A Functional Estimate of Covariation

J. O. Ramsay
Department of Psychology, McGill University, Ottawa, Ontario, Canada

ABSTRACT
The analysis of functional data calls for a bivariate functional covariance function \( \sigma(s, t) \) that may be evaluated at any discrete set of points to define a variance-covariance matrix \( \Sigma \). This article uses finite element methodology to construct a representation of a functional Choleski factor \( \Lambda \) to define \( \sigma(s, t) = \int \lambda(w, s) \lambda(w, t) \, dw \). An estimate of \( \Sigma^{-1} \) is especially important for applications and, where the eigenstructure of the covariance permits, this is readily available since the resulting \( \Sigma \) is almost always positive definite. A simulation study compares the performance of estimates of \( \Sigma \) and \( \Sigma^{-1} \) to those from the classic covariance matrix estimate and an estimate using the 

Precision surface

1. Introduction
This article presents a methodology for functional covariance estimation that uses a finite approximation of the bivariate covariance function with values \( \sigma(s, t) \). Finite element methods are widely used in the modelling of variation over spatial regions, and especially the models are defined by partial differential equations. A recent application of finite element smoothing is Sangalli, Ramsay, and Ramsay (2013).

An important application of the finite element approximate of \( \sigma(s, t) \) is the evaluation of \( \sigma(s, t) \) at arbitrary discrete points, \( \Sigma(s, t) \), where vectors \( s \) and \( t \) contain strictly increasing series of points. In the special case where \( s = t \) and the series is equally spaced, contains the boundaries of the functional observations and is of length \( n \), we will use the notation \( \Sigma(n) \), along with the notations \( \Sigma^{-1}(s, t) \) and \( \Sigma^{-1}(n) \) for their respective inverses, which will be referred to as the precision matrices.

This estimation strategy, is denoted here as the FE-\( \Sigma \) model, has many applications. An estimate of \( \sigma(s, t) \) can fill gaps in covariance values due to coarse sampling, missing data or uneven spacing of observation times as a preliminary to further analyses or graphical display. The representation almost always provides a nonsingular estimate of \( \Sigma(s, t) \) and, therefore, easy access to an estimate of \( \Sigma^{-1}(s, t) \). The computational efficiency of the estimation permits its use in simulation studies where it offers a new approach to the generation of random symmetric positive definite matrices. The number of parameters used in the estimation does not depend on the size of the functional data sample or the size of either \( s \) or \( t \), and as a consequence the amount of detail in the estimate can be controlled so as to avoid over-fitting the data as well as nonsingularity, which is important in the mixed effect modeling of longitudinal or functional data.

The estimation itself proceeds directly from the raw data available and, therefore, does not depend on a preliminary smoothing of the data, as suggested in Ramsay and Silverman (2005) or Yao, Müller, and Wang (2005).

In this introductory section the properties of the population \( \sigma \) are reviewed, with a particular focus on the properties and estimation of its functional inverse \( \sigma^{-1} \) and the corresponding matrix inverses of evaluations. The finite element approximation is briefly introduced. Some relationships between an evaluation \( \Sigma(s, t) \) of a smooth \( \sigma \) and its inverse are indicated and illustrated. Some historical notes conclude the section.

1.1 The Covariance Surface \( \sigma(s, t) \)
When an observation \( x \) is a random function with mean \( \mu \), the covariance kernel is a symmetric nonnegative definite bivariate function with values

\[
\sigma(s, t) = E[(x(s) - \mu(s))(x(t) - \mu(t))].
\]

In most applications, \( x \) and \( \sigma \) will be smooth in the sense of possessing a certain number of derivatives, and an example of \( \sigma \) drawn from spatial data analysis is in the upper left panel of Figure 1, which displays \( \Sigma(21) \) for the Matérn covariance \( \sigma(s, t) \) (Matérn 1960),

\[
\sigma(s, t) = \alpha \left( \frac{|s - t|}{\beta} \right)^{\nu} K_\nu \left( \frac{|s - t|}{\beta} \right)
\]

over the interval \([0, 20]\) for parameter values \( \alpha = 1, \beta = 0.1 \) and \( \nu = 4 \); and where \( K_\nu \) is the modified Bessel function of the second kind. The lower left panel shows the profile of the surface in the cross-diagonal direction, which is a plot of \( n \) points \((t_i, \sigma_i, n - i + 1)\) along with the \( n - 2 \) averaged
points \([(t_i + t_{i+1})/2, (\sigma_{i+1,n-i+1} + \sigma_{i,n-1})/2]\), \(i = 2, \ldots, n - 1\). Although the Matérn \(\sigma\) is atypical in having a constant diagonal height, it will serve in this article to illustrate a number of aspects of functional covariance estimation, including the need for caution in estimating the inverse evaluation.

### 1.2 The Inverse Covariance or Precision Surface \(\sigma^{-1}(s, t)\)

The inverse covariance matrix \(\Sigma^{-1}\) is in many ways more important than the covariance matrix \(\Sigma\) itself. The linear model requires it, the Gaussian, Wishart, and inverse Wishart distributions are defined in terms of \(\Sigma^{-1}\), the Mahalanobis distance \(D^2 = (x - \mu)'\Sigma^{-1}(x - \mu)\) defines the metric for the analysis of covarying observations and the precision matrix offers a conditional dependency perspective on the data since \((\Sigma^{-1})_{ij} = 0\) if and only if \(x_i\) and \(x_j\) are independent given the remainder of Gaussian observations \(x_i\), contrasting with the marginal dependency framework provided by \(\Sigma\). Recent applications of Gaussian Markov random fields have capitalized on sparse models for \(\Sigma^{-1}\) to develop fast methods for spatial data analysis (Held and Rue 2005; Lindgren, Rue, and Lindström 2011).

The functional analogue of the matrix equation \(\Sigma^{-1}\Sigma x = x\) is

\[
\int \sigma^{-1}(w, s)\text{cov}(f)(w)\,dw
= f(s) \text{ where } \text{cov}(f)(s) = \int \sigma(s, t)f(t)\,dt,
\]

\(\sigma^{-1}\) denotes the functional inverse of \(\sigma\) rather than its reciprocal, and \(\text{cov}(f)\) is the linear covariance operator for which \(\sigma\) is the kernel.

The compact operator \(\text{cov}\) is usually smooth in the sense of having a differentiable kernel, but this unfortunately implies that \(\sigma^{-1}\) is a kernel function so rough that it inhabits an overspace of \(\mathcal{L}^2\), and is referred to in functional analysis as a generalized function or a distribution (Aubin 2000). An image of the effect of inverting a smooth matrix can be seen in the right panels of Figure 1, where we see that the Matérn inverse surface exhibits rapidly decaying symmetric oscillations around the diagonal, with amplitudes that are about two orders of magnitude greater than the cross-diagonal variation of \(\Sigma\). The rate of decay as a function of the number of points on either side of the diagonal remains virtually unchanged as \(n\) increases, as does the overall shape. Thus, going to infinite \(n\), we can imagine the Matérn \(\sigma^{-1}\) as displaying five substantial oscillations infinitely close to the diagonal and infinitely large in magnitude.

This smooth/rough relation between a smooth linear operator and its inverse is familiar in the theory of linear differential equations and their associated Green’s functions, and also for other common integral transforms such as the Laplace transform, the Fourier transform, and convolutions. The linear inverse problem is that of finding a useful approximation to the inverse of such transforms, and successful strategies almost inevitably involve some degree of regularization.

The smoothness/roughness relation is also illustrated in Figure 2. The upper left panel displays the surface

\[
\sigma(s, t) = (1+2\exp(-(s + t)/4) + 0.5\exp(-(s + t - 9)^2/4)) \times \cos(2\pi|s - t|/4)\exp\left(-((s - t)/2)^2\right)/4.
\]

This is a stylized version of what will be estimated in Section 4 as the covariance structure of residuals from the smoothing of human growth data. We see in the cross-diagonal of the covariance an oscillation decaying to zero in about five units.
Figure 2. The upper panels show the hypothetical growth residual covariance surface and the corresponding precision surface, respectively. The lower left and right panels show the corresponding cross-diagonal variations.

The lower right panel shows the precision matrix oscillation, and we see that this decays much more rapidly than its counterpart in Figure 1. This illustrates that rougher covariance surfaces tend to correspond to more gentle cross-diagonal variation in their inverses.

**1.3 The Inversion of Perturbed Covariance Surfaces**

The smooth/rough relation between a smooth covariance surface and its inverse can create a second difficulty of great concern to statisticians whose data framework involves measurement error. We already know from numerical analysis that, if the condition number of a matrix is large, the computed inverse of the covariance matrix will have substantial rounding error. The condition numbers of the Matérn surface in Figure 1 for 11, 21, and 41 equal-spaced observations are $1.3 \times 10^3$, $6.0 \times 10^6$, and $3.1 \times 10^9$, respectively. These already dangerous levels illustrate that the higher the sampling rate, the more severe the conditioning problem will become. On the other hand, the condition number of the growth residual surface in Figure 2 is a safe 7.7.

In this article we consider a method for approximating a smooth covariance surface $\Sigma$ by a bivariate function $\hat{\Sigma}$ which we expect will be computed from data that are noisy. This implies two sources of error: that arising from the approximation procedure, and that due to the variability of the data. Figure 3 displays the consequences of these two sources of error for the respective inverses of the perturbed covariance matrices. In Figure 3 the top panel shows the approximations of the Matérn covariance and the direct approximation of its inverse, both resulting from evaluations at 21 equally spaced points. A visual comparison of these approximations with their exact counterparts in Figure 1 shows that the approximation to both $\Sigma_{(21)}$ and $\Sigma_{(21)}^{-1}$ are useful. For the covariance surface the square root of squared errors summed over 441 values is 0.026, and the corresponding value for the inverse matrix is 0.1, an impressive value when viewed against matrix values reaching about $\pm2000$ for $\Sigma^{-1}$.

But the bad news is seen in the lower two panels, which shows what happens when each approximation is inverted. Inverting the approximation wipes out most of the interesting detail seen in the exact inverse. The nature of the approximation is not the problem. For example, when the error arises by perturbing each surface by a random Wishart-distributed matrix having a small number degrees of freedom and rescaled so has to have amplitudes that are only 0.1% of that of the corresponding matrix diagonal, we can barely see the perturbations in a plot of $\Sigma_{(21)}$, but there is still severe degradation of shape from inverting the corresponding respective perturbed inverses. In spite of the popularity of the Matérn covariance model in spatial data analysis, its high condition number makes it an extreme case that we should rather hope to avoid in practice. A good account of the impact of perturbation on the inversion of matrix can be found in Stoer and Bulirsch (2010).

On the other hand, the inverse growth residual surface in Figure 2 is hardly affected at all by perturbations from these sources, and the inverse of the approximation to the surface is an excellent representation of the exact inverse. The conclusion is that one should estimate an inverse by inverting an estimate only after careful consideration of the condition number of the matrix to be estimated. To do so, however, is not a simple matter since we usually don’t have the real surfaces at our disposal. A comparison of the cross-diagonals of the Wishart and growth residual surfaces suggests that those with most of their variation concentrated around the diagonal combined with healthy condition numbers for their approximations are apt to have useful inverse-approximations.
These two difficulties, the smoothness/roughness tradeoff and the sensitivity of the inverse matrix to perturbation, are further compounded by the problems caused by the use of covariance estimation methods appropriate to multivariate statistical data in the functional situation. When a sample of $N$ observations $y_{ij}$, $j = 1, \ldots, n_i$; $i = 1, \ldots, N$ are available, the usual estimate of $\sigma$ involves an initial approximation of the discrete data by a set of functions $x_i$, and then the application of the functional version of the usual multivariate estimate, namely

$$\hat{\sigma}(s, t) = N^{-1} \sum_i [\hat{x}_i(s) - \bar{\hat{x}}(s)][\hat{x}_i(t) - \bar{\hat{x}}(t)],$$

where the $\hat{x}_i$'s and $\bar{\hat{x}}$ are estimates of the $x_i$'s and their pointwise mean, respectively. However, this estimation procedure has a number of undesirable features. It ignores the nature of the dependency of each $x_i$ on the corresponding data due to measurement errors, variation in the distribution number of the $y_{ij}$'s, the number $K$ of basis functions used to represent the $x_i$'s, and other aspects of the smoothing process. Moreover, positive definiteness of $\hat{\sigma}$ will be completely lost if $N$ is smaller than the $n_i$'s or $K$. The methodology presented in this article works directly with the discrete data rather than requiring an intermediate smoothing step and is virtually always positive definite.

1.4 An Overview of FE-$\Sigma$ Smoothed Covariance Surfaces

The FE-$\Sigma$ estimation strategy uses a functional version of the Choleski decomposition $S = L^T L$ of an estimate $\hat{S}$ of a covariance matrix derived from functional data. This is

$$\hat{\sigma}(s, t) = \int_{\Omega(s, t)} \hat{\lambda}(w, s)\hat{\lambda}(w, t) \, dw,$$

where the linear integral operator $\Lambda$ on function $f$

$$\Lambda f)(s) = \int_{\Omega} \hat{\lambda}(w, s)f(w) \, dw$$

is the functional analogue of matrix operation $Lf$ on vector $f$. The domain of $\lambda$ is a parallelogram, a natural modification in the functional context of the lower triangular structure of $L$, and is illustrated for a simple example in the left panel Figure 4. The width of the parallelogram can be chosen so as to construct band-structured covariance surfaces, and the right panel of the Figure displays the corresponding positive support of the estimate of $\sigma$.

Finite element approximations of surfaces over spatial domains are essentially the generalization of spline approximations over line segments. A regular triangular mesh is constructed over the parallelogram domain, where "regular" means that no vertex of a triangle is in the interior of another triangle's
Figure 4. The left panel contains the parallelogram domain of the function Cholesky factor $\lambda$ for $T = 4$ with values $\lambda(w, s), s \in [0, T], T - t - B \leq w \leq T - t$, and lag $B = 2$. The finite element triangulation into 16 triangles is outlined by the heavy solid lines. The horizontal dashed lines indicate three possible values of $s$ or $t$, and over this domain $\hat{\sigma}(0.5, 3.5) = 0$. There are 15 vertices, and consequently 15 basis functions $\phi_{k\ell}$, each basis function being nonzero only over the hexagonal region surrounding its vertex and within the parallelogram, an example being shaded in red. The right panel contains the corresponding domain and finite element triangulation for the covariance estimate $\hat{\sigma}(s, t)$.

edge, just as the design of spline representation of over a line begins with the subdivision of an interval into interior line segments. In Figure 4 this is achieved by subdividing the ordinate into $I = 4$ equal intervals, subdividing the abscissa into $J = 2$ intervals of the same width, and dividing each square within the parallelogram into two right triangles.

The second step is construction of basis functions having local support, analogous to the of a spline basis function over a small number of adjoining intervals. As with polygonal or order two splines constructed from linear combination of tent functions centered on knots, a simple but effective basis function is centered on a single vertex, is piece-wise linear, and nonzero only over the interior of the hexagonal region covered by triangles sharing that vertex location. For example, the basis function centered on the point at $(1, 2)$ in the left panel of Figure 4 has value one at that point and declines linearly to a value of zero on the hexagonal region surrounding that point. Further details are taken up in Section 2.1.

We designate such a basis function for building $\hat{\lambda}$ as $\phi_{k\ell}(w, s)$, the use of index pairs being natural for this application, so that the representing function is

$$\hat{\lambda}(w, s) = \sum_{k=0}^{I} \sum_{\ell=0}^{J} c_{k\ell} \phi_{k\ell}(w, s).$$

(4)

The number $(I + 1)(J + 1)$ of basis functions does not depend on the order of any $\Sigma$ or $\Sigma^{-1}$ to be estimated, so that $\hat{\sigma}(s, t|C)$ is defined by a coefficient matrix $C$ whose size can be adapted to the amount and accuracy of available data and the desired complexity of $\hat{\sigma}$.

Figure 5 displays the $\lambda$ surface corresponding to the Matérn covariance surface in Figure 1. The ridge, starting at a sharp peak

Figure 5. The Choleski factor $\lambda(w, s)$ corresponding the Matérn covariance surface in Figure 1.
near \((T, 0)\), and progressively broadening toward \((-10, T)\) is fairly typical of the topography of \(\lambda\) when the covariance surface is strongly diagonal.

### 1.5 Historical Notes and Overview

There is already a substantial literature in multivariate statistics on parameterizations \(\Sigma(\theta)\) of the covariance matrix that circumvent the greediness of the estimate \(\hat{\Sigma}\) with respect to degrees of freedom. Pinheiro and Bates (2000) described several patterned covariance structures, and Smith and Kohn (2002) have added to these by using the normalized Choleski decomposition \(\Sigma = L^TDL\), where the diagonal of the Choleski factor \(L\) consists of ones. Band-structured matrices, however, do have a design dependence on the order of the covariance matrix that is avoided by the approach in this article. Most structured covariance proposals also presume equally spaced times of observation.

FE-\(\Sigma\) smoothing shares some features with the graphical lasso that Friedman, Hastie, and Tibshirani (2008) developed to estimate large \(\Sigma^{-1}\) where \(n \gg N\). Their approach uses the \(\ell_1\) norm to impose a sparsity on the estimate. Their R package \texttt{glasso}, moreover, permits the constraint that values in \(\hat{\Sigma^{-1}}\) beyond a prespecified distance from the diagonal are zero. However, a functional version of \(\sigma\) is not produced that can be used to compute covariance matrices for any vector of time values; and, since the method does not use the spacings between \(t\)-values, \texttt{glasso} would only be appropriate for equally-spaced observation points. Moreover, the aim in this article is not sparsity, but rather smoothness; although admitted sparsity also acts as a regularization principle. Simulation-based comparisons of the performance of the \texttt{glasso} package and this article’s methods are provided in Section 3.

Yao, Müller, and Wang (2005) developed a technique for the principal components analysis of sparsely and irregularly sampled functional observations that has been successfully adapted to a considerable number of other functional estimation problems. They use conditional expectation to compute principal component scores, and this requires an estimate of \(\hat{\Sigma^{-1}}\) for each functional observation. In their software package \texttt{PACE}, a relatively involved and ad hoc bivariate kernel smoothing procedure is used to convert \(N\) one-dimensional smoothing matrices \(S_i\) into an estimate of \(\sigma(s, t)\) whose evaluation matrices can be inverted. Unfortunately, the covariance surfaces produced in this fashion are not guaranteed to be positive definite. In fact, using Matlab code supplied by one of the authors for estimating surfaces from data simulating those in Figures 1 and 2, it was observed that not one of 1024 estimated surfaces was positive definite. Section 3 provides further details.

Finite element methods were originally developed to approximate solutions to partial differential equations, and consequently there is a large literature on the topic in both numerical analysis and engineering. A readable introduction that focusses on both spline approximation and computation is Larson and Bengzon (2013) and a more comprehensive treatment of higher dimensional spline approximation is Lai and Schumaker (2007).

The following section provides details on how the functional Choleski factor \(\Lambda(w, s)\) is constructed using a finite element representation and the properties of the corresponding functional variance estimate, as well as the estimation of \(\hat{\Sigma}\) and \(\hat{\Sigma^{-1}}\) from a collection of sample covariance matrices. Section 3 contains the results of some simulation experiments, and this is followed by Section 4 describing a real-data application. The supplemental materials associated with this article provide further detail on computational aspects of FE-\(\Sigma\).

### 2. The FE-\(\Sigma\) Estimate \(\hat{\sigma}(s, t)\)

A method for fast computation of the corresponding representation of \(\hat{\sigma}(s, t)\) is presented, and extended to include the superposition of two or more covariance surfaces to permit varying levels of resolution over the domain of \(\sigma\). A roughness penalty is also specified to impose smoothness on the result.

#### 2.1 Linear Finite Element Basis Functions \(\phi_{k\ell}(w, s)\)

We assume without loss of generality that the functional data are defined over the interval \([0, T]\). The domain \(\Omega\) of Choleski factor \(\Lambda\) is the parallelogram defined by points \((T, 0), (0, T), (-B, T),\) and \((T - B, 0)\). Here \(B \geq 0\) is a lag such that, for \(|s - t| \geq B\), \(\sigma(s, t) = 0\). These features are illustrated in the left panel of Figure 4. The interval \(\Omega(s, t)\) of integration over \(w\) in (2) depends on \((s, t)\) and is

\[
\Omega(s, t) = [T - \max(s, t), T - \min(s, t) - B].
\]

But, when \(B > T\), the capacity of \(\Lambda\) to capture covariance as a far out as \(\hat{\sigma}(0, T)\) improves, and \(B = 2T\) corresponds to an effectively unconstrained covariance kernel. That is, variation in \(\Lambda\) over \(s\) tends to capture variation in \(\sigma\) along its diagonal, and variation over \(w\) does the same for its cross-diagonal shape. The domain for the nonzero portion of \(\sigma\) is plotted in the right panel of Figure 4.

The basis function expansion (4) is defined by subdividing the domain \(\Omega\) into triangular subregions, and the regular triangulation shown in Figure 4 is natural and convenient for computation, although by no means the only possible triangulation. Let \(I > 0\) be a number of equal intervals, each of length \(\delta = T/I\), spanning \([0, T]\), and define \(B = I\delta\), so that \(J > 0\) is the number of intervals of the same size spanning \([0, B]\). That is, \(T/B\) is a simple fraction, and in the Figure 4 is \(1/2\). The resolution of the triangulation, therefore, is completely defined by the number of intervals \(J\) and the lag number \(I\). There are \(2IJ\) right triangles covering the domain, divided equally into those with left- and right-orientations; and there are \(M = (I + 1)(J + 1)\) vertices. Each triangle is a \textit{finite element} in the terminology of finite element analysis used in the approximation of solutions of partial differential equations (Grossmann, Roos, and Stynes 2007).

Each basis function \(\phi_{k\ell}\) is associated with one of the \(M\) vertices, so that coefficient indices \(k = 0, \ldots, I\) and \(\ell = 0, \ldots, J\) are associated with the vertical and horizontal coordinates of nodes defining \(\lambda\), respectively. Let \(I + 1\) by \(J + 1\) matrix \(C\) contain these coefficients, and let \(c = \text{vec}(C)\) denote this matrix stored column-wise as a vector. The basis function \(\phi_{k\ell}\) is defined to be piecewise linear, to have value one at node \((k, \ell)\), to decline to value 0 at the opposite edges of the six triangles that share this node, and to be zero beyond the hexagonal region covered...
by these triangles. However, when a vertex is on the boundary of the domain, it can be assumed without loss of generality that the basis function height outside of the domain is 0, which permits the notational simplification of defining all integrations over \( w \) as being over \([-B, T]\). Defining the basis functions in this way implies that \( c_{k\ell} \) is the height of the \( \lambda \)-surface at node \((k, \ell)\).

The corresponding covariance surface (5) is also piecewise linear, and is defined over the triangular mesh shown in the right panel of Figure 4. Covariance \( \tilde{\sigma} = 0 \) outside of the triangulated region and we see, therefore, that \( \tilde{\sigma}(0.5, 3.5) = \tilde{\sigma}(3.5, 0.5) = 0 \).

The choice of \( I \) and \( J \) very much depends on whether one is approximating \( \Sigma(t) \) or \( \Sigma^{-1}(t) \), as well as on the number of sampling points and their spacing in \( t \). The number \( I \) of intervals over \([0, T]\) controls the resolution or amount of detail in the estimated surface. The quality of the resolution is of course also limited by the number and spacing of the sampling points. When \( \sigma \) is smooth, as will usually be the case, the shape of the estimated surface will not depend critically on \( I \), and therefore one can, as we do in spline smoothing, control the smoothness of the estimated surface by the choice of \( I \). That is, the smaller \( I \), the smoother the estimated surface, but at a cost of missing fine detail. On the other hand, we see in both the Matérn and growth surfaces the general principle that the shape of \( \Sigma^{-1}(t) \) will depend critically on the number and spacing in \( t \). Consequently, in the equal spacing case, it can be useful to set \( I \) to \( n - 1 \).

The number \( J \) plays the role of eliminating the estimation of the surface at lags \(|s - t|\) larger than those where the covariance has any appreciable deviation from zero. For the estimation of smooth covariance surfaces, this will have to be chosen on the basis of some preliminary exploration of the surface. As noted, \( J = I \) implies a surface equal to zero for lags of size \( T \); but this will not generally make sense for periodic records, where \( J \) may be required to permit an unrestricted covariance surface. On the other hand, the estimation of \( \Sigma^{-1}(t) \) will benefit from making \( J \) only large enough to catch the oscillations of this surface around the diagonal, which, for smooth covariance structure, will decay rapidly and at a rate that depends on \( n \). In other words, the need for \( J \) to be large in this inverse covariance case can be offset by the possibility of reducing \( I \).

### 2.2 Computing Inner Product Matrices \( R(s, t) \)

Equation (5) implies that the covariance surface \( \sigma(s, t) \) will be piecewise linear and differentiable to the first order. Defining the order \( M \) symmetric matrix \( R(s, t) \), called the mass matrix in the finite element literature, as containing the inner products

\[
r_{(k_1, \ell_1), (k_2, \ell_2)}(s, t) = \int_B^T \phi_{k_1, \ell_1}(w, s)\phi_{k_2, \ell_2}(w, t) \, dw,
\]

leads to an expression that is quadratic in \( c \),

\[
\tilde{\sigma}(s, t) = \mathbf{c}' \mathbf{R}(s, t) \mathbf{c}.
\]

The computation and storage of the \( \mathbf{R}(s, t) \) matrices will be a first step in an application prior to an optimizing of some fitting criterion with respect to \( c \). Fortunately, most of the elements in each matrix will correspond to nonoverlapping basis functions and, therefore, have the value 0 so that storing them in a sparse storage mode will greatly economize on disk space.

For example, an application involving daily annual weather data \((n = 365)\) and \( M = 72 \) would require about \( 8 \times 10^{10} \) bytes to store all the matrices in full storage model, but only a tolerable \( 6 \times 10^5 \) in sparse mode. A typical application will involve both a computation of \( R \)-matrices over a set of \( t \)-values at data points, and over a fine mesh for display purposes. If the sampling points vary in spacing and/or number from one record to another, a set of \( n\alpha(n_1 + 1)/2 \) \( R \)-matrices will be required for each record. The computation of these matrices requires considerable discussion, and is, therefore, detailed in the supplementary materials.

### 2.3 Combining Multiple Covariance Surfaces

The phenomenon of covariance can arise from multiple sources, and in particular it is not uncommon to observe strong local covariances associated with small lags \(|s - t|\), combined with a longer-range smoother covariance structure over larger lags. This is illustrated in the analysis of the Fels growth data in Section 4 and is also the principal rational for the factor analysis model in multivariate statistics. Covariance matrices with dominant diagonals in general have much lower condition numbers, so that enhancing the diagonal regions of an approximation can greatly improve the quality of the inverse of the approximation as an estimate of its true inverse.

Superimposing two or more covariance structures can be accommodated within this framework by combining two or more covariance surfaces within the additive structure

\[
\sigma(s, t|I, J) = \sum_{r=1}^p c_r \mathbf{R}(s, t|I_r, J_r) c_r.
\]

The short-range covariance would involve a term in this expansion with a fairly large \( I \) combined with a small \( J \), while the long-range covariance detail would be modeled with smaller values of \( I \) and \( J \) with \( J \) being equal to or even greater than \( I \).

### 2.4 A Roughness Penalty \( \gamma \mathbf{P}(\mathbf{c}) \) for \( \lambda(w, s) \)

The smoothness of \( \sigma \) is determined by that of \( \lambda \), which in turn can be quantified by

\[
P(\mathbf{c}) = \gamma \mathbf{c}' \mathbf{K} \mathbf{c},
\]

where the highly sparse order \((I + 1)(J + 1)\) matrix \( K \) contains the inner products

\[
k_{(k_1, \ell_1), (k_2, \ell_2)} = \iint_B^T \left[ \frac{\partial^2 \phi_{k_1, \ell_1}}{\partial w \partial s} \frac{\partial^2 \phi_{k_2, \ell_2}}{\partial w \partial s} + \frac{\partial \phi_{k_1, \ell_1}}{\partial s} \frac{\partial \phi_{k_2, \ell_2}}{\partial s} \right] \, dw \, ds.
\]

By augmenting a fitting criterion \( F(\mathbf{c}) \) to be minimized to \( G(\mathbf{c}) = F(\mathbf{c}) + \gamma \mathbf{P}(\mathbf{c}) \), finer control over the smoothness of \( \sigma \) can be achieved by manipulating the smoothing parameter \( \gamma \) than is possible by keeping \( I \) small.

### 2.5 Estimation from \( N \) Covariance Matrices

Let the observed data be in the form of \( N \) sample covariance matrices \( \mathbf{S}_i \) of order \( n_i \), respectively, possibly involving unequally
spaced $t$-values; and let $\Sigma_i(c)$ denote the estimated covariance matrix for case $I$ and let $G(\Sigma, S)$ be a loss function. Then the fitting criterion

$$F(c) = N^{-1} \sum_i^N G(\Sigma_i(c), S_i) + \gamma c' K c$$

measures the total fit to the data with an added penalty for the roughness, where the stiffness matrix $K$ is defined in (8) and $\gamma$ controls the size of the penalty on roughness. In the balanced design case where all records have the same observation points, $N$ can be taken as 1. The negative log-likelihood of the Wishart distribution in the loss function is equivalent to

$$F(c) = \text{trace}[S \Sigma^{-1}(c)] - \log |\Sigma^{-1}(c)| + \gamma c' K c$$

and its counterpart for inverse Wishart distribution has a negative sign on the second term. Note that this loss function is defined in terms of $\Sigma^{-1}$ rather than $\Sigma$ and, therefore, its natural application is to the estimation of $\Sigma^{-1}$.

3. A Simulation Experiment

This simulation experiment is designed to show how well $S^{-1}$, FE-$\Sigma$ and glasso perform for the more difficult problem of estimating $\Sigma^{-1}$ for data arising from two functional covariance structures representative of what might be seen in practice. Figure 6 displays the covariance surface for the logarithm of average annual precipitation for 35 Canadian weather stations, described by Ramsay and Silverman (2005). This surface is periodic with strong covariance patterns widely disbursed over the year, and has condition number 581.4. The second example is the stylized growth residual surface shown in Figure 2 which has covariance more concentrated about the diagonal. Both matrices are the result of evaluating the surfaces over 21 equally spaced points.

The results in Table 1 are based on 1024 covariate matrices $S$ sampled from the Wishart distribution with 50, 100, 200, and 400 degrees of freedom. The values of $I$ and $J$ used in the analyses are given in parentheses after the name of the method.

![Figure 6](image-url) The covariance surface $\Sigma_{(21)}$ for average annual log precipitation measured at 35 Canadian weather stations.
root-mean-squared-errors (RMSE) and squared multiple correlations or variance ratios $R^2 = \| \Sigma^{-1} - \Sigma_\text{true}^{-1} \|^2 / \| \Sigma_\text{true}^{-1} \|^2$. The accuracy of the estimate $S^{-1}$ provides a useful baseline for assessing performance. All the FE-$\Sigma$ analyses used the inverse Wishart criterion since the objective was to estimate $\Sigma^{-1}$.

We see in the top panel of the table that the FE-$\Sigma$ analysis with largest basis defined by $I = J = 20$ performs about as well as $\Sigma^{-1}$ for both surfaces, which is to be expected since it has the capacity to nearly interpolate each data matrix. Good recovery is obtained for $N = 200$ and 400, some value is in the $N = 100$ estimates, but $N = 50$ is not enough data to support a useful estimate.

There are three ways to impose some smoothness on the estimated surfaces: (1) use a coarser grid such as $I = J = 10$ shown in the second panel, (2) take advantage of the known strongly diagonal structure in $\Sigma^{-1}$ by reducing $J$ while keeping $I = 20$, as shown in the third and fourth panels where $J = 6$, and (3) use the roughness penalty in (9). Results with the roughness penalty failed to achieve any interesting improvement, and are not shown. We see that a compact band-structured estimate using FE-$\Sigma(20,6)$ was able to recover useful information about $\Sigma^{-1}$ for $N = 50$ and also improve the estimation for the growth surface right up to $N = 400$. The coarse grid approach did not work, however, because too much detail in the surface was sacrificed. The band-structured fit by gglasso failed to improve the estimation of the inverse relative to $\Sigma^{-1}$, primarily because it essentially interpolated the $\Sigma$ surface.

The recoveries of the surface $\Sigma$ itself using the Wishart criterion are not shown, since both FE-$\Sigma$ and gglasso performed approximately as well as $S$ itself, but with the exception that the coarse grid FE-$\Sigma(10,10)$ was substantially better for the growth surface. Analyses of simulated samples using the PACE code supplied by the authors were also carried out, but the quality of the estimates was found to be too poor to be competitive for recovery of $\Sigma$, and to be virtually always nonpositive definite. The PACE code compensates for this by adding a judicious constant to the diagonal of the surface. However, this seems undesirable given that it is an estimate of $\Sigma^{-1}$ that is used for estimation of principal component and other structures by that approach.

In summary, the simulation suggests that FE-$\Sigma$ will do at least as well as $S^{-1}$, but that building a strong band-structure into the estimate will be profitable, and especially for noisier data and smaller sample sizes. The use of coarser bases can also be helpful for smoothing small sample covariance matrices. R function gglasso seems not to have much to offer to the estimation of $\Sigma^{-1}$.

4. A Factor Analysis of the FELS Female Growth Residuals

The Fels growth data bank (Roche 1992) is one of the world’s largest collections of systematic measurements of the height of children, and now contains third generation records. Heights were measured at roughly six-month intervals after age 3 and up to a maximum of age 22; but the years from birth to three were more densely sampled. A variable indicating when the measurement was of supine body length was included. Ramsay, Altman, and Bock (1994) compared results from the Fels data base with those from the Zurich data of similar size, and noted that the only significant difference was that the average age of puberty in the Fels data was advanced by about six months. In the analyses described here, the focus is on detecting any residual covariance structure that may be present in a child’s height measurements after its growth curve has been estimated using the monotone smoothing process described in Ramsay (1998), Ramsay and Silverman (2005), and Ramsay, Hooker, and Graves (2009).

The variation from child to child in the age of puberty is the largest source of variation in growth data, and inevitably affects the size of the estimated residual variation. Consequently, the fits $\hat{y}_i(t)$, $i = 1, \ldots, 161$ to the data between ages 3 and 18 were registered by the continuous registration method of Ramsay and Li (1998). The registration process nonlinearly transforms the physiological or growth times $t_{ij}$, $j = 1, \ldots, n_i$ to the clock times $h_i(t_{ij})$ so that, even if the original measurement times were common across records (they were not for the Fels data), the spacings between observation times after registration inevitably varied within a girl’s record and from one girl to another.

After monotone smoothing and registration, the residual values for each girl $\hat{e}_{ij} = y_{ij} - \hat{y}_i[h_i(t_{ij})]$ were calculated and a rank 1 covariance matrix $S = e'e$ computed. The FE-$\Sigma$ estimation used the full objective function for unbalanced data described in Section 2.5. The model combined two FE-$\Sigma$ covariance structures as described in Section 2.3, the first defined by a coarse mesh with numbers of intervals $I = 15$, $J = 5$, and lag $B = 5$ years; and the second with a fine but strongly diagonal mesh with numbers of intervals $I = 30$, $J = 4$, and lag $B = 2$ years. In this way, short-range strongly diagonal variation could be combined with longer-range correlation while still using only a modest number of basis functions for the representation of the Choleski factor. Cross-validation was used to select a smoothing parameter $\gamma$ by using a random 2/3 of the cases as a training sample and the remainder as a validation sample. This selected $\gamma = 0$ as the optimal smoothing parameter, as was expected in view of the limited dimensionality of the basis system.

Figure 7 shows the composite residual covariance surface estimated at 31 equally spaced values in the upper left panel. Along the diagonal, the residual standard deviation is about 0.4 cm at age 3, declines to about 0.3 cm by age 8, moves back to 0.4 cm over the adolescent years, and then declines to 0.3 cm at 18 years. The standard deviation contributed by fine diagonal mesh contributes almost all of the total diagonal at age 3, a substantial proportion in later childhood, and again at age 18, but relatively little to the adolescent years. The covariance contributed by the fine mesh at lag one year is $-0.05$ cm at almost all ages. The cross-diagonal plot in the lower left panel, which is positioned at 13 years at the diagonal, indicates that covariance drops from its diagonal value of 0.2 to $-0.05$ over a lag of two years, returns to positivity at 3 years, and then drops to zero just beyond 4 years. The negative trench on either side of the diagonal persists over all but the last years of growth, which suggests a pulse-like character to growth. This is consistent with the oscillation in the first derivative of height noted in Ramsay and Silverman (2005) in the analysis of the data the growth of a ten year old boy (Thalange et al. 1996).

The condition number of the covariance matrix is 525.7, and the simulation results encourage the inspection of the inverse
of this matrix. The right panels display the residual precision matrix and its cross-diagonal, also positioned at 13 years, respectively. If used as a weight matrix, $\hat{\Sigma}^{-1}$ would weight the nearest neighbors of a wide neighborhood during the childhood years; but we recall that the data were registered before the analysis, and this probably had some effect on these results, and especially at the age of puberty.

5. Discussion and Conclusions

The primary benefit of the FE-$\Sigma$ methodology is to represent covariation in truly functional terms as a bivariate function $\sigma(s, t)$ constructed from a finite element approximation of its functional Choleski factor. That this $\sigma$-surface can be evaluated anywhere over the interval of observation is an aspect that is convenient for graphical display as well as filling in or gridding for unevenly distributed or sparse data. The FE-$\Sigma$ procedure also permits the estimation of composite surfaces composed from fits with different horizontal and vertical resolutions of $\lambda(w, s)$, as illustrated by the analysis of the Fels residual data. This aspect provides a promising alternative to the PACE method for covariation estimation of Yao, Müller, and Wang (2005).

A key insight during the development of this method was that a satisfactory approximation of covariance matrix does not necessarily imply a useful approximation of its inverse. A large condition number can mean that a small perturbation of $\Sigma$ can result in an arbitrarily large perturbation of $\Sigma^{-1}$. This is important for the many statistical methods using matrix inverses, including regression analysis and regression smoothing where the least squares estimate depends on $(X'X)^{-1}$, and where recent research has highlighted the surprisingly large impact of even fairly small errors in covariates on estimates of regression coefficients and the predictions that they define (Carroll et al. 2006; Ruppert, Wand, and Carroll 2003).

Nevertheless, where the eigenstructure of the matrix is reasonably well conditioned, as for the log precipitation and growth residual structures, the simulation results confirmed that the FE-$\Sigma$ can provide a more accurate estimate of $\Sigma^{-1}$ than either the inverse of $\mathbf{S}$ or the glasso estimate, which in any case is inapplicable for unevenly distributed sampling times such as those in the Fels data.

FE-$\Sigma$ methodology can be generalized in many ways. Although the use of a lattice structure for $\lambda(w, s)$ with $w$ and $s$ increments of the same size brings computational and programming advantages, lattices constructed to allow $J$ to vary over $s$ are feasible and under development by the author. Extensions to quadratic and higher order finite elements are described in many texts on finite element methodology.

Supplementary Materials

An appendix is provided with additional details on computational issues.

Acknowledgments

The author gratefully acknowledges funding by the Natural Science and Engineering Research Council of Canada.

References

Aubin, J.-P. (2000), Applied Functional Analysis (2nd ed.), New York: Wiley-Interscience. [161]
Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), Measurement Error in Nonlinear Models: A Modern Perspective (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC. [169]
Friedman, J., Hastie, T., and Tibshirani, R. (2008), “Sparse Inverse Covariance Estimation With the Graphical Lasso,” Biostatistics, 9, 432–441. [165]
Grossmann, C., Roos, H.-G., and Stynes, M. (2007), Numerical Treatment of Partial Differential Equations, Berlin: Springer. [165]
Held, L., and Rue, H. (2005), Gaussian Markov Random Fields, New York: CRC Press. [161]
Lai, M.-J., and Schumaker, L. L. (2007), Spline Functions on Triangulations, Cambridge: Cambridge University Press. [165]
Larson, M. G., and Bengzon, F. (2013), The Finite Element Method: Theory, Implementation and Applications, New York: Springer. [165]
Lindgren, F., Rue, H., and Lindström, J. (2011), “An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach,” Journal of the Royal Statistical Society, Series B, 73, 423–498. [161]
Matérn, B. (1960), “Spatial Variation,” Meddelanden Fran Statens Skogsforskningsinstitut, 49. [160]
Pinheiro, J. C., and Bates, D. M. (2000), Mixed-Effects Models in S and S-PLUS, New York: Springer. [165]
Ramsay, J. O. (1998), “Estimating Smooth Monotone Functions,” Journal of the Royal Statistical Association, Series B, 60, 365–375. [168]
Ramsay, J. O., Altman, N., and Bock, R. D. (1994), “Variation in Height Acceleration in the Fells Growth Data,” Canadian Journal of Statistics, 22, 89–102. [168]
Ramsay, J. O., Hooker, G., and Graves, S. (2009), Functional Data Analysis in Matlab and R, New York: Springer. [168]
Ramsay, J. O., and Li, X. (1998), “Curve Registration,” Journal of the Royal Statistical Association, Series B, 60, 351–363. [168]
Ramsay, J. O., and Silverman, B. W. (2005), Functional Data Analysis, New York: Springer. [160,167,168]
Roche, A. F. (1992), Growth, Maturation and Body Composition: The Fels Longitudinal Study 1929-1991, Cambridge: Cambridge University Press. [168]
Ruppert, D., Wand, J. P., and Carroll, R. J. (2003), Semiparametric Regression, Cambridge: Cambridge University Press. [169]
Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013), “Spatial Spline Regression Models,” Journal of the Royal Statistical Society, Series B, 75, 681–703. [160]
Smith, M., and Kohn, R. (2002), “Parsimonious Covariance Matrix Estimation for Longitudinal Data,” Journal of the American Statistical Association, 97, 1141–1153. [165]
Stoer, J., and Bulirsch, R. (2010), Introduction to Numerical Analysis (3rd ed.), New York: Springer. [162]
Thalange, N. K., Foster, P. J., Gill, M. S., Price, D. A., and Clayton, P. E. (1996), “Model of Normal Prepubertal Growth,” Archives of Disease in Childhood, 75, 427–431. [168]
Yao, F., Müller, H. G., and Wang, J. L. (2005), “Functional Data Analysis for Sparse Longitudinal Data,” Journal of the American Statistical Association, 100, 577–590. [160,165,169]