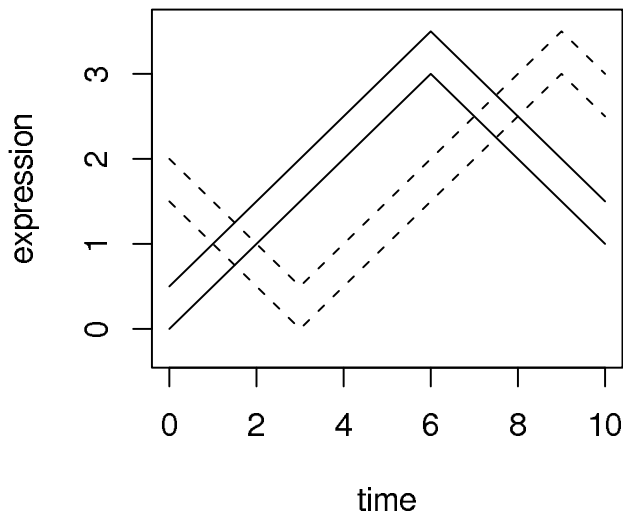
Figure 4:   Example with a fixed time lag between the two variables.

ACKNOWLEDGMENTS

REFERENCES

[1]   DIGGLE, P.J.; HEAGERTY P.J.; LIANG K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data, 2nd Edition*, Oxford: Oxford University Press.

[2]   DOBRA, A.; HANS, C.; JONES, B.; NEVINS, J. R.; YAO, G. and WEST M. (2004). Sparse graphical models for exploring gene expression data, *J. Multiv. Anal.*, **90**, 196–212.

[3]   DUBIN, J. A. and MÜLLER, H.-G. (2005). Dynamical correlation for multivariate longitudinal data, *J. Amer. Statist. Assoc.*, **100**, 872–881.

[4]   EDWARDS, D. (1995). *Introduction to Graphical Modelling*, New York: Springer.

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

[6]   EFRON, B. (2005). *Local false discovery rates*, Technical Report, Dept. of Statistics, Stanford University.

[7]   FRIEDMAN, N. (2004). Inferring cellular networks using probabilistics graphical models, *Science*, **303**, 799–805.

[8]   HARTEMINK, A.J.; GIFFORD, D.K.; JAAKKOLA, T.S. and YOUNG, R.A. (2002). Bayesian methods for elucidating genetic regulatory networks, *IEEE Intell. Systems*, **17**, 37–43.

[9]   RAMSAY, J.O. and SILVERMAN B.W. (2005). *Functional Data Analysis, 2nd Edition*, New York: Springer Verlag.

[10] RANGEL, C.; ANGUS, J.; GHAHRAMANI, Z.; LIOUMI, M.; SOTHERAN, E.; GAIBA, A.; WILD, D.L. and FALCIANI, F. (2004). Modeling T-cell activation using gene expression profiling and state space modeling, *Bioinformatics*, **20**, 1361–1372.

[11] SCHÄFER, J. and STRIMMER, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics*, **21**, 754–764.

[12] SCHÄFER, J. and STRIMMER, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Statist. Appl. Genet. Mol. Biol.*, **4**, 32.

[13] WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*, New York: Wiley.