Dynamical Correlation for Multivariate Longitudinal Data

Joel A. Dubin and Hans-Georg Müller

Joel A. Dubin is Assistant Professor, Division of Biostatistics, Yale University, New Haven, CT 06520 (e-mail: *joel.dubin@yale.edu*). Hans-Georg Müller is Professor, Department of Statistics, University of California, Davis, Davis, CA 95616 (e-mail: *mueller@wald.ucdavis.edu*). This research was supported in part by NSF Grants DMS99-71602, DMS02-04869 and DMS03-54448. The authors wish to thank Prof. G. Kaysen for many discussions on the biomedical aspects of the longitudinal protein data. Also, the authors very much appreciate the comments and recommendations of the editor, associate editor, and referees, which significantly improved the paper.

ABSTRACT

Nonparametric methodology for longitudinal data analysis is becoming increasingly popular. The analysis of *multivariate longitudinal data,* where data on several time courses are recorded per subject, has received considerably less attention, in spite of its importance for practical data analysis. In particular, there is a need for measures and estimates to capture dependency between the components of vector-valued longitudinal data. We propose and analyze a simple and effective nonparametric method to quantify the covariation of components of multivariate longitudinal observations, which are viewed as realizations of a random process. This includes the notion of a correlation between derivatives and time-shifted versions. The concept of *dynamical correlation* is based on a scalar product obtained from pairs of standardized smoothed curves. The proposed method can be utilized when observation times are irregular and not matching between subjects or between responses within subject. For higher-dimensional data, one may construct a dynamical correlation matrix which then serves as a starting point for standard multivariate analysis techniques such as principal components. Our methods are illustrated via simulations as well as with data on five acute phase blood proteins measured longitudinally from a study of hemodialysis patients.

KEY WORDS: Acute phase proteins; Curve data; Dependency; Random effects model; Smoothing; Stochastic process.

1. INTRODUCTION

Longitudinal data modeling is essential to describe both trend and variation for biological processes, such as growth curves, effects over time of medical intervention on physiological characteristics, monitoring human exposure to carcinogens, etc. Methods and approaches are described in texts such as Jones (1993), Hand and Crowder (1996), Verbeke and Molenberghs (2000), and Diggle *et al.,* (2002).

A promising approach for functional data analysis is to treat longitudinal pathways as realizations of a smooth stochastic process (Ramsay and Silverman, 1997). This concept proved useful for describing the effects of certain treatments on a response curve (Church 1966), and naturally progressed to the modeling of a collection of random curves (Rice and Silverman, 1991), and semiparametric and nonparametric models for the effects of time-dependent covariates on longitudinal observations (Martinussen and Scheike, 1999 and 2000).

In this paper, we describe an approach for capturing the correlation structure between multivariate longitudinal responses, leading to the notion of *dynamical correlation* to describe the correlation amongst multivariate longitudinal curves. A classical approach to describe the correlation between subsets of elements of random vectors is canonical correlation (Hotelling, 1936). Canonical correlation has been extended to the case of multivariate time series (Brillinger, 1975), under the assumption of stationarity, and extension of canonical correlation to functional data has been proposed in Leurgans, Moyeed, and Silverman (1993), where the need for regularization was pointed out. Moreover, functional canonical correlation requires restrictive assumptions in order to be well defined, as it corresponds to an inverse problem (He, Müller, and Wang, 2003). For these reasons, simple and efficient alternative measures to describe the dependency of multivariate functional data are needed (Service, Rice, and Chavez 1998).

Dynamical factor analysis has been discussed in the psychological literature to investigate intra-individual variation and lagged relationships for multivariate longitudinal data (Molenaar, 1985). However, these methods require restrictive designs and are applied to a single individual only; extensions to samples of subjects have not been established. Methods based on the notion of causality between the components of multivariate stochastic processes have been discussed by Boudjellaba, Dufour, and Roy (1992) and Sy, Taylor, and Cumberland (1997). Like the approaches by Molenaar (1985) and Brillinger (1975), the method of Boudjellaba *et al.,* (1992) is for multivariate time series and does not generalize to multiple subject longitudinal data; Sy *et al.,* (1997) rely on fairly restrictive assumptions on the nature of the correlations and subsequent determination of causality. A spline-based method for modeling bivariate longitudinal data, including investigation of the correlation between responses, has been presented by Wang, Guo, and Brown (2000).

It is the purpose of this article to define simple, efficient, non-parametric correlation measures for multivariate longitudinal data, which include derivatives and lags. These measures are first obtained at the subject level, and consistent estimates for population dynamical correlations are

then easily obtained by averaging over the subjects in the sample.

An advantage of the proposed dynamical correlation over functional canonical correlation is its explicit representation given by (5), whereas functional canonical correlation is implicitly defined by (12) via the solution of a maximization problem, and therefore does not have a comparably simple interpretation as an average of individual correlations. Section 5.2 below is devoted to a more detailed comparison of the practical performance of dynamic and functional canonical correlation. An additional benefit of dynamical correlation is its stability. As we show below, even if a pre-smoothing step is included, dynamical correlation is quite insensitive to the choice of bandwidth for this step, while the estimation of functional canonical correlation critically depends on regularization, as it corresponds to an ill-defined inverse problem (Leurgans *et al.,* 1993). Also, as shown below, these regularized correlation estimates easily break down and then are not useful, while dynamic correlation generally is found to lead to reliable results. Related time-averaged correlation measures between two regression functions were discussed by Heckman and Zamar (2000). A limitation of the proposed dynamical correlation, especially in the application to longitudinal studies, is that the number of repeated measurements per subject cannot be too small, and that the times of measurements need to fill out the domain of the random trajectories for which the correlation measure is desired.

The data used to demonstrate the methods in this paper come from a nephrological study (Kaysen *et al.,* 2000). Thirty-five hemodialysis patients were followed for up to 230 days, with measurements of five acute phase blood proteins taken longitudinally. Observed repeated measurements for two serum proteins, albumin and C reactive protein, for a randomly selected subject, are shown in Figure 1. The graphs are suggestive of a negative relationship over time for these two acute phase proteins. While such simple graphical representations are useful, it is important to have quantitative summary measures of correlation taking the entire variation over time into account.

The paper is organized as follows. The underlying model and basic definition of dynamical correlation between two components of random multivariate longitudinal curves are discussed in Section 2. In Section 3, we describe how to estimate the dynamical correlation between two sets of longitudinal data obtained for a sample of independent subjects. In Section 4.1, we provide extensions of dynamical correlation to derivatives of curves and time-shifted curves. In Section 4.2, we discuss a two-stage bootstrap procedure for obtaining a non-parametric interval estimate for the correlation measure when smoothing of the original data is required. We report results of simulation studies in Section 5, including a sensitivity analysis of the correlation estimate under a range of bandwidth choices, and a comparison of the performance of functional canonical correlation and dynamical correlation. The application of the proposed methods to the blood protein data is the topic of Section 6. Concluding remarks can be found in Section 7. Finally, additional details and proofs are provided in the Appendix.

2. DEFINING DYNAMICAL CORRELATION

In this section, we define dynamical correlation and discuss some basic properties. The setting is as follows: For a randomly selected subject (experimental unit), we observe p random functions or curves, $f_1, ...f_p$, $p \geq 2$, where $f_k \in L^2(dw)$, the space of square integrable functions with $E\{\int f_k^2(t)w(t)dt\} < \infty$ for $1 \leq k \leq p$, with respect to a measure $dw = w\ dt$, where dt is Lebesgue measure and w is a nonnegative weight function with $\int w(t)dt = 1$ and $\int w^2(t)dt < \infty$ (see Ash and Gardner, 1975). Usually, w will be chosen to have compact support.

A simple and convenient choice for the weight function w is an indicator function, $w(t) = \frac{1}{b-a} I_{[a,b]}$, for $a < b$. Other choices may relate to a varying degree of uncertainty with which the functions are observed over time, to a non-constant variance function of the underlying stochastic process, or to inhomogeneities in the design, i.e., the timing of the longitudinal measurements.

The notions of inner products and angles between functions that we will use are extensions from the familiar multivariate concepts to Hilbert space (Conway, 1985; Ramsay and Silverman, 1997). Using the notation $\langle f, g\rangle = \int f(t)g(t)w(t)dt$ in $L^2(dw)$, our approach is based on the "functional inner products"

$$M_k = \langle f_k, 1\rangle, \quad M_{k,l} = \langle f_k, f_l\rangle, \quad 1 \leq k, l \leq p. \tag{1}$$

The basic assumptions ensure that the moments EM_k, EM_k^2, and $EM_{k,l}$ are all well-defined.

Any given L^2 function can be represented uniquely in L^2 through the following random effects model, without any restriction, and it will be convenient to define dynamical correlation within the framework of this representation. In particular the k-th functional component always can be represented as

$$f_k(t) = \mu_k(t) + \mu_{0,k} + \sum_{r=0}^{\infty} \varepsilon_{r,k}\eta_r(t) = \mu_k(t) + (\mu_{0,k} + \varepsilon_{0,k}) + \sum_{r=1}^{\infty} \varepsilon_{r,k}\eta_r(t), \quad 1 \leq k \leq p. \tag{2}$$

Here, μ_k is a fixed mean function with $\mu_k \in L^2(dw)$ and $\langle \mu_k, 1\rangle = 0$, and $(\mu_{0,k} + \varepsilon_{0,k})$ is an intercept term, the "static random part" of the model, which includes a constant term $\mu_{0,k}$, and a random variable $\varepsilon_{0,k}$, neither of which depend on time, with $E(\varepsilon_{0,k}) = 0$ and $\mathrm{var}(\varepsilon_{0,k}) = \sigma_{0,k}^2 < \infty$, where $\eta_0(t) \equiv 1$. Lastly, $\sum_{r=1}^{\infty} \varepsilon_{r,k}\eta_r(t)$ is the "dynamic random part" of the model, where the random variables $\varepsilon_{r,k}$, $r \geq 1$, satisfy $E\varepsilon_{r,k} = 0$ for all r, $E(\varepsilon_{r,k}^2) = \sigma_{r,k}^2 < \infty$, $0 < \sum_{r=0}^{\infty} \sigma_{r,k}^2 < \infty$, and $\varepsilon_{0,k}$ is uncorrelated with $\varepsilon_{r,k}$, $r \geq 1$. The functions η_r, $r = 0, 1, ...$, are assumed to form an orthonormal basis of $L^2(dw)$, i.e., $\langle \eta_i, \eta_j\rangle = 0$ for $i \neq j$ and $\langle \eta_i, \eta_j\rangle = 1$ for all $i = j$. The functions μ_k, η_r, $1 \leq k \leq p$, $0 \leq r < \infty$, are furthermore assumed to be smooth, say twice continuously differentiable. We also assume that all random variables $\varepsilon_{r,j}, 0 \leq r < \infty, 1 \leq k \leq p$ have a joint distribution.

As we will demonstrate below, the dependence between the functional components f_k and f_l is essentially captured by the quantity $S_{k,l} = \sum_{r=1}^{\infty} \varepsilon_{r,k}\varepsilon_{r,l}$. We note that in model (2), for $1 \le k \le p$, $E(f_k(t)) = \mu_k(t) + \mu_{0,k}, \quad E(\langle f_k, 1\rangle) = \mu_{0,k},$ and

$$M_k = (\mu_{0,k} + \varepsilon_{0,k}) + \sum_{r=1}^{\infty} \varepsilon_{r,k}\langle \eta_r, 1\rangle = \mu_{0,k} + \varepsilon_{0,k}, \tag{3}$$

since $\langle \eta_r, \eta_0\rangle = 0$ for $r \ge 1$. This implies that the centered version, $f_k(t) - M_k = \mu_k(t) + \sum_{r=1}^{\infty} \varepsilon_{r,k}\eta_r(t)$, of f_k does not anymore involve the intercept term.

Defining standardized curves

$$f_k^*(t) = \frac{f_k(t) - M_k - \mu_k(t)}{(\int (f_k(t) - M_k - \mu_k(t))^2 w(t)dt)^{1/2}}, \tag{4}$$

we then obtain the representation

$$f_k^*(t) = \sum_{r=1}^{\infty} \varepsilon_{r,k}\eta_r(t) / (\sum_{r=1}^{\infty} \varepsilon_{r,k}^2)^{1/2}.$$

Since $\langle \eta_r, 1\rangle = 0$, this implies $\langle f_k^*, 1\rangle = 0$, and the fact that $\langle \eta_i, \eta_j\rangle = 0$, $i \ne j$, and $\langle \eta_i, \eta_i\rangle = 1$ implies $\langle f_k^*, f_k^*\rangle = 1$. In this sense, $f_k^*(t)$ is a standardized version of f_k. We note that this standardization applies to each single realization of f_k. If $\mu_k \equiv$ const., i.e., the mean function is a constant, which is a reasonable approximation for many applications, then $\langle \mu_k, 1\rangle = 0$ implies that $\mu_k \equiv 0$. In this case the standardization simplifies to

$$f_k^*(t) = \frac{f_k(t) - M_k}{(\int (f_k(t) - M_k)^2 w(t)dt)^{1/2}}.$$

The proposed dynamical correlation between f_k and f_l is now defined as the expected cosine of the angle between the standardized versions f_k^* and f_l^*,

$$\rho_{k,l} = E\langle f_k^*, f_l^*\rangle = E\frac{\langle f_k^*, f_l^*\rangle}{\{\langle f_k^*, f_k^*\rangle \langle f_l^*, f_l^*\rangle\}^{1/2}}. \tag{5}$$

As shown in the Appendix, dynamical correlation can be represented as follows:

Theorem 1.

$$\rho_{k,l} = E\{\sum_{r=1}^{\infty} \varepsilon_{r,k}\varepsilon_{r,l} / [(\sum_{r=1}^{\infty} \varepsilon_{r,k}^2)^{1/2} (\sum_{r=1}^{\infty} \varepsilon_{r,l}^2)^{1/2}]\}. \tag{6}$$

We find that $\rho_{k,l}$ captures the dependency between f_k and f_l through a standardized form of $S_{k,l} = \sum_{r=1}^{\infty} \varepsilon_{r,k}\varepsilon_{r,l}$, namely $\rho_{k,l} = E\frac{S_{k,l}}{[S_{k,k}S_{l,l}]^{1/2}}$. If we denote the angle in $L^2(dw)$ between the dynamic random parts $(\sum \varepsilon_{r,k}\eta_r, \sum \varepsilon_{r,l}\eta_r)$ by $\zeta_{k,l}$, then $\rho_{k,l} = E(\cos\, \zeta_{k,l})$. Hence, the dynamical correlation between random curves f_k and f_l can be directly linked to the cosine of an angle between functions in the Hilbert space L^2. It immediately follows from (6) that $-1 \le \rho_{k,l} \le 1$ by the Cauchy-Schwarz inequality.

We may interpret dynamical correlation as a measure of average concordant or discordant behavior of pairs of random trajectories, in the sense that if both trajectories tend to be mostly

on the same side of their time-average (a constant), then dynamical correlation is positive; if the opposite occurs, then negative. For example, if we deal with pairs of random trajectories f_k, f_l, for which we may assume $M_k = M_l = 0$ and $\int f_k^2 = \int f_l^2 = 1$ so that $f_k^* = f_k$, $f_l^* = f_l$, then $\rho_{k,l} = 1$ if $f_k = f_l$ and $\rho_{k,l} = -1$ if $f_k = -f_l$. As a special case, consider $f_k(t) = \sqrt{2}\sin(2\pi(t - Y_k))$, $f_l(t) = \sqrt{2}\sin(2\pi(t - Y_l)), t \in [0, 1]$, where Y_k, Y_l are random variables that determine the corresponding random trajectories. A simple calculation shows $\rho_{k,l} = E(\cos[2\pi(Y_k - Y_l)])$, and if, for example, $(Y_k - Y_l) \sim \text{Uniform}[-\frac{1}{2}, \frac{1}{2}]$, then the time shift between the two sine curves relative to each other is uniform, and indeed $\rho_{k,l} = 0$ for this case. If instead $(Y_k - Y_l) \sim \text{Uniform}[-\frac{1}{4}, \frac{1}{4}]$, then the shift between the sine curves is limited and in this case we find $\rho_{k,l} = \frac{2}{\pi}$. Further examples can be constructed along these lines.

We conclude this section with two remarks. First, defining dynamical correlation with the expectation as the "outside" operator, in contrast to the usual definition of correlation, leads to a stable, simple and computationally fast procedure for estimation that is easy to implement and helps avoid the calculation of an inverse which may lead to an ill-posed problem such as in functional canonical correlation. Second, definition (5) is independent of representation (2), and there is no need to actually identify the functions $\eta_r,\ r \geq 0$, which merely serve as components of the underlying modeling framework.

3. DYNAMICAL CORRELATION ESTIMATION

To estimate the dynamical correlation, $\rho_{k,l}$ (5), we assume one has multivariate longitudinal data collected for a sample of n subjects. For the i-th subject, one observes random curves $f_{i,1}(t), f_{i,2}(t), ..., f_{i,p}(t)$ on $[0, T_i]$. We then define the estimated k-th standardized curve for the i-th subject $(1 \leq k \leq p, 1 \leq i \leq n)$ by

$$\hat{f}_{i,k}^*(t) = \frac{\breve{f}_{i,k}(t) - \breve{M}_{i,k}}{(\int(\breve{f}_{i,k}(t) - \breve{M}_{i,k})^2 w(t)dt)^{1/2}}\ , \tag{7}$$

where $\breve{f}_{i,k}(t) = f_{i,k}(t) - (1/n)\sum f_{i,k}(t)$ and $\breve{M}_{i,k} = \langle \breve{f}_{i,k}, 1\rangle$. Since $\breve{f}_{i,k}(t) - \breve{M}_{i,k} = f_{i,k}(t) - M_{i,k} - (1/n)\sum(f_{i,k}(t) - M_{i,k})$ and $E(f_{i,k}(t) - M_{i,k}) = \mu_k(t)$, (7) provides a plausible estimate for (4). Here, $\breve{M}_{i,k}$ and $\int(\breve{f}_{i,k} - \breve{M}_{i,k})^2 w(t)dt$ are evaluated by numerical integration.

We then obtain an estimate for the i-th individual's dynamical correlation for components k and l,

$$\hat{\rho}_{k,l,i} = \langle \hat{f}_{i,k}^*, \hat{f}_{i,l}^* \rangle \tag{8}$$

and an estimate of the overall dynamical correlation by averaging the individual dynamical correlations over subjects:

$$\hat{\rho}_{k,l} = \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}_{i,k,l} = \frac{1}{n}\sum_{i=1}^{n} \langle \hat{f}_{i,k}^*, \hat{f}_{i,l}^* \rangle. \tag{9}$$

An added benefit of this approach is that one obtains a measure of dynamical correlation at the individual level. Note that we may estimate $\mu_k(t)$ by $\hat{\mu}_k(t) = \frac{1}{n}\sum_{i=1}^{n}(f_{i,k}(t) - M_{i,k})$.

In some applications, such as gait analysis (Olshen *et al.,* 1989), one observes entire trajectories $f_{i,k}$, while in other cases, such as the application to longitudinal dynamical relationships between acute phase proteins that we explore below, the data are only available in the form of discrete measurements. We then include a pre-smoothing step; the effect of pre-smoothing is explored in the simulations reported in Section 5. If a pre-smoothing step is included, we denote the resulting quantities by a superscript, so that $f_{i,k}, \hat{f}^*_{i,k}, \hat{\rho}_{k,l,i}, \hat{\rho}_{k,l}$ become $f^S_{i,k}, \hat{f}^{*,S}_{i,k}, \hat{\rho}^S_{k,l,i}, \hat{\rho}^S_{k,l}$.

Our main result refers to the asymptotic distribution of dynamical correlation $\hat{\rho}^S_{k,l}$:

Theorem 2. Under regularity conditions (see Appendix), if $0 < |\rho_{k,l}| < 1$, and if $\mu_k(t)$, for all k, is constant or known, then

$$n^{1/2}(\hat{\rho}^S_{k,l} - \rho_{k,l}) \longrightarrow N(0, \delta^2_{k,l})$$

in distribution, as $n \longrightarrow \infty$, where $\delta^2_{k,l} = \mathrm{var}(\langle f^*_k, f^*_l \rangle)$. If $\mu_k(t)$ is neither constant nor known, then $n^{1/2}|\hat{\rho}^S_{k,l} - \rho_{k,l}| = O_p(1)$.

The proof is in the Appendix. We may estimate the variance of $\hat{\rho}_{k,l}$ simply by the empirical variance of $\{\hat{\rho}_{k,l,i}, i = 1, ..., n\}$, possibly including extra variance due to smoothing. This may be used to calculate large sample confidence intervals and tests of, for example, $H_o : \rho = 0$. Alternatively, one may construct bootstrap confidence intervals, obtaining bootstrap samples by resampling with replacement from $\{\hat{\rho}_{k,l,i}, i = 1, ..., n\}$ (Efron and Tibshirani, 1993). The construction of bootstrap confidence intervals is discussed in Section 4.2. Such resampling will be reasonable if $\rho_{k,l}$ is not too extreme, i.e., close to 1 in absolute value. If $\rho_{k,l}$ is extreme, one strategy might be to employ the Fisher transformation (Fisher, 1921), $z = (1/2)[\ln(1 + r) - \ln(1 - r)]$, to the $\{\hat{\rho}_{k,l,i}, i = 1, ..., n\}$, where r here is the estimated dynamical correlation (see also Efron, 1998).

In the assumptions of Theorem 2, the number of repeated measurements is required to be large compared to the number of subjects from which dynamical correlation is calculated. This is needed to accommodate a sufficiently fast rate of convergence for the smoothing step. This in itself of course does not mean that dynamical correlation will not work if this assumption is not satisfied. In fact, the small sample simulation reported near the end of Section 5.1 shows that the results can be quite good for moderate numbers of repeated measurements as we encounter them in the data example in Section 6. However, caution is indicated when the number of repeated measurements is truly small. Additional comments on this point can be found toward the end of the Discussion section.

Observe that case weights can be easily incorporated in the estimation of $\rho_{k,l}$. For example, one may use a weighted average in lieu of the sample mean when estimating $\rho_{k,l}$, where the weights might reflect the number of measurements available for patient i. Then, greater weight

would be given to those subjects for whom one has more repeated measurements when calculating the dynamical correlation estimate. Another comment is that in the estimation of $\rho_{k,l}$, there is no need to specify the limit of the sum or define the exact form of the orthonormal basis in $\sum \varepsilon_{r,k}\eta_r(t)$.

4. EXTENSIONS INCLUDING DERIVATIVES, LAGS, AND INTERVAL ESTIMATION AFTER SMOOTHING

4.1 Incorporating derivatives and lag

We can extend the concepts introduced in section 2 to incorporate derivatives of curves as well as a lag term. Specifically, for derivative orders ν, ν_1 and ν_2, and lag term τ, we can more generally define the quantities in (1) as $M_k = \langle f_k^{(\nu)}, 1\rangle$ and $M_{k,l} = \langle f_k^{(\nu_1)}, f_{l,\tau}^{(\nu_2)}\rangle$, where $f_{l,\tau}^{(\nu_2)} = f_l^{(\nu_2)}(t-\tau)$. In the case of lag $\tau \neq 0$, some caution needs to be exercised near the ends of the data, in order to ensure that both $f_k^{(\nu_1)}(t)$ and $f_l^{(\nu_2)}(t-\tau)$ are well-defined. For example, a suitable weight function for lag τ would be $w_\tau(t) = \frac{1}{(b-a-|\tau|)} I_{[\max(a,a+\tau),\min(b+\tau,b)]}(t)$. To choose a common weight function for all possible values of τ, one would choose the maximally occurring τ value.

The steps to obtain a standardized curve, as outlined in Sections 2 and 3, are then repeated with $f_{i,k}$ replaced by $f_{i,k}^{(\nu)}$. The curves $f_{i,k}^{(\nu)}$ would typically be obtained by a nonparametric derivative estimate, obtained by one of various derivative estimation methods, for example using the kernel method (Müller, Stadtmüller, and Schmidt, 1987) or local polynomial fitting (Fan and Gijbels, 1996). The dynamical correlation between two random curves (or derivatives) incorporating lag τ is then defined as

$$\rho_{k,l,\nu_1,\nu_2,\tau} = E\{\langle f_k^{(\nu_1)*}, f_{l,\tau}^{(\nu_2)*}\rangle\}, \tag{10}$$

where $M_{l,\tau}^{(\nu)} = \langle \hat{f}_{l,\tau}^{(\nu)}, 1\rangle$, and
$\hat{f}_{l,\tau}^{(\nu)*}(t) = \{f_{l,\tau}^{(\nu)}(t) - M_{l,\tau}^{(\nu)} - \mu_{l,\tau}^{(\nu)}(t)\}/\{\int (f_k^{(\nu)}(t) - M_{l,\tau}^{(\nu)} - \mu_{l,\tau}^{(\nu)}(t))^2 w(t)dt\}^{1/2}$.

We estimate the dynamical correlation between two standardized random curves or derivatives with lag τ analogously as before, by first defining the dynamical correlation for the i-th individual, $\hat{\rho}_{k,l,i,\nu_1,\nu_2,\tau} = \langle f_{i,k}^{(\nu_1)*}, f_{i,l,\tau}^{(\nu_2)*}\rangle$, and then obtaining the estimate

$$\hat{\rho}_{k,l,\nu_1,\nu_2,\tau} = \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}_{k,l,i,\nu_1,\nu_2,\tau}. \tag{11}$$

A quantity of interest is the lag τ at which the absolute value of the correlation $\rho_{k,l,\nu_1,\nu_2,\tau}$ is maximized, i.e., $\tau_{k,l,\nu_1,\nu_2} = \text{argmax}_\tau\, |\rho_{k,l,\nu_1,\nu_2,\tau}|$. A natural estimate for this is $\hat{\tau}_{k,l,\nu_1,\nu_2} = \text{argmax}_\tau\, |\hat{\rho}_{k,l,\nu_1,\nu_2,\tau}|$. We may estimate the variance of $\hat{\rho}_{k,l,\nu_1,\nu_2,\tau}$ in the same manner as for $\hat{\rho}_{k,l}$, described in Section 3.

4.2 Interval estimation after smoothing

If no smoothing of the longitudinal data is necessary (e.g., as with fully observed random trajectories), then the standard bootstrap percentile interval method, implemented via sampling

from replacement from the $\hat{\rho}_{i,k,l}$, provides interval estimates for the dynamical correlation measure. If an additional pre-smoothing step is necessary to handle data that are collected at discrete time points, then an appropriate interval estimate will reflect the additional uncertainty of the smoothing step. For this purpose we propose a two-stage bootstrap approach:

In a first smoothing step, one obtains continuous random trajectories. As will be shown in Section 5.1, the estimation of dynamical correlation is robust to the smoothing bandwidth selection. At the time points of the originally collected measurements, obtain residuals by subtracting the smoothed from the original responses. In a second step, obtain a random sample of individuals of size n, by sampling with replacement from all individuals. Then, for each individual from this sample, sample with replacement from the residuals obtained in the first smoothing step. If the same individual is sampled more than once, obtain a new random sample of residuals, for each appearance of that individual in the sample. Then "recreate the original data" by adding the resampled residuals to the original smoothed curves. We repeat this step a total of B times. We recommend $B = 500$ as a reasonable size for the number of bootstrap samples constructed in this way.

One then smooths the recreated data, using the same bandwidth choice as in the first smoothing step; this choice worked well in our applications. The resulting smooth curves reflect the additional uncertainty from the initial smoothing step, due to the additional resampling from residuals. Finally, one obtains an overall dynamic correlation estimate, following the specifications above, for each bootstrap sample. Ordering these across the B samples, one may obtain a $100(1-\alpha)\%$ two-stage bootstrap percentile interval. Alternatively, a studentized bootstrap interval can be generated.

4.3 Additional extensions

Other applications of dynamical correlation are *dynamical principal components* and *dynamical factor analysis*. Upon having obtained the *dynamical correlation matrix* $R = (\hat{\rho}_{k,l,\nu_1,\nu_2,\tau})_{1\le k,l\le p}$, by applying (11) to all pairs of components (k, l), we then apply the corresponding classical multivariate techniques (principal components analysis, factor analysis) to this correlation matrix.

In addition to the *dynamical correlation* discussed above, model (2) also contains the notion of a *static correlation*, which would be represented by the correlation of the $\varepsilon_{0,k}$. Here, $\gamma_{k,l} = \mathrm{corr}(\varepsilon_{0,k}, \varepsilon_{0,l}) = \frac{E[(\varepsilon_{0,k}-1)(\varepsilon_{0,l}-1)]}{[\mathrm{var}(\varepsilon_{0,k})\mathrm{var}(\varepsilon_{0,l})]^{1/2}}$ can be estimated by $\hat{\gamma}_{k,l} = \{\frac{1}{n}\sum_{i=1}^{n}(M_{i,k} - \overline{M}_k)(M_{i,l} - \overline{M}_l)\}/\{\widehat{\mathrm{var}}(M_{i,k})\widehat{\mathrm{var}}(M_{i,l})\}^{1/2}$, as according to (3), $E(M_{i,k}) = \mu_{0,k}$, and $\mathrm{corr}(M_{i,k}, M_{i,l}) = \gamma_{k,l}$.

5. SIMULATION STUDY

We performed two sets of simulations. In the first set, our goal was to determine the sensitivity of the correlation estimate when using different bandwidths for implementation of a pre-smoothing

step to longitudinal data. In the second set, our focus was on the comparison between a form of functional canonical correlation and dynamical correlation.

5.1 Smoothing Parameter Selection Sensitivity Analysis

For $n = 50$ subjects, we generated two sets of $N = 100$ equidistant points t_j on the grid $[0, 1]$, for each of a set of 100 simulations. The responses $Y_{ik}(t_j)$ were generated as follows:

1. $Y_{ik}(t_j) = f_{i,k}(t_j) + e_{ikj}$,
   where $f_{i,k}(t_j)$ is the k^{th} random function or curve for the i^{th} subject, and e_{ikj} is the within-subject error term, with $i = 1, 2, ..., 50$, $j = 1, 2, ..., 100$, and $k = 1, 2$.

2. Following (2), we define the random function $f_{i,k}(t_j)$ as:
   $f_{i,k}(t_j) = \mu_k(t_j) + (\mu_{0,k} + \varepsilon_{0,k}) + \sum_{r=1}^{\infty} \varepsilon_{r,k}\eta_r(t_j) = 1 + \sum_{r=0}^{2} \varepsilon_{r,k}\eta_r(t_j)$,
   where, without loss of generality for the purposes of the simulation, we set $\mu_k(t_j) \equiv 0$, $\mu_{0,k} = 1$, and assume that only the random components $\{\varepsilon_{0,k}, \varepsilon_{1,k}, \varepsilon_{2,k}\}$ are of significance.

3. The three orthonormal functions on $[0, 1]$ that we employ to represent the functions $f_{i,k}$ are $\eta_0(t) \equiv 1, \quad \eta_1(t) = 2\sqrt{3}(t - 1/2), \quad \eta_2(t) = 6\sqrt{5}(t - 1/2)^2 - (1/2)\sqrt{5}\}$.

4. The $\varepsilon_{r,k}$ were generated as follows:

$$\begin{pmatrix} \varepsilon_{0,1} \\ \varepsilon_{1,1} \\ \varepsilon_{2,1} \\ \varepsilon_{0,2} \\ \varepsilon_{1,2} \\ \varepsilon_{2,2} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{01}^2 & 0 & 0 & \sigma_{01,02} & 0 & 0 \\ 0 & \sigma_{11}^2 & 0 & 0 & \sigma_{11,12} & 0 \\ 0 & 0 & \sigma_{21}^2 & 0 & 0 & \sigma_{21,22} \\ \sigma_{01,02} & 0 & 0 & \sigma_{02}^2 & 0 & 0 \\ 0 & \sigma_{11,12} & 0 & 0 & \sigma_{12}^2 & 0 \\ 0 & 0 & \sigma_{21,22} & 0 & 0 & \sigma_{22}^2 \end{bmatrix} \right),$$

   where $\varepsilon_{0,k}, \varepsilon_{1,k}, \varepsilon_{2,k}$ are the random terms for the k-th component function. The $\varepsilon_{r,k}$ are uncorrelated for different r. We choose $\sigma_{01}^2 = 1$, $\sigma_{11}^2 = 1/2$, $\sigma_{21}^2 = 1/3$, $\sigma_{02}^2 = 1/2$, $\sigma_{12}^2 = 1/3$, $\sigma_{22}^2 = 1/4$. The covariance terms were chosen as $\sigma_{01,02} = 1/3$, $\sigma_{11,12} = 1/4$, $\sigma_{21,22} = 1/6$.

5. The within-subject errors were assumed to be distributed as $e_{ikj} \sim N(0, \sigma_0^2)$, where we set $\sigma_0^2 = 1/4$, and assume that the e_{ikj} are all uncorrelated.

After generating the $Y_{ik}(t_j)$ once, we used local linear smoothing (see Appendix A.1 for more details) to obtain smooth trajectories. In order to determine the sensitivity of the resulting dynamical correlation estimate to bandwidth choice, we generated the smoothed curves over a grid of seventeen equidistant bandwidths h: {.010, .035, .060, ..., .410}. We then obtained dynamical correlation estimates for each bandwidth choice with a uniform weight function, using (9). The

results averaged over 100 simulation runs are in Table 1, noting that the target correlation ρ_{target} was chosen as 0.5.

Table 1. Results of Sensitivity Analysis Simulation
Using Different Bandwidths

h	.010	.035	.060	.085	.110	.135	.160	.185	.210	.235	.260	.285	.310	.335	.360	.385	.410
corr	.392	.461	.478	.485	.489	.491	.492	.493	.493	.493	.493	.492	.491	.490	.488	.487	.485
sd	.067	.081	.085	.087	.089	.090	.090	.091	.091	.091	.091	.091	.091	.091	.091	.092	.092

Here, *corr* represents the Monte Carlo mean of $\hat{\rho}$, while *sd* represents the Monte Carlo standard deviation of $\hat{\rho}$, over all 100 simulations.

The correlation estimates are seen to be robust over a wide range of reasonable bandwidths. The small downward bias of the estimates is likely due to the discrete nature of the measurements. Fortunately, the bias of the estimate is very small, only .007 for bandwidths .185 through .260 and not more than .012 for bandwidths .110 through .360. The only poor results are for the very small bandwidths, particularly for the lowest bandwidth of .010.

We also ran a similar simulation, except using a smaller number of repeated measurements per individual, i.e., 15, instead of 100. This may better reflect the number of repeated measurements people see in a longitudinal data set, and indeed represents the average number of repeated measures per patient in the protein application discussed in Section 6. Also, like this application, we limited the number of individuals in the simulation to 35. The results of this smaller sample simulation were almost identical to the larger one, except that the downward bias is more pronounced, though still small, at .024, versus .007 for the 50 individual, 100 repeated measures case. The bandwidth robustness remains, with any bandwidth choice away from the extremes showing very little increase in bias.

5.2 Functional Canonical Correlation vs. Dynamical Correlation

The goal of this simulation is to evaluate dynamical correlation versus functional canonical correlation (Leurgans *et al.,* 1993) on the basis of each method's ability to handle an increasingly large number of repeated measurements per subject or observation unit. Before describing the simulation, we should point out that the interpretations of dynamical correlation and functional canonical correlation are not the same. Using the example of two curves, dynamical correlation is attempting to describe the expected cosine of the angle in the function vector space between the random components of the two curves, whereas functional canonical correlation is attempting to describe the maximum correlation between linear projections of these curves, that are based on canonical weight functions. Both correlation methods can be used to describe multivariate relationships between curves and, hence, a comparison of their performance in practice is useful.

Formally, the (first) coefficient of functional canonical correlation $\rho^C_{k,l}$ between random functions f_k, f_l can be defined as follows, referring to He *et al.,* (2003) for further details and theoretical background. Let

$$\rho^C_{k,l} = \sup_{g_k, g_l \in L^2(dw)} \text{cov}(\langle g_k, f_k \rangle, \langle g_l, f_l \rangle) \tag{12}$$

where the *canonical weight functions* g_k and g_l are subject to

$$\text{var}(\langle g_k, f_k \rangle) = 1, \text{and } \text{var}(\langle g_l, f_l \rangle) = 1.$$

As estimating functional canonical correlation corresponds to an ill-posed inverse problem, one needs to implement some kind of regularization. As it turns out, this need for regularization is the crux of this method.

We use here a simple version of regularized functional canonical correlation for our comparisons. The setup of this simulation is the same as the initial larger sample simulation in Section 5.1. Given the previous results, for the smoothing steps we use a bandwidth of 0.21, which lies in the middle of the range of the originally considered bandwidth choices. We ran 250 simulations, generating two sets of repeated measurements for each of 50 patients, with target dynamical correlation $\rho_{target} = 0.50$. The dynamical correlation method is implemented with a smoothing step as described above. The mean dynamical correlations when assuming 100, 1000 or 5000 repeated measurements were as follows: $\hat{\rho}_{100} = 0.494$, $\hat{\rho}_{1000} = 0.500$, $\hat{\rho}_{5000} = 0.500$. Obviously, dynamical correlation is quite insensitive to the original number of repeated measurements, even when the number becomes large.

As the definition of functional canonical correlation clearly differs from that for dynamical correlation, we cannot expect the first canonical correlation to be 0.5. What is of main interest is to see how many repeated measurements the method can handle before breaking down. We evaluated canonical correlation in two ways: (1) Applying multivariate canonical correlation directly to the vector of repeated measurements; (2) Implementing a regularized version by incorporating an initial smoothing step as described above. This latter method is discussed and evaluated under the rubric "smoothing and subsequent subsampling" (Method 1, FCA-LP) in He *et al.,* (2004), where various implementations of functional canonical correlation are compared. This method is described there as a simple and successful regularization procedure.

With both implementations (1) and (2), functional canonical correlation was found to break down for fairly small numbers of repeated measurements. For even just 30 repeated measurements, the unregularized canonical correlation results were 1.0 for the first 11 (of 30) canonical correlations. For 49 repeated measurements, each and every canonical correlation was 1.0. It was even more surprising that the same was found for the functional canonical correlations that had been obtained through regularization by smoothing.

Canonical correlation may behave well for up to several repeated measurements, possibly up to twenty, which is the number of repeated measurements that were used in the Leurgans *et al.,* (1993) paper. However, we replicated the smaller sample simulation as described at the end of

section 5.1, and we found the first 4 (of 15) canonical correlations all were at least 0.9. This is difficult to interpret, and shows the signs of breakdown even with as few as 15 repeated measures. Our results strongly suggest that canonical correlation cannot be depended upon when the number of repeated measurements is in the medium to large range. This problem is clearly avoided by the dynamical correlation measure.

6. APPLICATION TO LONGITUDINAL PROTEIN DATA

The longitudinal acute phase protein measurements mentioned in section 1 consist of measurements for 35 subjects with varying designs, each with between 12 and 18 repeated measurements for the multivariate set of five proteins. Two of the proteins, albumin (alb) and transferrin (trf), fall in the class of negative acute phase proteins (NAPP's). The other three proteins, C-reactive protein (crp), α-aminoglobulin (aag), and ceruloplasmin (cer), belong to the class of positive acute phase proteins, or simply, acute phase proteins (APP's). Increased values of these latter proteins are indicators of infections and the APPs are generally thought to move in the opposite direction of the NAPP's. We note that there exists more between-subject variability than within-subject variability among the responses for these patients.

A central biomedical question of this study (Kaysen *et al.,* 2000) was to determine in what way the acute phase blood proteins are correlated longitudinally, specifically to establish whether APP's are negatively correlated over time with NAPP's, and to analyze the nature of the correlation. We show how these questions can be addressed with the proposed dynamical correlation, including its extension to derivatives and lag effects.

The pre-smoothing step to obtain continuous protein level trajectories was implemented as follows: Assume f_k is defined on $[0, T_i]$ and that for subject i, measurements $x_{i,k}(t_{i,j})$ are recorded at times $t_{i,j}$, with $j = 1, ..., n_i$, $1 \le k \le p$, and $0 \le t_{i,j} \le T_i$. We set $f_k(t) = \hat{m}(t)$, where $\hat{m}(t)$ is the value of a specified non-parametric smoother at argument t. We apply the smoother to the scatterplot $(t_{i,j}, x_{i,k}(t_{i,j}))$, with the smoothed values computed on a dense grid of points. To perform the smoothing, we chose local linear regression (e.g., Fan and Gijbels, 1996) with kernel function $(1 - x^2)^2 1_{[-1,1]}$. For the basic approach, also see Appendix A.1. For this specific data analysis, we used a monotonically increasing design-adaptive bandwidth,

$$h_t = \begin{cases} a & : \quad t < u \\ a + \{(t-u)/(v-u)\}(c-a) & : \quad u \le t \le v \\ c & : \quad t > v \end{cases},$$

where, for the standardized protein curves, $u = 25$ days and $v = 165$ days; u and v were chosen taking available data and the resulting sufficient smoothness of the curves across the span of the follow-up period for the subjects into account. This adaptive bandwidth choice reflects the nonequidistant designs in which, roughly, weekly repeated measurements in the first six weeks of

Table 2. Estimated Dynamical Correlation Matrix
for Longitudinal Protein Data

	alb	crp	aag	cer	trf
alb	1.000	-0.298* (.004)	-0.326* (.036)	-0.166 (.276)	0.247 (.060)
crp		1.000	0.549* (0)	0.387* (.024)	-0.215 (.072)
aag			1.000	0.686* (0)	-0.096 (.616)
cer				1.000	0.107 (.256)
trf					1.000

* = 95% two-stage bootstrap CI does not contain 0.

Bootstrap p-values in parentheses.

observation were subsequently followed by monthly measurements. In general, we should note the lack of sensitivity in the resulting correlation estimate based on bandwidth choice, except toward the extremes, as demonstrated in the simulation study in Section 5. Though not shown in the results here, such lack of sensitivity of bandwidth selection was also demonstrated for the data in this example.

The standardized versions of the protein curves are shown in Figure 2 for the same randomly selected subject as in Figure 1. There is some appearance of a positive relationship over time between alb and trf, as well as between crp and aag, with a negative correlation between these two sets. It appears that cer is not consistently correlated with any other protein over time. One would like to ascertain whether the longitudinal protein relationships seen here for one individual are consistent across subjects.

Using the methods of Section 3, we calculated the dynamical correlation matrix shown in Table 2. We used weighted averages such that individual dynamical correlations between proteins were assigned weights of $N_{i,k}/\sum_{i=1}^{n} N_{i,k}$ for the overall dynamical correlation values, where $N_{i,k}$ is the number of original observed repeated measurements of protein k for patient i.

The estimated dynamical correlations support the hypothesis that the negative acute phase proteins (NAPP's), alb and trf, are negatively correlated over time with the positive acute phase proteins (APP's), crp, aag, and cer. Significance was assessed based on 95% two-stage bootstrap percentile confidence intervals, as described in Section 4.2. Associated bootstrap p-values, based on the empirical bootstrap distribution of the correlation estimates, are also provided in Table 2. The largest observed dynamical correlation occurs between aag and cer, two slow-moving positive acute phase proteins, at 0.686.

Using the dynamical correlation matrix of Table 2, we performed a corresponding dynamical principal components analysis to see which linear combinations of the proteins would best explain the longitudinal variability in the data. The first dynamical principal component accounted for 80% of this variability, with loadings given in Table 3. The principal component accounting for most of

Table 3. Loadings of First Dynamical Principal Component;
Based on Estimated Dynamical Correlation Matrix

	alb	crp	aag	cer	trf
1st PC	0.501	-0.481	-0.516	-0.375	0.333

Table 4. Estimated Static Correlation Matrix
for Longitudinal Protein Data

	alb	crp	aag	cer	trf
alb	1.000	-0.214	-0.173	-0.259*	-0.337*
crp		1.000	0.597*	0.448*	-0.164
aag			1.000	0.433*	0.002
cer				1.000	0.121
trf					1.000

* = 95% bootstrap CI does not contain 0

the longitudinal variability of the proteins thus seems to be simply a linear contrast between the NAPP's and APP's, verifying the a priori hypothesis of the medical investigator.

In order to identify "typical" subjects exemplifying the correlation structure, we search for protein curves which are most aligned with the direction of the vector of loadings for the first dynamical principal component. This is possible since the proposed methodology allows for the construction of dynamical correlation matrices and corresponding principal components (PC's) for individuals. We simply locate the individual(s) whose loadings enclose the smallest angle with the loadings of Table 3. It may also be of interest when analyzing multivariate functional data to identify outlying subjects, i.e., individual(s) with the largest such angle(s). The standardized curves for the two subjects whose loadings gave rise to the smallest angles are presented in Figure 3. The strong negative dynamical correlation between the APP's and NAPP's is nicely demonstrated in these two subjects.

We also compared the dynamical correlation results to the "static correlation" discussed in section 4.3. This is equivalent to a cross-sectional analysis and is based on the correlations between the intercept terms in (2), i.e., M_k and M_l, for two proteins of interest, k and l, $1 \leq k, l \leq p$. The results from this approach, listed in Table 4, turned out to be similar to the longitudinal dynamical correlation approach, except for correlations related to trf. For example, for biological reasons, trf and alb, both negative acute phase proteins, would be expected to be positively correlated, a result seen in the dynamical correlation approach. However, trf and alb were seen to be significantly negatively correlated with this static cross-sectional approach. This indicates a preference for dynamical correlations.

Another analysis focused on the dynamical relationship between curves, derivatives, and time-

Table 5. Maximum Dynamical Correlation Results for alb and crp for Derivatives of Orders 0 (alb0, crp0) and 1 (alb1, crp1) and Lags (in days; alb leading crp), Ranging between -30 and 30 Days

		lag (days)	corr at lag	corr for no lag
alb0	crp0	-12	-0.348	-0.298
alb1	crp1	-10	-0.368	-0.300

shifted versions for alb and crp. These results are presented in Table 5, including the amount of lag for which the highest correlation was achieved, the value of the corresponding dynamical correlation, and the correlation at lag 0, for both functions and first derivatives. The results for the first derivative closely mimic those for the functions themselves. Inclusion of a lag increases correlation somewhat but not by much as compared to the correlation for no lag. The correlation as a function of lag is shown in Figure 4, for alb and crp. It appears that changes in alb are anticipated by changes in crp, a relationship that is biologically plausible, but the converse is not true. Though not shown here, a similar relationship exists for lags of the first derivatives between alb and crp.

A final analysis concerned the performance of functional canonical correlation in this real data setting. We use the same regularization procedure as described in Section 5.2 and the same estimated curves as created for the dynamical correlation analysis. We looked at 15 equidistant points from each estimated curve, which reflected the average number of repeated measurements per patient, and found that 5 (of 15) canonical correlations were 0.94 and above. This result renders this approach rather uninterpretable, and in concordance with the findings from the simulations in Section 5.2, indicates problems with using functional canonical correlation when the number of repeated measurements is in the medium to large range.

7. DISCUSSION

We have introduced a model for multivariate longitudinal data that allows for the definition of a simple measure of correlation between various longitudinal components. The corresponding estimates of dynamical correlation have reasonable asymptotic properties and allow for establishing dynamical correlation at the individual level. Usual multivariate techniques such as principal components analysis can be applied once a dynamical correlation matrix has been obtained. In addition, we can also determine whether the dynamical correlation is maximized at certain lags between the component curves. Such information will lead to a better understanding of the relationships between the components of multivariate longitudinal data.

There exist many possible extensions of this research. For example, the proposed dynamic correlation measures similarity between functions, quantifying how they move together through

time. However, one may have responses that follow differing patterns of correlation depending on which interval of follow-up one is investigating. An example might be the consideration of the correlation between two functions before and after introduction of some stimuli or treatment. This can be handled with a more flexible weight function. For example, one could include a local window type weight function $w(t, z, h) = (1/2h)\ I_{[z-h,z+h]}(t)$, where h is a specified bandwidth centered at the moving target z.

Another extension would be to investigate our correlation measure as a function of a covariate or covariate function. Morris *et al.,* (2001) have explored this, developing a functional data analytic correlation measure to determine the dependency between carcinogen-induced DNA adduct levels in two parts of the colon as a function of the relative cell position within the crypt. Our method could as well be extended to reflect the effect of a discrete or continuous covariate on the dynamical correlation. For example, we could consider determining the correlation of the blood proteins as a function of gender; we would then generate the curves and subsequent correlations in the two groups separately.

Although we establish an asymptotic normality result for the estimated dynamical correlation, in the proof, we do require the number of repeated measurements to increase to ∞. Though a very large number of repeated measurements is typically not observed in longitudinal data sets, the performance of dynamic correlation has been shown to be rather good in cases of a smaller number of repeated measurements, both in an application with an average of 15 repeated measures per individual, and in simulations. Because curves need to be estimated before implementing this method, we do recommend using caution when the number of repeated measurement is small, say below 10, and generally recommend against its use when the number of repeated measurements is as small as 3 or 4.

Finally, though we have demonstrated that the correlation estimate is not sensitive to bandwidth choice, an automatic bandwidth selection procedure is desirable. One possibility is a leave-one-individual-out cross-validation approach that is based upon the integral defined by (10). This is an area of future research. Our simulations and application to longitudinal protein data clearly demonstrate the usefulness and potential of the dynamic correlation method. Among its main attractions are easy interpretability, simplicity, stability, robustness, and reasonably fast computation.

APPENDIX

A.1. Local linear smoothing.

For a scatterplot (X_j, Y_j), $j = 1, \ldots, M$, and a point x in the domain of the X_i, we define local linear smoothers targeting the regression function $E(Y|X = x)$ as follows: Given a nonnegative

kernel function K and a sequence of bandwidths h, minimize the weighted sum of squares

$$\sum_{j=1}^{M} K\left(\frac{x - X_j}{h}\right)[Y_j - \{\beta_0 + \beta_1(x - X_j)\}]^2$$

with respect to β_0, β_1, obtaining $\hat{\beta}_0, \hat{\beta}_1$. The estimated regression function at x is $\widehat{E}(Y|X = x) = \hat{\beta}_0$. In our application we set $X_j = t_j, Y_j = Y_{ik}(t_j)$ and $M = N$.

A.2. Proof of Theorem 1.

Note that according to (a) orthonormality of the η_r $(r = 0, 1, ...)$, (b) $\int w(t)dt = 1$, and (c) $\int \mu_k(t)w(t)dt = 0$, we have $\int (f_k(t) - \mu_k(t) - M_k)(f_l(t) - M_l - \mu_l(t))w(t)dt =$
$\int (\sum_{r=1}^{\infty} \varepsilon_{r,k}\eta_r(t))(\sum_{r=1}^{\infty} \varepsilon_{r,l}\eta_r(t))w(t)dt = \int \sum_{r=1}^{\infty} \varepsilon_{r,k}\varepsilon_{r,l}w(t)dt = \sum_{r=1}^{\infty} \varepsilon_{r,k}\varepsilon_{r,l}$.

Analogously, $\int (f_k(s) - \mu_k(t) - M_k)^2 w(s)ds = \sum_{r=1}^{\infty} \varepsilon_{r,k}^2$, and (6) follows. □

A.3. Regularity conditions for Theorem 2.

The smooth curve estimates $\hat{f}_{k,i}^S$ are obtained in a pre-smoothing step from the observed discrete data. We make here a few simplifying assumptions. First of all, we assume that at least N measurements are available for each curve which is to be smoothed, secondly we assume that data for all curves are sampled according to the model $Y_{ikj} = f_{i,k}(x_{ikj}) + e_{ikj}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, p, \quad j = 1, \ldots, N$, where the errors e_{ikj} are i.i.d. and satisfy $\mathrm{var}(e_{ikj}) = \sigma_e^2$, $E(e_{ikj}) = 0$.

Additional assumptions are as follows: The sampling points x_{ikj} follow a design density g_{ikj} (see (2.4) in Müller and Stadtmüller, 1987, from here on referred to as MS). This family of design densities is equi-continuous and uniformly bounded away from 0 for all i, k, j. All design densities have the same compact support. Furthermore, the errors satisfy Assumption B of MS, with uniform bound on $E(|e_{ikj}|^s)$ in i, k, j for a given $s > 2$. The smoother that is used may be written as $\hat{f}_{i,k}^S(t) = \sum W_j(t)Y_{ikj}$, i.e., it is a linear smoother. The weight functions $W_j(t)$ employed by this smoother satisfy conditions (5.3)-(5.6) of MS, uniformly in i, k, j.

We note that these assumptions are satisfied for kernel smoothers with smooth (Lipschitz-continuous) non-negative kernels or local linear smoothers with smooth weight functions and suitable choices of bandwidths h. For twice differentiable functions, setting $k = 2$ and $\zeta = 0$ in Theorem 5.1 in MS, these conditions amount to $\liminf Nh^2 > 0$, $\liminf (\frac{Nh}{\log N})^{1/2} N^{-2/(s-\eta)} > 0$ for an η satisfying $0 < \eta < 2$.

We also assume that the smoother, when using bandwidth h, satisfies

$$E(\hat{f}^S(t) - f(t)) = h^2 f^{(2)}(t)C(1 + o(1)),$$

for a constant $C > 0$, where the o-term is uniform over equi-continuous families of functions $f^{(2)}$. Additional conditions specific to the situation at hand are

(C1) All functions $f_{i,k}$ are twice continuously differentiable, with equi-continuous second derivatives, and

$$\sup_{i,k,t} |f_{i,k}^{(2)}(t)| = O_p(1).$$

(C2) The minimum number of sampled observations per curve N increases with sample size n in such a way that

$$N \to \infty, \quad n(\log N/N)^{4/5} \to 0.$$

(C3) It holds that

$$n(\frac{\log N}{Nh})^{1/2} \to 0, \quad \sqrt{n}h^2 \to 0.$$

A.4. Proof of Theorem 2.

Following step by step the proof of Lemma 5.2 in MS, one finds that (C2) is sufficient to apply the exponential inequality to invoke the Borel-Cantelli lemma, and (C1) implies uniform bounds for the bias terms. Setting $\xi_N = (\frac{\log N}{Nh})^{1/2} + h^2$, we obtain from this lemma that

$$\sup_{i,k,t} |\hat{f}_{i,k}^S(t) - f_{i,k}(t)| = O_p(\xi_N),$$

and therefore also

$$\sup_t |\frac{1}{n}\sum_{i=1}^n \hat{f}_{i,k}^S(t) - \frac{1}{n}\sum_{i=1}^n f_{i,k}(t)| = O_p(\xi_N).$$

This in turn implies

$$\sup_{k,l,i} |\hat{\rho}_{k,l,i}^S - \hat{\rho}_{k,l,i}| = O_p(\xi_n),$$

and therefore

$$|\hat{\rho}_{k,l}^S - \hat{\rho}_{k,l}| = O_p(\xi_n).$$

Then (C3) implies that

$$|\hat{\rho}_{k,l}^S - \hat{\rho}_{k,l}| = o_p(\frac{1}{\sqrt{n}}).$$

This means that the dynamical correlation estimator based on pre-smoothed functions and the estimator obtained from full processes have the same asymptotic distribution.

Next, we show $\sqrt{n}(\hat{\rho}_{k,l} - \rho_{k,l}) \longrightarrow N(0, \delta_{k,l}^2)$, in distribution, as $n \to \infty$, for $0 < |\rho_{k,l}| < 1$, and that $var(\hat{\rho}_{k,l}) = \delta_{k,l}^2$ is finite and bounded between 0 and 1.

Let $\tilde{f}_{i,k}^*$ and $\tilde{f}_{i,l}^*$ be the standardized curves for responses k and l, respectively, for patient i when μ_k and μ_l are known or unknown but constant. Then, $\tilde{\rho}_{k,l} = \frac{1}{n}\sum_{i=1}^n \int \tilde{f}_{i,k}^*(t)\tilde{f}_{i,l}^*(t)w(t)dt$ is the average of n iid random variables, so that $E\tilde{\rho}_{k,l} = \rho_{k,l}$. Under the assumption that $\delta_{k,l}$ is finite, the central limit theorem provides the desired result.

Since $\tilde{\rho}_{k,l,i} = \int \tilde{f}_{i,k}^*(t)\tilde{f}_{i,l}^*(t)w(t)dt = \sum_{r=1}^\infty \varepsilon_{i,r,k}\varepsilon_{i,r,l}/[\sum_{r=1}^\infty \varepsilon_{i,r,k}^2 \sum_{r=1}^\infty \varepsilon_{i,r,l}^2]^{1/2}$, based on (6), this implies

$$\delta_{k,l}^2 = E\tilde{\rho}_{k,l}^2 - (E\tilde{\rho}_{k,l})^2 = E\{(\frac{\sum_{r=1}^\infty \varepsilon_{r,k}\varepsilon_{r,l}}{[\sum_{r=1}^\infty \varepsilon_{r,k}^2 \sum_{r=1}^\infty \varepsilon_{r,l}^2]^{1/2}})^2\} - (E\{\frac{\sum_{r=1}^\infty \varepsilon_{r,k}\varepsilon_{r,l}}{[\sum_{r=1}^\infty \varepsilon_{r,k}^2 \sum_{r=1}^\infty \varepsilon_{r,l}^2]^{1/2}}\})^2. \tag{A.1}$$

The second term on the r.h.s. of (A.1) is simply $\rho_{k,l}^2$ and hence is bounded between 0 and 1, while the first term, by Cauchy-Schwarz, is also bounded between 0 and 1. As long as $\rho_{k,l}$ is not equal to 0, -1, or 1, then $\delta_{k,l}^2$ is bounded, positive, and not equal to 0. A degenerate result occurs when $\rho_{k,l}$ is equal to 0, -1, or 1, since then $\delta_{k,l}^2 \equiv 0$. Therefore, we require $0 < |\rho_{k,l}| < 1$ for the asymptotic normality result. In the case when μ_k and μ_l are neither constant nor known, we obtain the weaker result that $\hat{\rho}_{k,l} = \rho_{k,l} + O_p(n^{-1/2})$. □

REFERENCES

Ash, R.B. and Gardner, M.F. (1975), *Topics in Stochastic Processes.* New York: Academic Press.

Boudjellaba, H., Dufour, J.M., and Roy, R. (1992), "Testing Causality Between Two Vectors in Multivariate Autoregressive Moving Average Models," *Journal of the American Statistical Association,* 87, 1082-1090.

Brillinger, D. (1975), *Time Series: Data Analysis and Theory.* New York: Holt Rinehart, and Winston.

Conway, J. B. (1985), *A Course in Functional Analysis.*New York: Springer-Verlag.

Church, A. (1966), "Analysis of Data When the Response Is a Curve," *Technometrics,* 8, 229-246.

Diggle, P.J., Heagerty, P., Liang, K.Y., and Zeger, S.L. (2002), *Analysis of Longitudinal Data,* 2^{nd} ed. New York: Oxford University Press.

Efron, B. (1998), "R.A. Fisher in the 21st century," *Statistical Science,* 13, 95-114.

Efron, B. and Tibshirani, R.J. (1993), *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications.* London: Chapman and Hall.

Fisher, R.A. (1921), "On the probable error of a coefficient of correlation deduced from a small sample," *Metron,* 1, 3-32.

Hand, D.J. and Crowder, M.J. (1996), *Practical Longitudinal Data Analysis.* London: Chapman and Hall.

He, G., Müller, H.G., Wang, J.L. (2003), "Functional Canonical Analysis for Square Integrable Stochastic Processes," *Journal of Multivariate Analysis*, 85, 54-77.

He, G., Müller, H.G., Wang, J.L. (2004), "Methods of Canonical Analysis for Functional Data," *Journal of Statistical Planning and Inference* **122**, 141-159.

Heckman, N. and Zamar, R. (2000), "Comparing the Shapes of Regression Functions", *Biometrika* **87**, 135-144.

Hotelling, H. (1936), "Relations between Two Sets of Variables," *Biometrika,* 28, 321-377.

Jones, R.M. (1993), *Longitudinal Data with Serial Correlation: A State-Space Approach.* London: Chapman and Hall.

Kaysen, G.A., Dubin, J.A., Müller, H.G., Rosales, L.M., Levin, N.W., and the HEMO Study Group (2000), "The Acute Phase Response Varies with Time and Predicts Serum Albumin Levels in a Longitudinal Study of Hemodialysis Patients," *Kidney International,* 58, 346-352.

Leurgans S.E., Moyeed R.A., and Silverman B.W. (1993), "Canonical Correlation Analysis When the Data Are Curves," *Journal of the Royal Statistical Society,* Ser. B, 55, 725-740.

Martinussen, T., and Scheike, T.H. (1999), "A Semiparametric Additive Regression Model for Longitudinal Data," *Biometrika,* 86, 691-702.

Martinussen, T., and Scheike, T.H. (2000), "A Nonparametric Dynamic Additive Regression Model for Longitudinal Data," *The Annals of Statistics,* 28, 1000-1025.

Molenaar, P.C.M. (1985), "A Dynamic Factor Model for the Analysis of Multivariate Time Series," *Psychometrika,* 50, 181-202.

Morris, J.S., Wang, N., Lupton, J.R., Chapkin, R.S., Turner, N.D., Hong, M.Y., and Carroll, R.J. (2001), "Parametric and Nonparametric Methods for Understanding the Relationship between Carcinogen-Induced DNA Adduct Levels in Distal and Proximal Regions of the Colon," *Journal of the American Statistical Association*, 96, 816-826.

Müller, H.G. and Stadtmüller, U. (1987), "Estimation of Heteroscedasticity in Regression-Analysis," *Annals of Statistics,* 15, 610-625.

Müller, H.G., Stadtmüller, U., and Schmitt, T. (1987), "Bandwidth Choice and Confidence Intervals for Derivatives of Noisy Data," *Biometrika,* 74, 743-750.

Olshen, R.A., Biden, E.N., Wyatt, M.P., and Sutherland, M.P. (1989), "Gait Analysis and the Bootstrap," *Annals of Statistics,* 17, 1419-1440.

Ramsay, J.O. and Silverman, B.W. (1997), *Functional Data Analysis.* New York: Springer.

Rice, J.A., and Silverman, B.W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society*, Ser. B, 53, 233-243.

Service, S.K., Rice, J.A., and Chavez, F.P. (1998), "Relationship between Physical and Biological Variables During the Upwelling Period in Monterey Bay, CA," *Deep-Sea Research,* II, 1669-1685.

Sy, J.P., Taylor, J.M.G., and Cumberland, W.G. (1997), "A Stochastic Model for the Analysis of Bivariate Longitudinal AIDS Data," *Biometrics,* 53, 542-555.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Wang, Y., Guo, W., and Brown, M.B. (2000), "Spline Smoothing for Bivariate Data with Applications to Association between Hormones," *Statistica Sinica,* 10, 377-397.

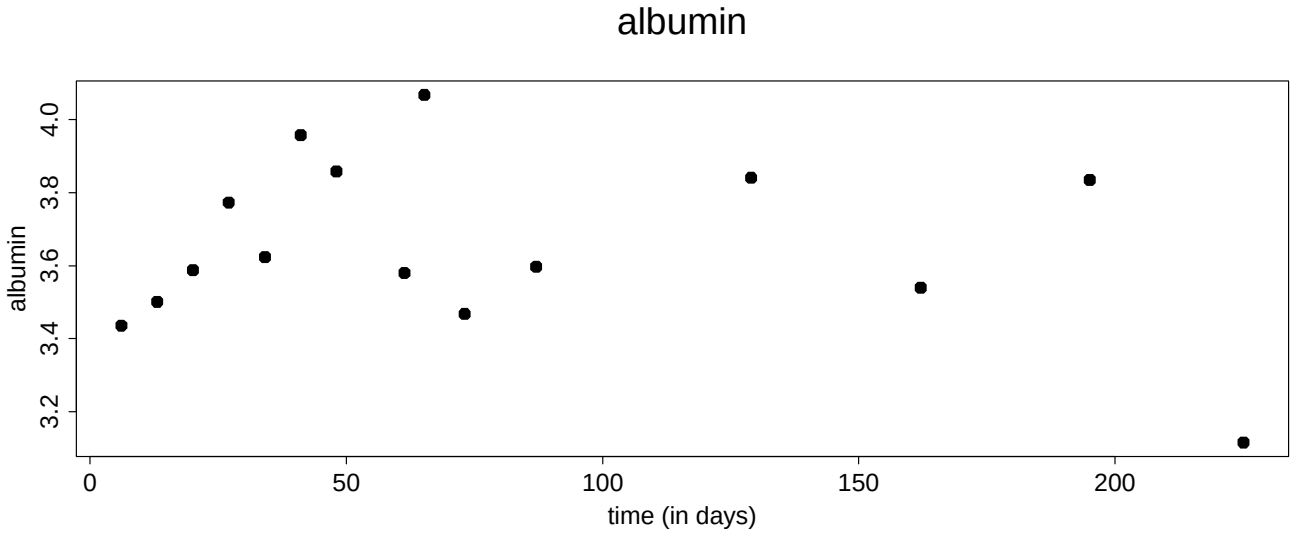
FIGURE CAPTIONS

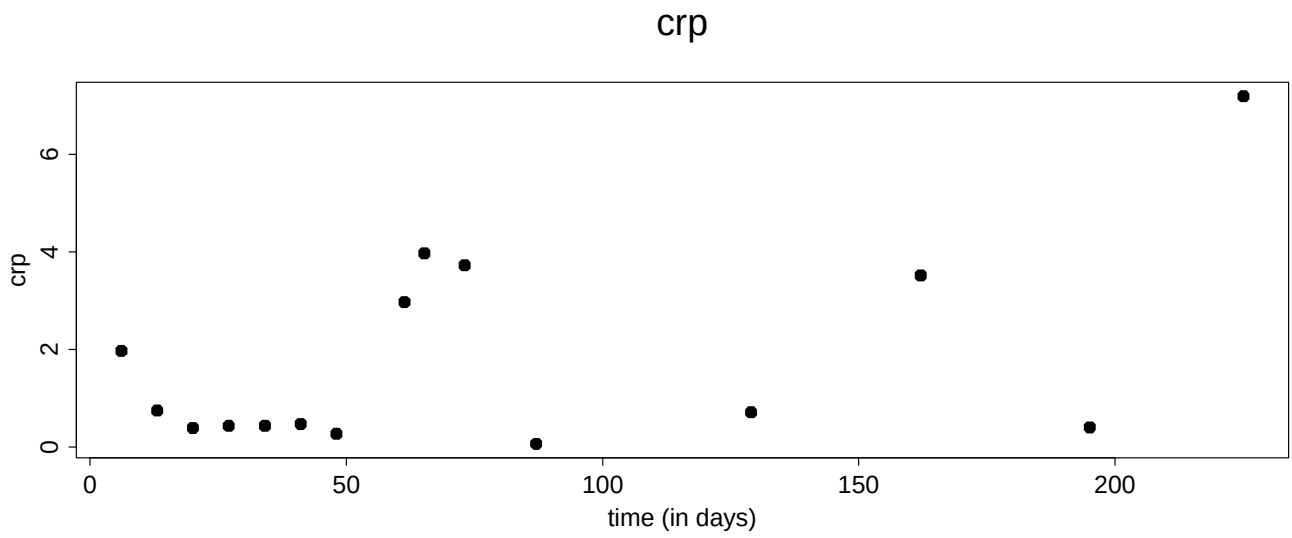
Figure 1. Observed repeated measurements for albumin (alb) and C-reactive protein (crp) for a randomly selected subject. There is evidence for a negative correlation over time between alb and crp.

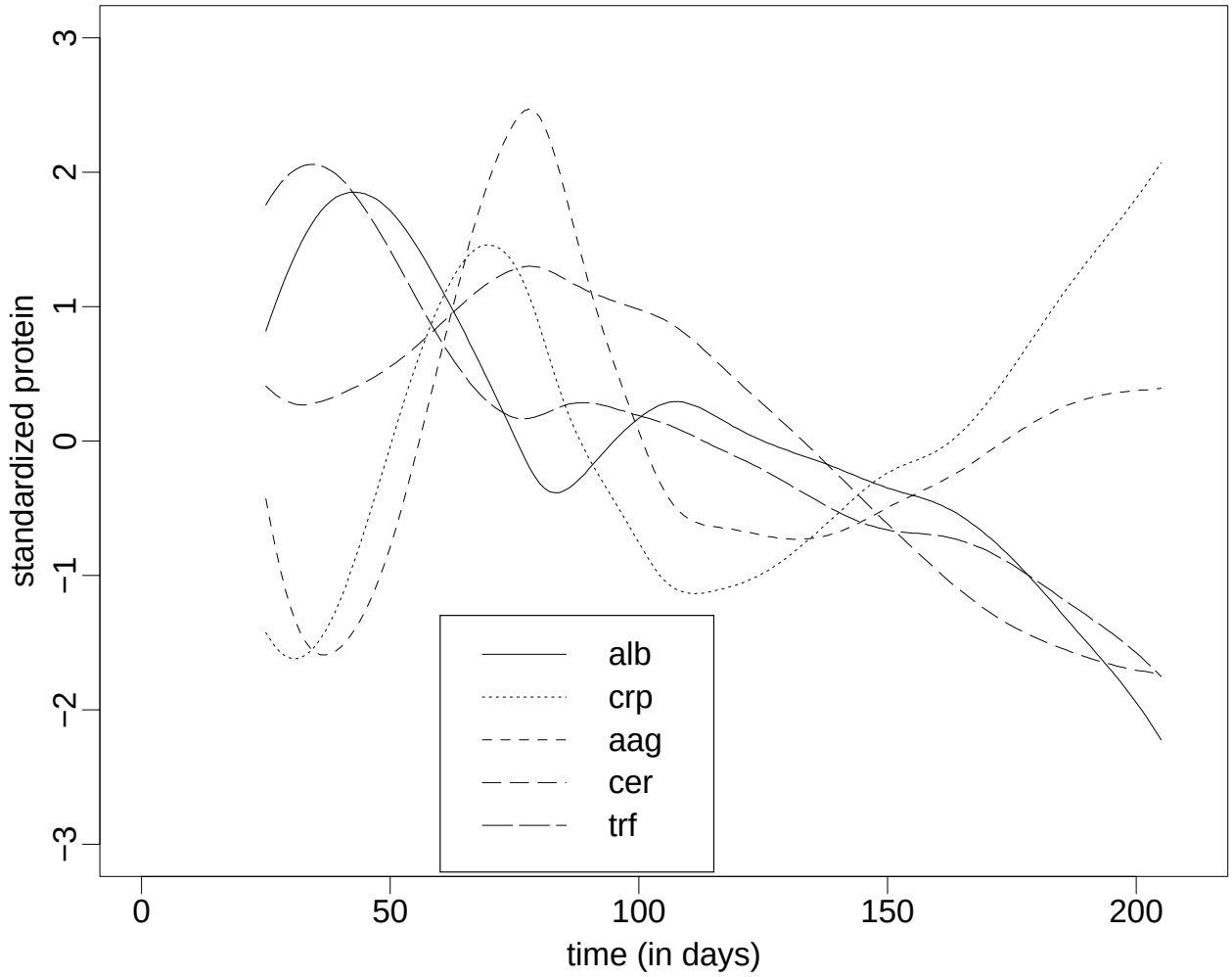
Figure 2. Smoothed standardized protein curves for the five proteins (albumin (alb); C-reactive protein (crp); α-aminoglobulin (aag); ceruloplasmin (crp); transferrin (trf)) for the same subject shown in Figure 1. The smoothed curves were created using local linear regression as described in the text.

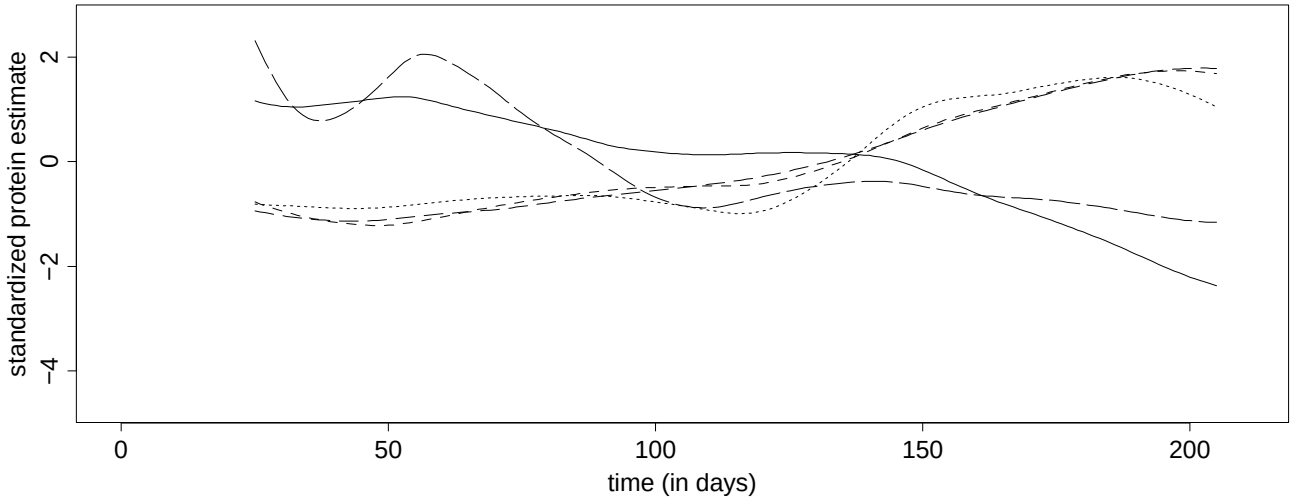
Figure 3. Smoothed standardized proteins for two subjects whose correlation structure is most closely aligned with the first principal component of the estimated dynamical correlation matrix.

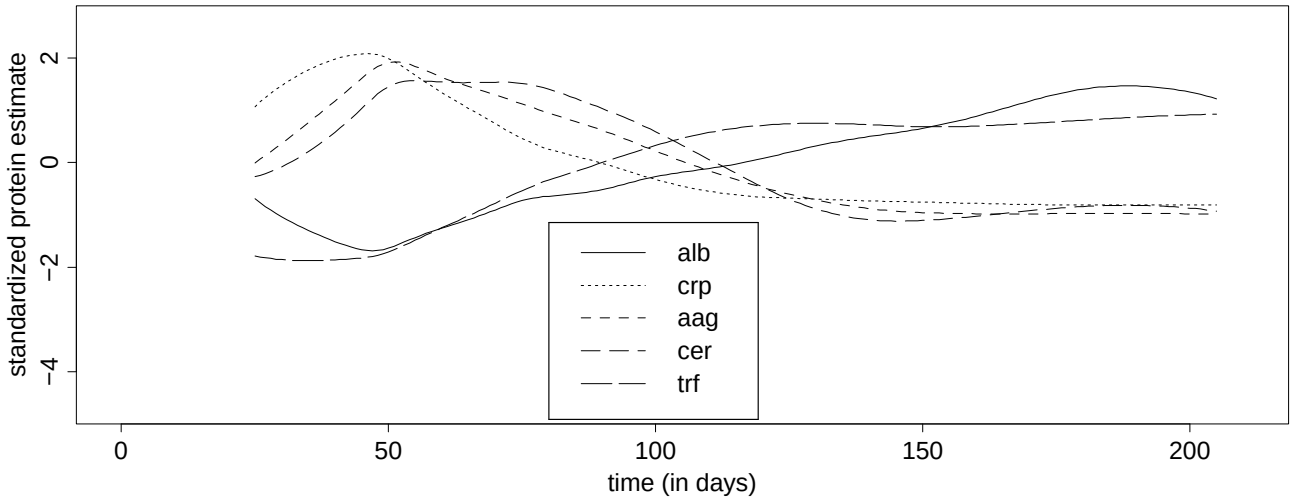
Figure 4. Dynamical correlation between alb and crp as a function of lag term. Dynamical correlations were calculated over a grid of lags, in days. When the lag term is negative (e.g., -x days), this implies crp is leading alb by x days, while alb is leading crp by x days when the lag term is positive.

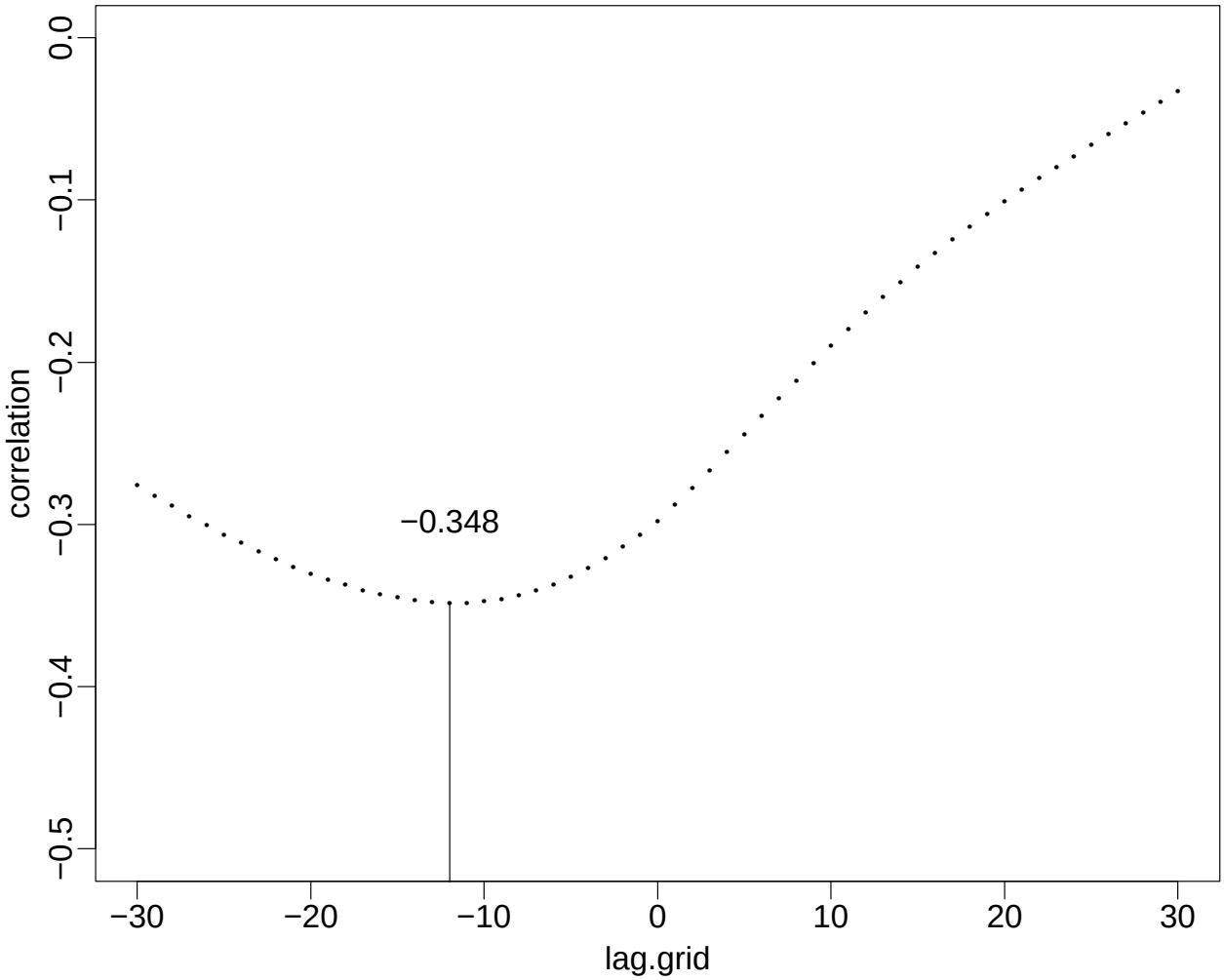

albumin
albumin
3.2
3.4
3.6
3.8
4.0
0
50
100
150
200
time (in days)


crp
crp
0
2
4
6
0
50
100
150
200
time (in days)


standardized protein
3
2
1
0
−1
−2
−3
0
50
100
150
200
time (in days)
alb
crp
aag
cer
trf


standardized protein estimate
2
0
−2
−4
0
50
100
150
200
time (in days)


standardized protein estimate
2
0
−2
−4
0
50
100
150
200
time (in days)
alb
crp
aag
cer
trf


correlation
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.348
−30
−20
−10
0
10
20
30
lag.grid