Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

August 6, 2017

Abstract

With high dimensional longitudinal and functional data becoming much more common, there is a strong need for methods of estimating large covariance matrices. Estimation is made difficult by the instability of sample covariance matrices in high dimensions and a positive-definite constraint we desire to impose on estimates. A Cholesky decomposition of the covariance matrix allows for parameter estimation via unconstrained optimization as well as a statistically meaningful interpretation of the parameter estimates. Regularization improves stability of covariance estimates in high dimensions, as well as in the case where functional data are sparse and individual curves are sampled at different and possibly unequally spaced time points. By viewing the entries of the covariance matrix as the evaluation of a continuous bivariate function at the pairs of observed time points, we treat covariance estimation as bivariate smoothing.

Within regularization framework, we propose novel covariance penalties which are designed to yield natural null models presented in the literature for stationarity or short-term dependence. These penalties are expressed in terms of variation in continuous time lag and its orthogonal complement. We present numerical results and data analysis to illustrate the utility of the proposed method.

keywords: non-parametric, covariance, longitudinal data, functional data, splines, reproducing kernel Hilbert space

1 Introduction

Covariance estimation with high dimensional data has has recently gained growing interest; it is generally recognized that there are two primary hurdles responsible for the difficulty in covariance estimation: the instability of sample covariance matrices in high dimensions and a positive-definite constraint we wish estimates to obey. Estimation of population covariance matrices from samples

*The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201
†The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201

of multivariate data has been important for methods in classical multivariate data analysis, time series analysis, spatial statistics and, more recently, the growing field of statistical learning. Covariance estimates play a critical role in establishing independence or conditional independence through graphical models, constructing discriminant functions as in linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) for the classification of Gaussian data, building confidence intervals for component means and contrasts, and constructing a low-dimensional representation of data via principal components analysis (PCA). One may note that the last two techniques require an estimate of the covariance matrix, and the first two require estimation of the inverse.

Prevalent technological advances in industry and many areas of science make high dimensional longitudinal and functional data a common occurrence, arising in numerous areas including medicine, public health, biology, and environmental science with specific applications including fMRI, spectroscopic imaging, gene microarrays among many others, presenting a need for effective covariance estimation in the challenging situation where parameter dimensionality is possibly much larger than the number of observations. Additional difficulty due to constraints required to yield positive definite estimates make covariance estimation a potentially complex optimization problem. Further, most existing approaches to covariance estimation require data to be sampled at regular grid (time) points, with subjects sharing a set of common observation points. However, in many practical situations, data are irregularly sampled, and subjects may share few common observation times, and methods are needed to accommodate for data collected in this way.

To address the challenge of enforcing positive definiteness, several have considered modeling various matrix decompositions including variance-correlation decomposition, spectral decomposition, and Cholesky decomposition. The Cholesky decomposition has received particular attention, as it which allows for a statistically meaningful interpretation as well as an unconstrained parameterization of elements of the covariance matrix. This parameterization allows for estimation to be accomplished as simply as in least squares regression.

It is well known that the sample covariance matrix is unstable in high dimensions, and there is an extensive existing body of work addressing the issue of high dimensionality in the context of covariance estimation. See Pourahmadi [2011] for a survey of approaches to covariance estimation from the generalized linear modeling and regularization perspectives. However, much of this work addresses high dimensionality arising from functional or times series data sampled on a dense, regular grid. With such data, it is typical that the number of time points is larger than the number of observations. Few have addressed the challenges posed by sparse longitudinal data where measurement times may be almost unique yet sparsely distributed within the observed time range for each individual in the study. In this case, high dimensionality may not be a consequence of having more measurements per subject than the number of subjects themselves, but rather because when pooled across subjects, the total number of unique observed time points is greater than the number of individuals. Incomplete and unbalanced data arise when measurement schedules with targeted time points which are not necessarily equally spaced or if there is missing data. Sparse longitudinal data arise when the measurement schedule has arbitrary or almost unique time points for every individual. A given time point may have very few individuals with corresponding measurements.

We sidestep both issues of high dimensionality and irregularly sampled data by viewing the

response as a stochastic process having continuous covariance function. Recent work outlines the use of function estimation for smoothing elements of the covariance matrix, including Wu and Pourahmadi [2003],Huang et al. [2007]. To our knowledge, however, no previous work has applied smoothing to both dimensions of the Cholesky factor; we model the generalized autoregressive parameters using tensor product splines. Viewing covariance modeling as bivariate function estimation both accommodates irregularly sampled curved and permits interpolation and extrapolation of the covariance function between two measurements at any pair of time points within the time interval of interest rather than at observed pairs of time points only. The Cholesky decomposition enables covariance estimation through the estimation of a varying coefficient model. A transformation of the design point axes allows for an ANOVA-like decomposition of the coefficient function into two components, corresponding to the lag between time points and an additive component. Through this general framework, we can easily impose penalties on fitted functions to yield natural null models presented in the literature.

2 Cholesky Decomposition of Σ

To present a comprehensive overview our estimation procedure, we begin with the representation of the inverse covariance matrix, $\Omega = \Sigma^{-1}$, in terms of its Cholesky decomposition (see Pourahmadi [2007] for a detailed discussion.) In the section to follow, we will demonstrate that this parameterization of the precision matrix is particularly attractive due to both the computational advantages as well as the convenient modeling interpretation it permits. For any positive definite matrix Σ, there exists a unique unit lower triangular matrix T with diagonal entries equal to 1 which diagonalizes Σ:

$$T\Sigma T' = D$$

If we assume that the data having covariance matrix Σ follow an autoregressive model, then the entries of the Cholesky factor T and D enjoy a useful interpretation. Let $Y = (Y_1, Y_2, \ldots, Y_m)'$ be defined on a probability space with some probability measure $\mathcal{P}$ corresponding to the multivariate Normal distribution with mean 0 and covariance Σ, and let $Y_1, Y_2, \ldots, Y_m$ have associated measurement times

$$t_1 < t_2 < \cdots < t_m.$$

Let $\hat{y}_t$ denote the linear least-squares predictor of y_t based on its predecessors $y_1, \ldots, y_{t-1}$, and let e_t denote its prediction error with variance $Var\,(e_t) = \sigma_t^2$, $Cov\,(e_i, e_j) = 0$ for $i \neq j$. Standard regression theory gives us that we can find coefficients ϕ_{tj} such that

$$y_j = \sum_{k=1}^{j-1} \phi_{jk} y_k + \sigma_j e_j, \qquad j = 2, \ldots, m, . \tag{1}$$

Let T denote the $m \times m$ unit lower triangular matrix with elements

$$T_{jk} = \begin{cases} -\phi_{jk} & j > k \\ 1 & j = k \\ 0 & otherwise, \end{cases}$$

for $j, k = 1, \ldots, m$. In matrix notation, model 1 may then be written

$$e = TY, \tag{2}$$

Taking covariances on both sides of 2, we have

$$D = T\Sigma T' \tag{3}$$

An attractive feature of this reparameterisation is that, regardless of the modelling approach, the estimated covariance matrix is guaranteed to be positive definite. The unconstrained regression coefficients $\{\phi_{jk}\}$ are referred to as the *generalized autoregressive parameters* (GARPs). The $\{\sigma_j^2\}$ are called the *innovation variances* (IVs.) Unconstrained estimation of the $\{\sigma_k^2\}$ is achieved by log transformation; we leave these details for section 2. Expressing the precision matrix in terms of the GARPs and IVs, we have

$$\Omega = \Sigma^{-1} = T' D^{-1} T.$$

Rather than estimating a specific covariance matrix for data observed on a fixed, regular grid, we aim to estimate a smooth covariance function. This accomodates data which may consist of observations on multiple subjects measured at potentially unequally spaced and individual-specific times. In estimation of the means μ of p-vectors of i.i.d. variables, the Gaussian white noise model [9] is the appropriate infinite-dimensional model into which all objects of interest are embedded. In estimation of matrices, a natural analogue is the space B(l2,l2), which we write as B, of bounded linear operators from l2 to l2. These can be represented as matrices [cite *Regularized estimation of large covariance matrices by Bickel and Levina - section 4.*]

Rather than m-dimensional vectors, consider Y and e as the values of the stochastic processes $Y(t)$ and $e(t)$ at the set of observation times. We assume that $Y(t)$ is equipped with covariance function $G(s,t)$, and

$$e(s) \sim \mathcal{N}(0,1)$$

is a zero mean Gaussian process with unit variance. We assume that $G(s,t)$ satisfies some smoothness conditions, where smoothness is defined in terms of square integrability of certain derivatives. [TODO: clean up statement about smoothness of covariance function; integrability of covariance function of a stochastic process?] The entries of Σ, then, correspond to G evaluated at the distinct pairs of observed time points. Similarly, we treat the elements of the precision matrix Ω as the evaluation of some smooth function, $\omega(s,t)$.

Viewing the elements of D and the elements of the Cholesky factor T as the evaluation of smooth functions $\sigma(\cdot)$ and $\phi^*(\cdot,\cdot)$ leads us to the varying coefficient (VC) models first introduced in the seminal work of Hastie and Tibshirani. The procedures presented by Fan and Zhang [2000] and Huang et al. [2002] utilize varying coefficient models for modeling the mean of longitudinal data; parameterizing the covariance matrix according to 3 allows us to exploit these models in covariance estimation for such data as well. A generalization of traditional linear regression models,

varying coefficient models offer more flexibility than their static analogues by allowing the effect of covariates to change smoothly with the value other variables. Both regressors and response variables are assumed to vary according to an *indexing variable*, which permits interpolation of regressors and response variables at values of this indexing variable where there is either missing data of only a single observation and slope estimation is not feasible. Replacing $\{\phi_{jk}\}$ and $\{\sigma_j\}$ with smooth functions

$$\begin{array}{ll} \phi(t,s) & 0 \leq s < t \leq 1, \\ \sigma^2(t) & 0 \leq t \leq 1, \end{array}$$

the autoregressive model becomes

$$y(t_j) = \sum_{k=1}^{j-1} \phi(t_j, t_k)\, y(t_k) + \sigma(t_j)\, \epsilon(t_j) \quad j = 1, \ldots, m, \tag{4}$$

for $t_1 < t_2 < \cdots < t_m$.

We represent the varying coefficient function and the innovation variances using tensor product smoothing splines and penalized tensor product B-splines. By appending penalties to the negative log likelihood, we can seamlessly control the fit of ϕ and σ^2 to produce final covariance estimates exhibiting the desired null structure.Recasting the problem as the estimation of model 4 allows us access to the existing set of tools developed in the bivariate smoothing literature; our approach provides a flexible, comprehensive framework for covariance estimation.

3 Estimating ϕ via penalized maximum likelihood

We employ maximum likelihood for the estimation of the varying coefficient function $\phi(t,s)$ and the innovation variance function $\sigma(t)$, though neither the derivation the form of model 1 nor model 4 via the Cholesky decomposition rely on any assumptions about the distribution of Y.

For fixed $\left\{\sigma_j^2\right\}$, as a function of ϕ_{jk} the negative log-likelihood for a sample of N i.i.d. observations $Y_1, Y_2, \ldots, Y_N$ from a multivariate Gaussian distribution is proportional to the usual error sums of squares:

$$-2L\left(y_1, \ldots, y_N, \phi^* | \boldsymbol{\sigma^2}\right) \propto \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma(t_j)^{-2} \left(y_{ij} - \sum_{k=1}^{j-1} \phi(t_{ij}, t_{ik})\, y_{ik} \right)^2 \tag{5}$$

where

$$y_i = \left(y_{i1}, y_{i2}, \ldots, y_{i,m_i}\right), \quad i = 1, \ldots, N$$

denotes the vector of observations for subject i with corresponding measurement times

$$t_{i1} < t_{i2} < \cdots < t_{i,m_i}.$$

The form of the likelihood of $y_1, \ldots, y_N$ indicates that we allow both the number of measurements as well as the observation times to varying across subjects. The $\{t_{ij}\}$ need not be evenly-spaced

within or across individuals. Denote the innovation variance function evaluated at the vector of observed time points by $\boldsymbol{\sigma^2}$, and similarly let $\boldsymbol{\phi^*}$ denote the resulting vector when evaluating ϕ^* at the observed grid of time points, transformed to the l-m axis. Estimation of the varying coefficient function and the innovation variance function may be accomplished in an iterative fashion:

I. Fix $\boldsymbol{\sigma^2} = \boldsymbol{\sigma^2}_0$;

II. find $\widehat{\boldsymbol{\phi^*}} = \underset{\boldsymbol{\phi^*}}{\arg\max} - 2L\left(y_1, \ldots, y_N, \boldsymbol{\phi^*} | \boldsymbol{\sigma^2}_0\right)$.

III. Fix $\boldsymbol{\phi^*} = \widehat{\boldsymbol{\phi^*}}$;

IV. find $\widehat{\boldsymbol{\sigma^2}} = \underset{\boldsymbol{\sigma^2}}{\arg\max} - 2L\left(y_1, \ldots, y_N, \boldsymbol{\sigma^2} | \widehat{\boldsymbol{\phi^*}}\right)$.

V. Iterate until convergence.

For ease of exposition, we first focus our attention on the estimation of ϕ assume that $\sigma^2(t)$ is fixed and known. Estimation of the innovation variance function is presented in Section **??**. In the case that subjects share a common set of observation times $t_1 < \cdots < t_m$, it is well known that the MLE for Σ, $S = \sum_{i=1}^{N} y_i y_i'$ is highly unstable in high dimensions, a condition that is potentially worsened when one or more subjects has at least one observation time that is unique from the set of observation times common across subjects. To mitigate instability due to high dimensionality and simultaneously permit the estimation of $\phi(\cdot, \cdot)$ as a smooth bivariate function, we obtain a covariance estimator by applying bivariate smoothing of the elements of the Cholesky factor.

Estimating the varying coefficient function ϕ^*, however, is quite different from the usual problem of estimating an arbitrary bivariate function. In the case of the latter, we most typically treat both arguments equally in terms of regularization, but in the case of covariance estimation and the generalized coefficient function equal treatment of l and m in terms of penalization perhaps is not the most appropriate approach. The lag component, l, has particularly significant meaning in terms of the covariance function and thus also in terms of ϕ^* and is of considerable more interest than the orthogonal component, m. We parameterize ϕ in terms of the transformed domain:

$$l = t - s, \qquad m = \frac{1}{2}(s + t),$$

so that the following relationship holds:

$$\phi(s, t) = \phi^*\left(s - t, \frac{1}{2}(s + t)\right) = \phi^*(l, m)$$

with

$$\frac{l}{2} < m < 1 - \frac{l}{2}, \quad 0 < l < 1. \tag{6}$$

The likelihood can be written in terms of the reparameterized varying coefficient function:

$$\begin{aligned} -2L\left(y_1,\ldots,y_N,\phi^*|\boldsymbol{\sigma^2}\right) &= \sum_{i=1}^{n}\sum_{j=2}^{m_i}\sigma_{ij}^{-2}\left(y_{ij}-\sum_{k=1}^{j-1}\phi\left(t_{ij},t_{ik}\right)y_{ik}\right)^2 \\ &= \sum_{i=1}^{n}\sum_{j=2}^{m_i}\sigma_{ij}^{-2}\left(y_{ij}-\sum_{k=1}^{j-1}\phi^*\left(l_{ijk},m_{ijk}\right)y_{ik}\right)^2 \end{aligned} \tag{7}$$

We equip the l and m axes each with a B-spline basis to construct a basis for the varying coefficient function ϕ in **??** by taking the tensor product of the two marginal bases. Let

$$B_1\left(l\right),\ldots,B_K\left(l\right) \text{ and } B_1\left(m\right),\ldots,B_L\left(m\right)$$

denote the B-spline bases for l and m, each having a set of equally spaced knots along their respective domain. It is worth noting that while we have chosen not to distinguish between $\{B_k\}$ and $\{B_l\}$ for the sake of brevity, one is free to specify a different basis for each dimension either by using different order B-spline or, of course, using different numbers of knots, and hence entirely different knot sequences since P-splines rely on bases with equally spaced knots. The tensor product basis functions

$$T_{jk}\left(l,m\right)=B_j\left(l\right)B_k\left(m\right)$$

carve the l-m domain into rectangles. Figure 2 shows a thinned tensor product basis $\{T_{kl}\}$; a portion of the basis was omitted to eliminate overlapping of the basis functions so that the reader can identify individual tensor products. Each "hill" in Figure 2 is associated with an unknown coefficient θ_{ij} which determines the height of the hill. For a given knot grid, we can approximate a surface by

$$\phi^*\left(l,m\right)=\sum_{i=1}^{K}\sum_{j=1}^{L}\theta_{ij}B_i\left(l\right)B_j\left(m\right), \tag{8}$$

and the function evaluated at the observed (l_{ijk},m_{ijk}) may be written

$$\boldsymbol{\phi^*}=B_m\Theta B_l'$$

where Θ denotes the $K\times L$ matrix of tensor product coefficients, with elements θ_{ij}.

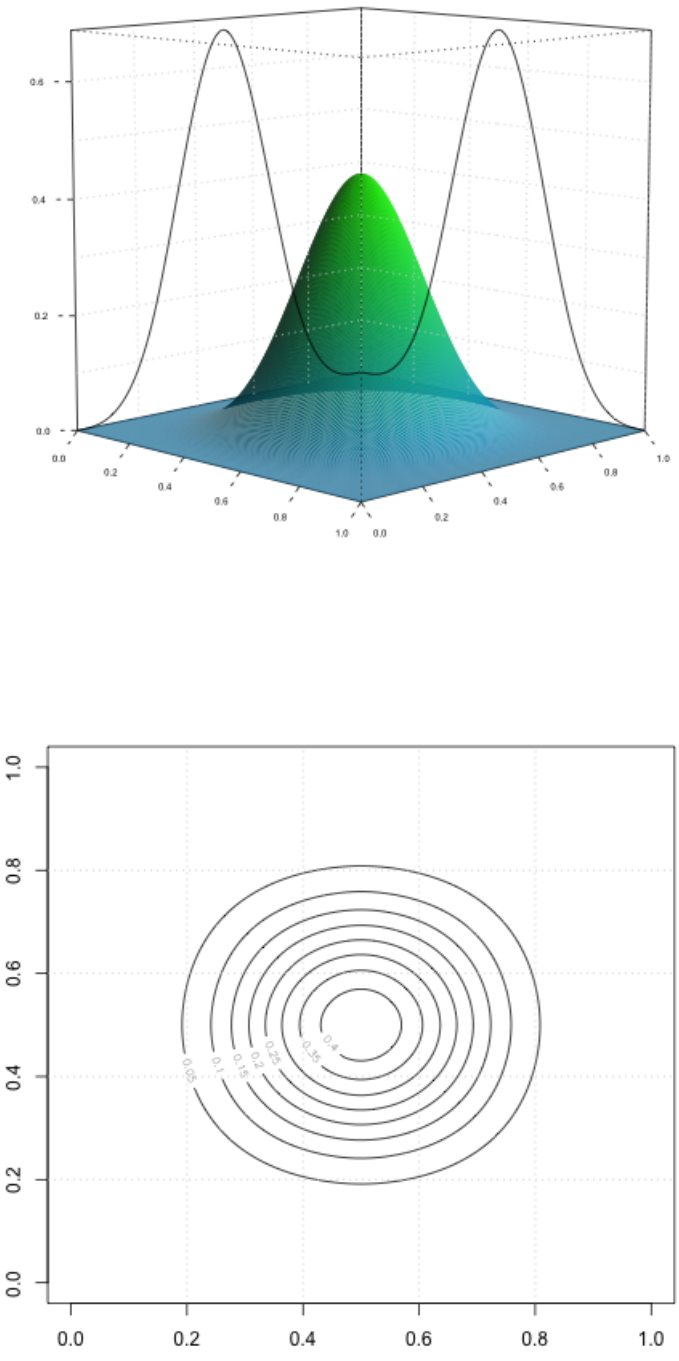


Figure 1: Tensor product of two cubic B-splines

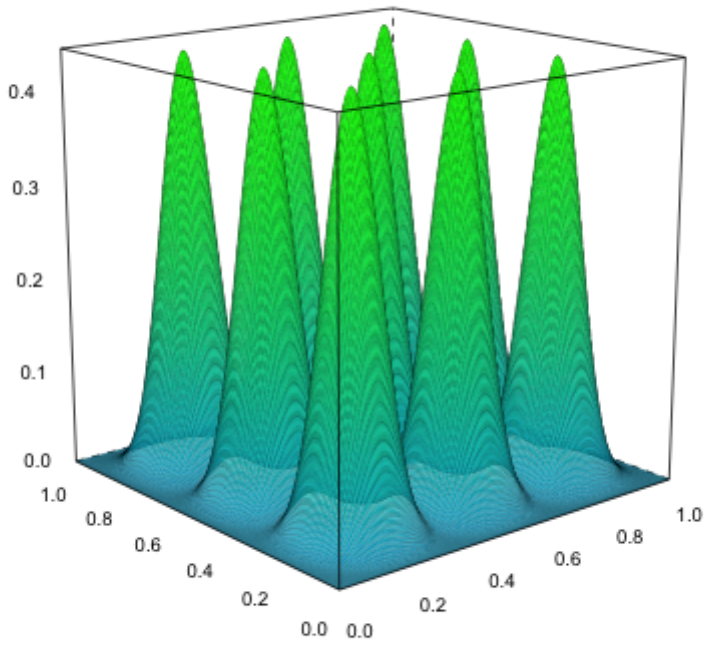


Figure 2: A subset of a full bivariate basis of cubic B-splines

3.1 Regularization with difference penalties

The minimizer of **??** honors the fidelity to the data, so to balance the complexity of the fitted function with the goodness of fit to the data, we can append a penalty to the negative log likelihood to control the fitted function. By using rich B-spline bases for l and m alongside discrete difference penalties on the spline coefficients, we can achieve as much smoothness of the fitted function in both the l and m dimensions as desired. O'Sullivan [1986] was the first to propose using a rich B-spline basis and using a penalty to restrict the flexibility of the fitted curve, like Wahba [1990] applying a penalty to the second derivative of the fitted curve:

$$J = \int_0^1 \left[f''(l)\right]^2 \, dx.$$

For a B-spline of the form

$$f(x) = \sum_{j=1}^{n} \theta_i B_j(x),$$

one can derive a banded matrix P using the properties of B-splines such that

$$J = \theta' P \theta$$

where $\theta = (\theta_1, \ldots, \theta_n)$. The i-j^{th} element of P is given by

$$p_{ij} = \int_0^1 B_i'' (x) \, B_j'' (x) \; dx.$$

In some applications, it is useful to work with third and fourth order differences, since for large values of λ, the fitted curve approaches a parametric polynomial model. This may be of particular interest in the context of estimating the elements of the Cholesky factor, as many have proposed simple parametric functions of lag only for ϕ^*, such as low order polynomials. See Pourahmadi [1999]. However, with the use of higher order derivatives, the computation of P is nontrivial and becomes very tedious. Eilers and Marx [1996] were the first to propose P-splines, or *penalized B-splines*, as an approach to nonparametric regression. P-splines circumvent complexity associated with constructing such penalty matrices by omitting derivatives and integrals altogether, replacing them with finite differences and sums.

Instead, flexibility of the fitted function is controlled by using a discrete penalty matrix based on finite difference formulas. Smoothness of the fitted function is achieved by first using a rich B-spline basis with equally spaced knots to purposefully overfit the smooth coefficient vectors; this eliminates the difficulty of choosing the optimal set of knots. Then by attaching a difference penalty to the goodness of fit measure, one may prevent overfitting and make a potentially ill-conditioned fitting procedure a well-conditioned one. The finite difference penalty is simple to compute and can be handled mechanically for any order of the differences. Since it is easily introduced into regression equations, it is feasible to evaluate the impact of different orders of the differences on the fitted model. Using the properties of B-splines, it is straightforward to show that the difference penalty of order d is a good discrete approximation to the integrated square of the d^{th} derivative, so little is lost by replacing the derivative-based penalty with

$$J_d (f) = \sum_{j=d}^{n} \left(\Delta^d \theta_j \right)^2 \tag{9}$$

where $\theta = (\theta_1, \ldots, \theta_n)$. Let D_d denote the matrix difference operator: $D_d \theta = \Delta^d \theta$, where

$$\Delta \theta_j = \theta_j - \theta_{j-1}, \quad \Delta^2 \theta_j = \Delta \left(\Delta \theta_j \right) = \theta_j - 2\theta_{j-1} + \theta_{j-2}$$

In general,

$$\Delta^d \theta_j = \Delta \left(\Delta^{d-1} \theta_j \right).$$

Then, 9 can be written in terms of the squared norm of the difference operator applied to the vector of B-spline coefficients:

$$\begin{aligned} J_d (f) &= ||D_d \theta||^2 \\ &= \theta' P_d \theta \end{aligned} \tag{10}$$

where $P_d = D_d' D_d$. To examine the connection between the second-derivative penalty to the penalty on second-order differences of the B-spline coefficients, we only need to employ straightforward calculus and exploit the recursive property of the B-spline basis functions:

$$\int_0^1 \left[f''(x)\right]^2 \, dx = \int_0^1 \left\{ \sum_{j=1}^{n} \theta_j B''_{j,3}(l) \right\}^2 \, dl.$$

The derivative properties of B-splines permits this to be written as

$$\int_0^1 \left[f''(x)\right]^2 \, dx = \int_0^1 \left[\sum_{j=1}^{n} \sum_{k=1}^{n} \Delta^2\theta_j \Delta^2\theta_k B_{j,1}(l) \, B_{k,1}(l) \right] dl.$$

Most of the cross products of $B_{j,1}(x)$ and $B_{k,1}(x)$ vanish since B-splines of degree 1 only overlap when j is $k-1$, k, or $k+1$. Thus, we have that

$$\begin{aligned}
\int_0^1 \left[f''(x)\right]^2 \, dx &= \int_0^1 \left[\left\{ \sum_{j=1}^{n} \Delta^2\theta_j B_j(l,1) \right\}^2 + 2 \sum_{j} \Delta^2\theta_j \Delta^2\theta_{j-1} B_j(l,1) \, B_{j-1}(l,1) \right] dl \\
&= \sum_{j=1}^{n} \left(\Delta^2\theta_j\right)^2 \int_0^1 B_j^2(l,1) \, dl \\
&\qquad + 2 \sum_{j=1}^{n} \Delta^2\theta_j \Delta^2\theta_{j-1} \int_0^1 B_j(l,1) \, B_{j-1}(l,1) \, dl
\end{aligned} \tag{11}$$

which can be written as

$$\int_0^1 \left[f''(x)\right]^2 \, dx = c_1 \sum_{j=2}^{n} \left(\Delta^2\theta_j\right)^2 + c_2 \sum_{j=3}^{n} \Delta^2\theta_j \Delta^2\theta_{j-1} \tag{12}$$

Given a set of equidistant knots, the constants c_1 and c_2 are given by

$$\begin{aligned}
c_1 &= \int_0^1 B_{j,1}^2(x) \, dx \\
c_2 &= \int_0^1 B_{j,1}(x) \, B_{j-1,1}(x) \, dx.
\end{aligned} \tag{13}$$

This gives us that the traditional smoothness penalty on the squared second derivative can be written as a linear combination of a penalty on the second-order differences of the B-spline coefficients 9 and the sum of the cross products of neighboring second differences. The second term in 12 leads to a complex objective function when minimizing the penalized likelihood, where seven adjacent spline coefficients occur, as opposed to five if only the first term in 12 is used in the penalty. The added complexity is a consequence of overlapping B-splines, which quickly increases when using higher order differences and higher order B-splines. We can seamlessly augment the

likelihood with the difference penalty to achieve smooth fitted functions without the complexity posed by the derivative-based penalty.

A smoother sequence of coefficients leads to a smoother curve, as illustrated in Figure **??**. The relationship between P-spline curves and their coefficients is easily characterized if we consider the coefficients as the skeleton of the function, and draping the B-splines over them puts the flesh on the bones. As long as the coefficient sequence is smooth, the number of basis functions (and coefficients) is unimportant since the penalty ensures the smoothness of the skeleton and that the fitting procedure is well-conditioned. Figure **??** illustrates this utility of the penalty for simulated data; there are $m = 10$ observations and 60 cubic B-splines. This property of P-splines cannot be overly appreciated because it frees us from the concern of choosing the optimal set of knots. Unless computational constraints are of concern, which is possible with large models, it is prudent to use even more B-splines. Figure **??** shows how the fitted function changes as the tuning parameter varies when the data are sparsely sampled. P-splines enjoy a number of additional advantageous properties, many of which are inherited from the attractive properties of B-splines. See Eilers and Marx [1996] for a detailed presentation.

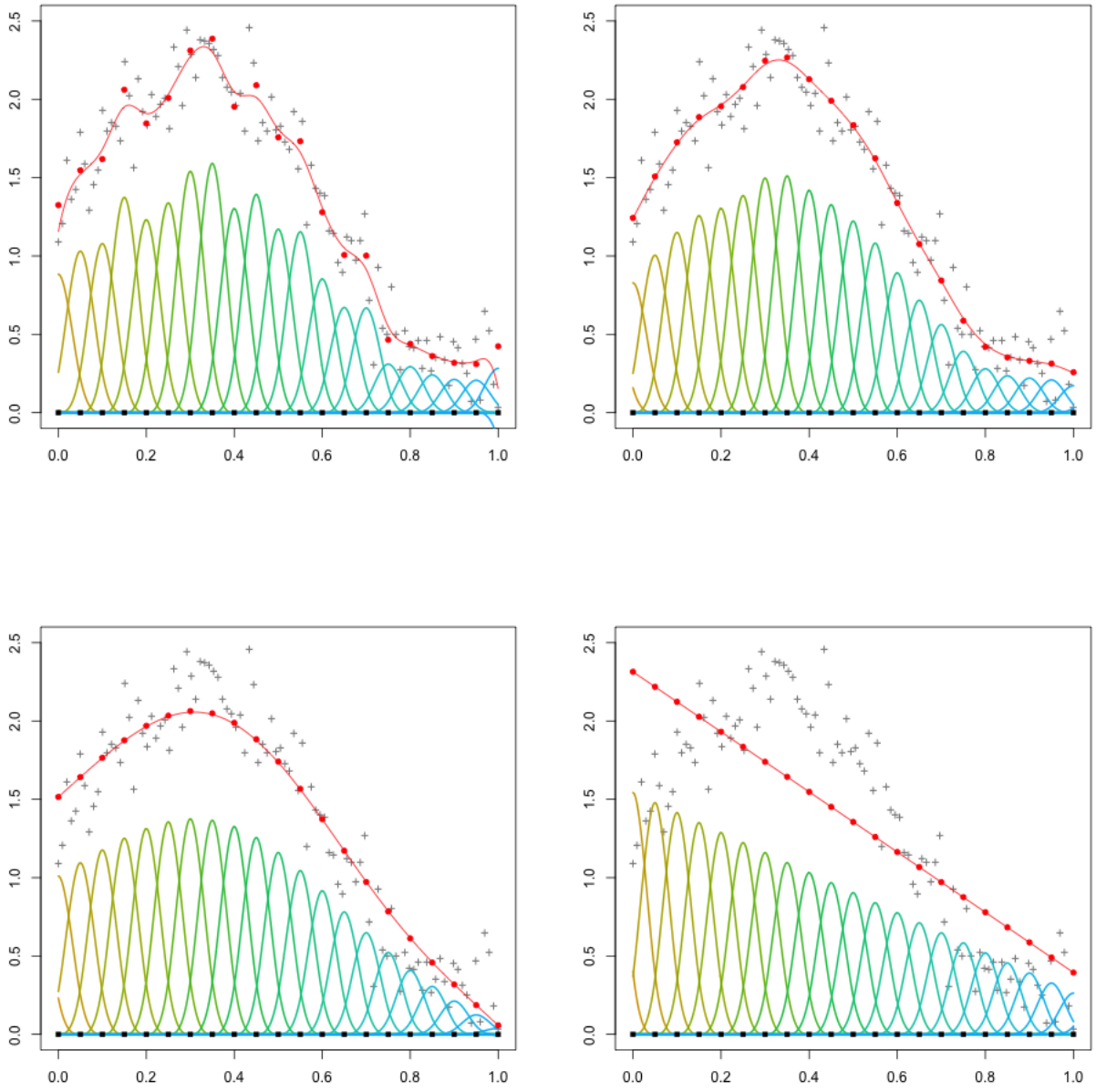
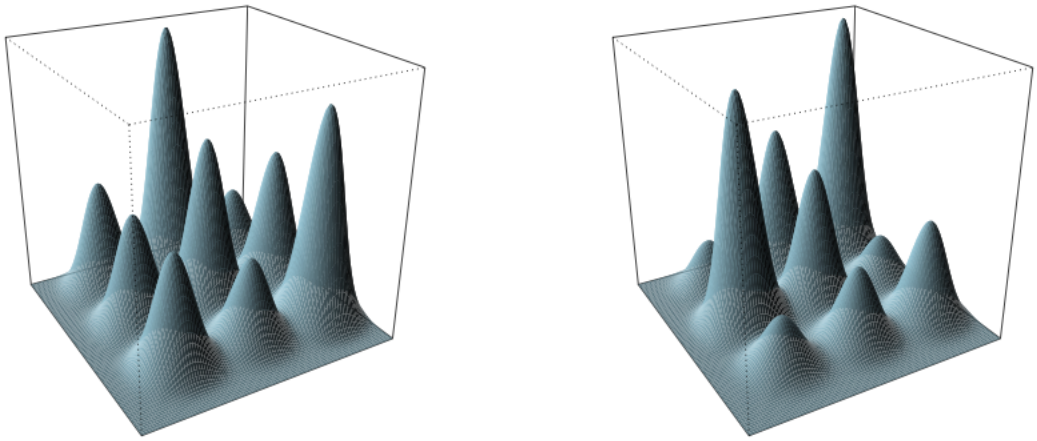


Figure 3: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The fitted curve in the upper left plot is the most "wiggly" of any of the fits, as the penalty plays the weakest roll in the fitted coefficients there. The red circles are the values of each of the B-spline coefficients; as the penalty increases, they form as smoother sequence as we move across the four plots, which results in a smoother fitted function. As the penalty parameter approaches infinity, the fit approaches a linear function as shown in the bottom right plot.*

To extend these results to the bivariate setting for regularizing of ϕ^*, the only modification to the differencing procedure in one dimension necessary is the addition of a second difference penalty, one for each variable l and m. We append the pair of penalties to the negative log likelihood **??** and take the estimator of ϕ^* to correspond to the B-spline coefficients minimizing

$$-2L + J(\phi^*) = \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma(t_j)^{-2}\left\{ y_{ij} - \sum_{k=1}^{j-1}\left(\sum_{r=1}^{L}\sum_{s=1}^{K}\theta_{rs}B_r(l_{ijk})\,B_s(m_{ijk})\right) y_{ik}\right\}^2 + \lambda_l \sum_{r'=1}^{K} |D_{d_l}\theta_{r'\cdot}|^2 + \lambda_m \sum_{s'=1}^{L} |D_{d_m}\theta_{\cdot s'}|^2. \tag{14}$$

where $\theta_{k\cdot}$ and $\theta_{\cdot l}$ denote the k^{th} row and l^{th} column of Θ, respectively. The first term in**??** imposes a difference penalty of order d_l on the rows of the coefficient matrix while the second term places a difference penalty (of possible different order d_m) on the columns. We give each direction its own smoothing parameter to permit anisotropic smoothing; however, one could opt to use a single smoothing parameter for both directions and dodge the added work of optimizing the amount of smoothing with two separate parameters. Figure **??** shows a potential result of heavy column penalization (left) and heavy row penalization (right) under a second order difference penalty on each row and each column for large values of λ_l and λ_m. The figure demonstrates that the limiting behaviour of each row and column is linear, but the resulting surface may exhibit slope reversals from one row (column) to the next.



(a) *heavy column penalty* (b) *heavy row penalty*

Figure 4: *Nine cubic B-spline tensor products with heavy linear column penalty and heavy linear row penalty*

We take the estimator of ϕ^* to be the minimizer of It is computationally advantageous to express the coefficient matrix in "unfolded" notation, which allows us to express the varying coefficient function at the observed coordinate grid as in the usual multiple regression form:

$$\text{vec}\left\{\phi^*\left(l,m\right)\right\} = B\theta$$

Stacking the columns of Θ gives the vectorized coefficient matrix $\theta = \text{vec}\left(\Theta\right)$. The $p \times KL$ tensor product basis B is constructed from the tensor product of the marginal B-spline bases defined in Eilers et al. [2006] as the *row-wise Kronecker product* of the individual bases:

$$B = B_l \square B_m = \left(B_m \otimes 1_K'\right) \odot \left(1_L' \otimes B_l\right). \tag{15}$$

The operator $\odot$ denotes the element-wise matrix product; 1_K (1_L) denotes the column vector of ones having length K (L.) The operations in 15 construct B such that the i^{th} row of $B_m \square B_l$ is the Kronecker product of the corresponding rows of B_m and B_l. The penalty in **??** can also be compactly expressed:

$$\lambda_l ||P_l\theta||^2 + \lambda_m ||P_m\theta||^2$$

where $P_l = I_L \otimes D_{d_l}' D_{d_l}$ and $P_m = D_{d_m}' D_{d_m} \otimes I_K$. We define the matrix W of historical regressors so that 14 can be written in matrix form as

$$-2L + J\left(\phi^*\right) = \left(Y - WB\theta\right)' D^{-1}\left(Y - WB\theta\right) + \lambda_l ||P_l\theta||^2 + \lambda_m ||P_m\theta||^2, \tag{16}$$

with $\hat{\theta}$ solving the system of equations

$$\left[\left(WB\right)' D^{-1} WB + \lambda_l P_l + \lambda_m P_m\right]\theta = \left(WB\right)' D^{-1} Y \tag{17}$$

From 17, we note that the system of equations depends on basis coefficients remains fixed at KL, even as the number of observations increases. The grid of regression coefficients can be recovered by arranging the elements of $\hat{\theta}$ into a matrix of L columns having length K.

This recipe for constructing a tensor product basis for ϕ^* is an easy and convenient way to construct a two-dimensional basis for a bivariate function with domain corresponding to the unit square. However, the domain of the autoregressive coefficient function, specified in 6, lies on the lower triangle of the unit square:

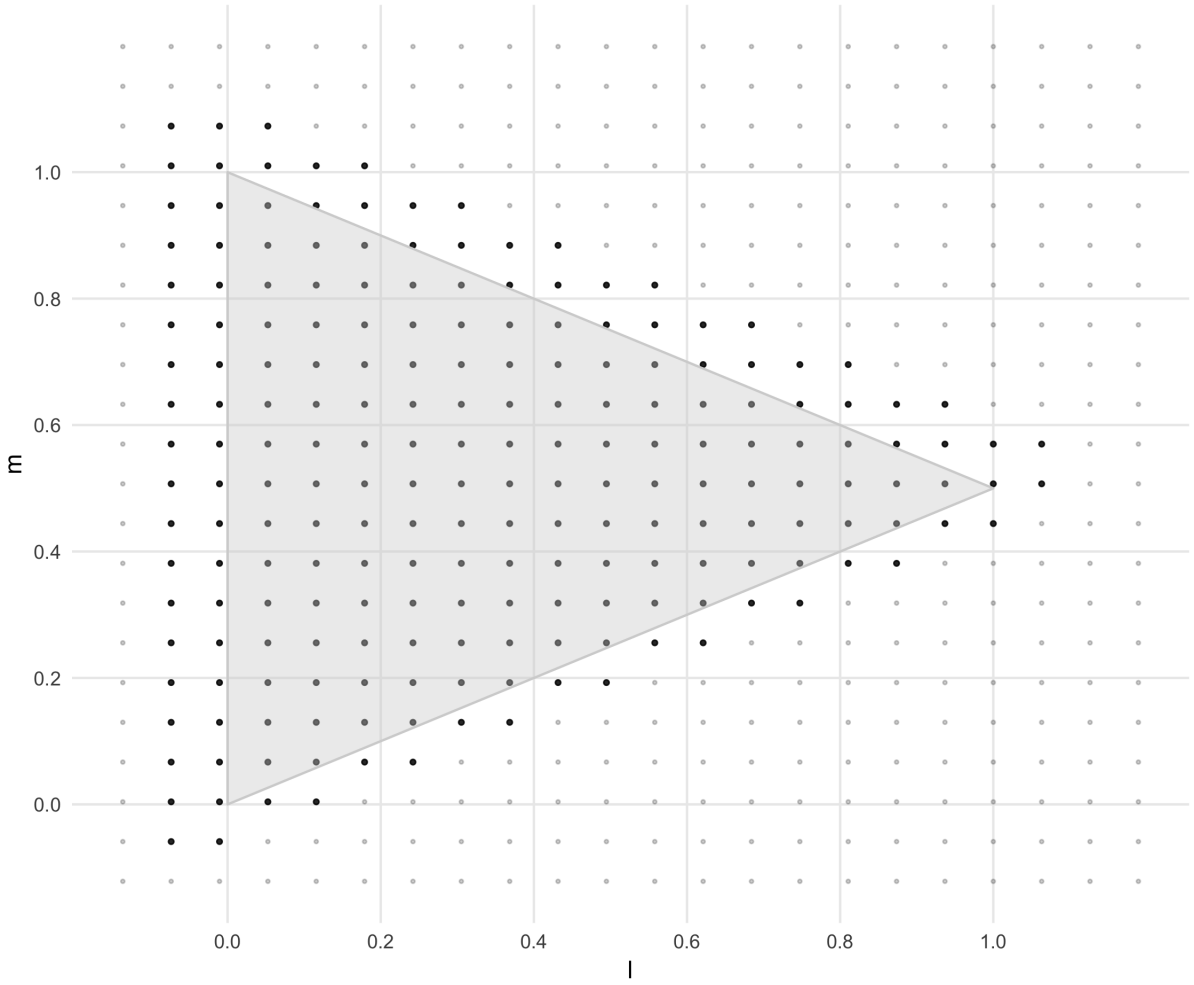


Figure 5: $\frac{l}{2} < m < 1 - \frac{l}{2}, \quad 0 < l < 1.$

The triangular domain of ϕ^* cannot be modeled by the tensor product basis as constructed due to singularities resulting from the large number of basis functions anchored at knots near which we have no data, and hence little information about the corresponding basis coefficient. Much work in computer graphics has been done proposing methods of smoothing over arbitrary function domains, which are approximated by triangulations. See Dahmen et al. [1992] and Seidel [1991] for details. These methods are, however, quite computationally intensive; to correct for the instability in the smoothed surface, we can simply remove the knots corresponding to tensor products functions which do not overlap with the function domain from the basis, B, and trimming the penalty matrices P_l and P_m as needed. With the trimmed basis and penalties, we can carry out optimization as previously discussed.

4 Model selection and tuning parameter estimation

4.1 The limiting behaviour of H_λ

The inspection of the hat matrix

$$H_\lambda = WB\left(WB'WB + \lambda_l P_l + \lambda_m P_m\right)^{-1}(WB)' D^{-1}.$$

and its properties are integral for assessing model complexity and selecting the optimal values of the tuning parameters λ_l and λ_m. Summarizing the complexity of a fitted P-spline is far from a trivial task; one must simultaneously consider the value of the smoothing parameter, the number of basis functions in the B-spline basis, as well as the order of the difference penalties. We follow Eilers and Marx [1996] andMarx and Eilers [2005] assess model complexity as discussed in citehastie1990generalized, who proposed to use the trace of the smoother matrix as an approximation to the effective dimensions of linear smoother. The *effective dimension* is easily obtained and combines the effect of all three of these elements:

$$\begin{aligned} \text{ED} &= \text{tr}\left[H_\lambda\right] \\ &= \text{tr}\left[\left[WB\,(WB)' D^{-1} WB + \lambda_l P_l + \lambda_m P_m\right]^{-1}(WB)' D^{-1}\right] \end{aligned} \tag{18}$$

When the number of basis functions is significantly smaller than the sample size, it is computationally advantageous to use the cyclic property of the trace:

$$\text{tr}\left[\left[(WB)' D^{-1} WB + \lambda_l P_l + \lambda_m P_m\right]^{-1}(WB)' D^{-1} WB\right],$$

which requires computing the trace of a $KL \times KL$ matrix. The effective dimension approaches $d_l + d_m$, the order of the differencing operator, as λ increases, where d_l and d_m denote the orders of the difference penalties in the l and m directions, respectively. Let

$$Q = (WB)' D^{-1} WB \qquad \text{and} \qquad Q_\lambda = P.$$

Using properties of the matrix trace, we can write

$$\begin{aligned} \text{tr}\left(H_\lambda\right) &= \text{tr}\left[\left(Q + Q_\lambda\right)^{-1} Q\right] \\ &= \text{tr}\left[Q^{1/2}\left(Q + Q_\lambda\right)^{-1} Q^{1/2}\right] \\ &= \text{tr}\left[\left(I + Q^{-1/2} Q_\lambda Q^{-1/2}\right)^{-1}\right] \end{aligned}$$

Define $L \equiv Q^{-1/2} Q_\lambda Q^{-1/2}$. Then

$$\text{tr}\left(H_\lambda\right) = \text{tr}\left[\left(I + \lambda L\right)^{-1}\right] = \sum_{j=1}^{n} \frac{1}{1 + \lambda\gamma_j}$$

where γ_j, $j = 1, \ldots, n$ are the eigenvalues of L. Q_λ has exactly $d_l + d_m$ eigenvalues equal to zero. Hence, L has $d_l + d_m$ zero eigenvalues. For large λ, only the $d_l + d_m$ terms with $\gamma_j = 0$ contribute to the sum which gives the trace of H, so that

$$\lim_{\lambda \to \infty} \text{tr}\,(H) = d_l + d_m.$$

An alternative approach to model selection is to minimize the information criterion (IC):

$$\text{IC}\,(\lambda) = -2\left(y, \hat{\beta}_\lambda\right) + c\ \text{dim}\left(\hat{\beta}_\lambda\right)$$

Special cases are when $c = 2$, yielding the Akaike information criterion (AIC) and when $c = \log(m)$, yielding the Bayesian information criterion (BIC). The IC assesses the quality of a model by adjusting the log likelihood of the data under the fitted model to account for some measure of model complexity. In the classical linear model setting, the degrees of freedom serves as the measure of model complexity, is used for obtaining an unbiased estimate of the error variance, and thus, is necessary for comparing the performance of different models. In the case of the normal likelihood, AIC becomes

$$\text{AIC}\,(\lambda) = \tag{19}$$

One may use CV to select the optimal value of λ, and then using the corresponding residuals gives a natural choice to use as an estimate of σ^2 for the computation of AIC (λ). It is practical to work with modified versions of CV (λ) and GCV (λ), with values that can be interpreted as estimates of the cross-validation standard deviation:

$$\begin{aligned} \bar{\text{CV}}\,(\lambda) &= \sqrt{\text{CV}\,(\lambda)} \\ \bar{\text{GCV}}\,(\lambda) &= \sqrt{m\text{GCV}\,(\lambda)} \end{aligned} \tag{20}$$

Alternatively, one might note that an (approximately) unbiased estimate of the error variance is given by the sum of squared residuals divided by the error degrees of freedom, $m - \text{ED}$.

5 *Simulations:*

5.1 $\Sigma = 0.3^2 I$

$N = 30$**,** $M = 20$**,** $d_l = d_m = 0$

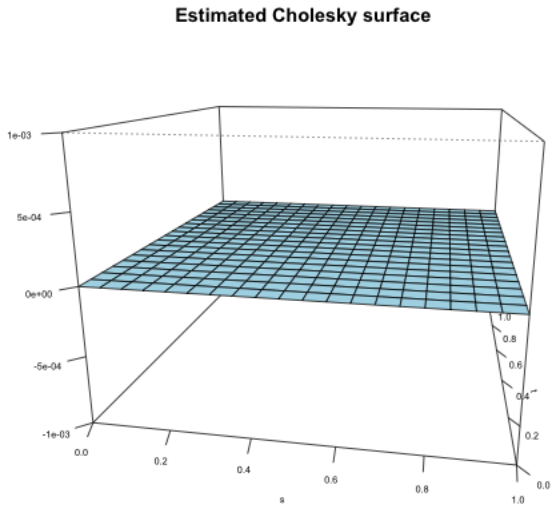
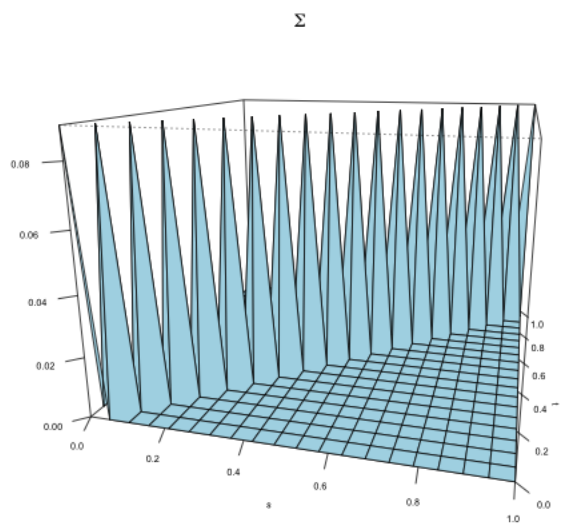
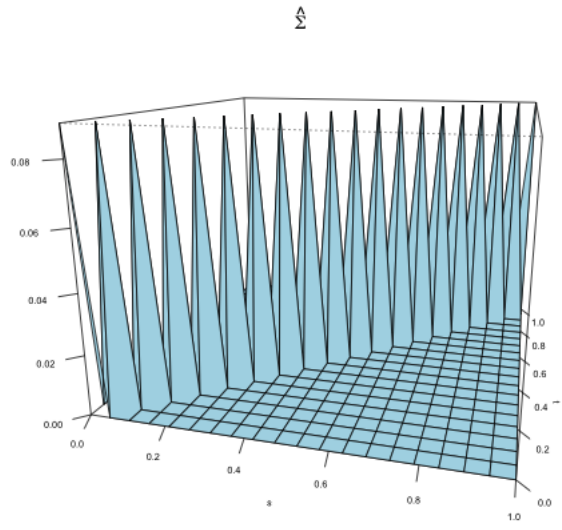


Figure 6: Estimated T

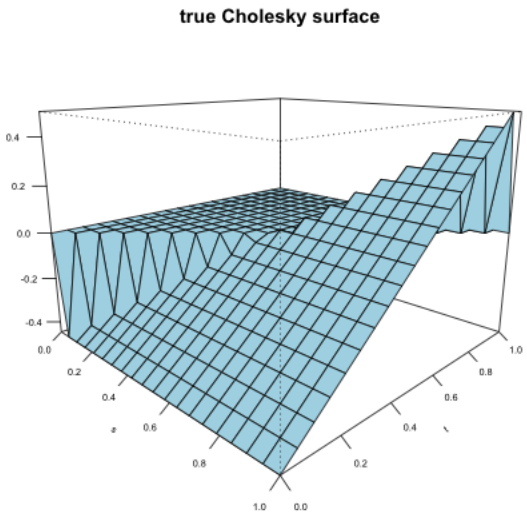


(a) True Σ

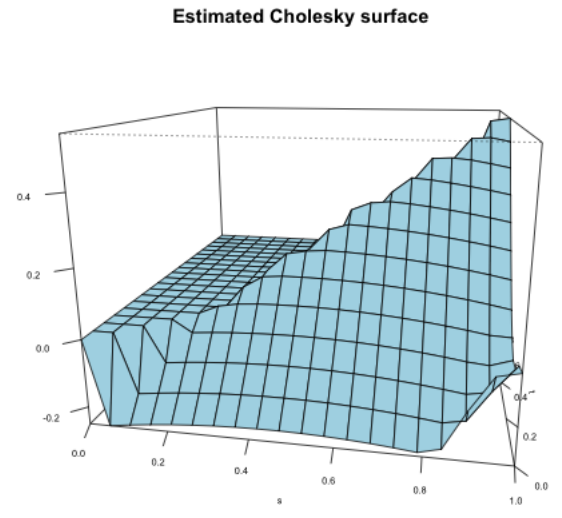


(b) $\hat{\Omega}^{-1}$

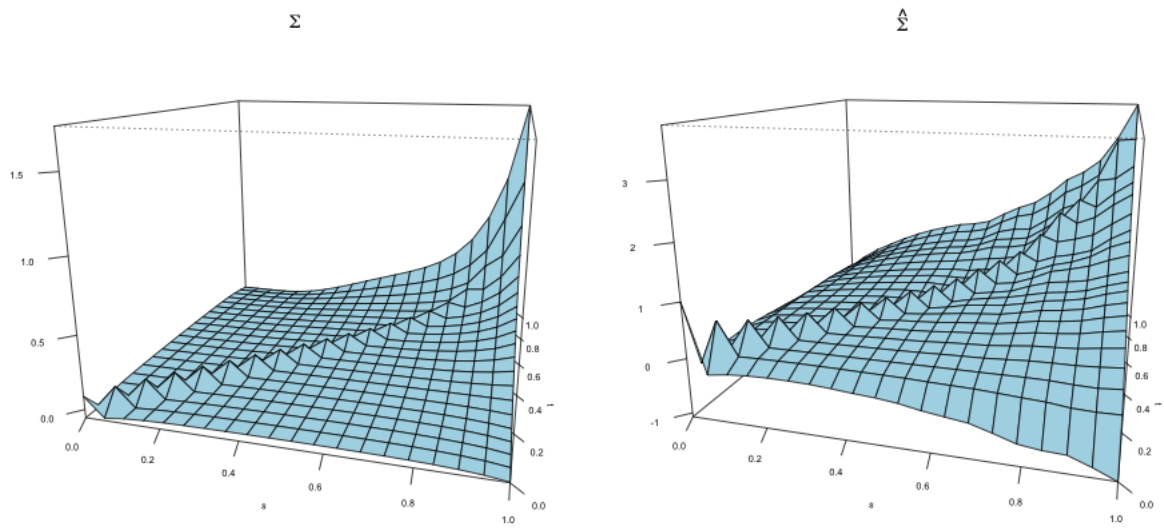
5.2 $\phi(s,t) = s - \frac{1}{2}$**,** $\sigma^2 = 0.3^2$
$N = 30$**,** $M = 20$**,** $d_l = d_m = 2$



(a) True T

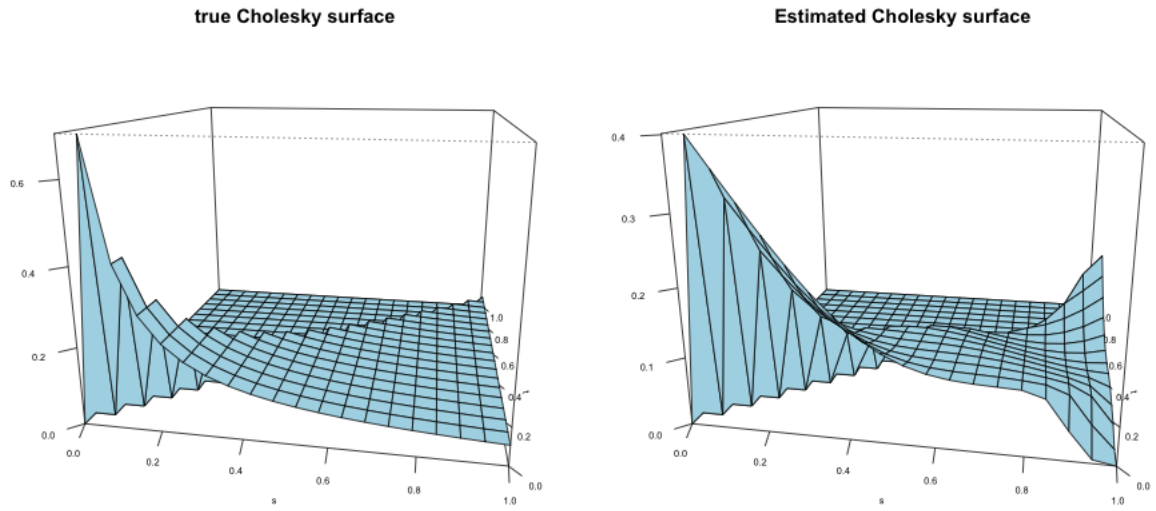


(b) Estimated T



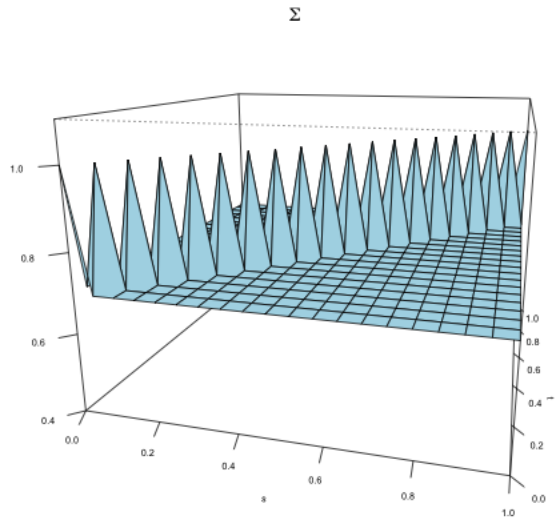
(a) Σ (b) $\hat{\Omega}^{-1}$

5.3 $\Sigma = 0.7J + 0.3I$
$N = 30$**,** $M = 20$**,** $d_l = 2$**,** $d_m = 1$

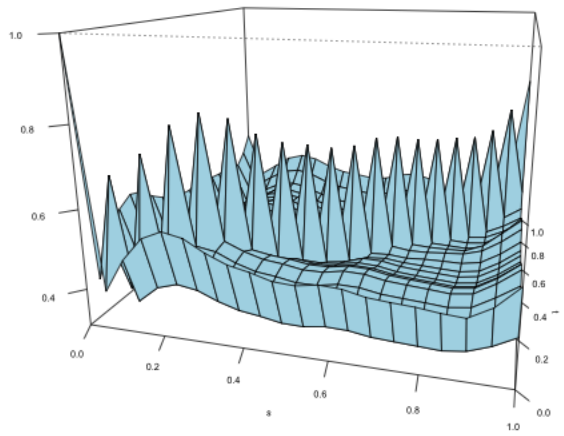


(a) True T

(b) Estimated T



(a) Σ



(b) $\hat{\Omega}^{-1}$

5.4 Banding the Cholesky factor

A recent upsurge in the application of regularization to large empirical covariance matrices includes the work Huang et al. [2006], Furrer and Bengtsson [2007], Fan et al. [2008], Ledoit and

Wolf [2004]. Furrer and Bengtsson [2007] propose stabilizing the sample covariance matrix by "tapering," or gradually shrinking the off-diagonal elements to zero. d'Aspremont et al. [2008] propose a sparse estimator by applying an L_1 penalty directly to the elements of the covariance matrix. Instead of regularizing the covariance matrix itself, some have opted to regularize its inverse; Wu and Pourahmadi [2003] band the Cholesky decomposition by setting certain diagonals of the Cholesky factor to zero. Huang et al. [2006] and Levina et al. [2008] use L_1 penalties to achieve parsimony of the entries of the Cholesky factor; sparsity of the Cholesky factor, however, does not necessarily imply sparsity in the inverse covariance matrix. Levina et al. [2008] propose banding the Cholesky factor using a nested Lasso penalty which yields sparse estimators of the precision matrix.

Pourahmadi [1999] presented a hueristic argument that the GARPs, $\phi_{t,t-l}$ should be monotonically decreasing in absolute value as l increases. The following proposition establishes the relationship between zeros in the Cholesky factor and zeroes in the inverse covariance, connecting the regularization of ϕ^* to the structure in the resulting inverse covariance matrix.

Proposition 5.1. *Let Ω denote a $m \times m$ positive definite matrix with elements ω_{ij} with modified Cholesky decomposition $T'D^{-1}T$, where T is unit lower triangular. Let t_{ij} denote the ij^{th} element of T. For any column j and row $r(j) > j$, $\omega_{mj} = \cdots = \omega_{r(j),j} = 0$ if and only if $t_{mj} = \cdots = t_{r(j),j} = 0$.*

The proof is left to the appendix. Proposition 5.1 maintains that the modified Cholesky factor T with arbitrary column band lengths corresponds to inverse covariance matrix Ω with the same column band lengths, and hence the inverse covariance matrix is K-banded if and only if its Cholesky factor is K-banded. That is, if ϕ^* is zero or nearly zero for large $|i-j|$, then y_i and y_j are conditionally uncorrelated (or nearly so). This suggests that the effect of y_{t-l} on y_t through model 4 should decrease as the time between the two measurements increases, such that

$$\phi^*(l, m) \approx 0$$

for large l is a reasonable way to institute regularization.

This notion of sparsity is quite different from the notion of smoothness imposed by the usual derivative-based penalties or the analog difference penalties for P-splines; however, the local support of the B-spline basis functions permits the intuitive expression of a penalty to be used for banding the Cholesky factor. The B-spline basis functions are non-negative on their support; therefore, if a B-spline (as in **??**) is zero for $l > l_0$, then the coefficients of the B-splines contributing to that region of the domain are also zero. For functions represented using order k B-spline basis functions $\{B_i\}$, $i = 1, \ldots, n$, one may truncate the function to zero at some truncation point $l_0 \in (0, 1)$ by penalizing the size of the coefficients corresponding to any basis functions having support on $l > l_0$. This naturally leads to penalties of the form

$$J_B = \sum_{i=i^*-k}^{n} |\theta_i|^p \tag{21}$$

where

$$i^* = \max_i \{\ell_i : \ell_i \leq l_0\}$$

is the index of the largest knot which is smaller than the truncation point l_0. Setting $p = 1$ in 21 corresponds to putting a LASSO penalty (see Tibshirani [1996]) on the coefficients contributing to the function value on $l > l_0$, while setting $p = 2$ corresponds to the usual ridge regression setting.

Several have proposed regularization of the inverse covariance matrix by "banding" the Cholesky factor T: setting all elements of T beyond the K^{th} off-diagonal to zero, i.e. setting $\phi_{t,t-l} = 0$ for $l > K$ for some choice of K. (See Pourahmadi [1999], Wu and Pourahmadi [2003], Bickel and Levina [2008], and Huang et al. [2007].) In terms of model 1, this is equivalent to regressing y_t on only its K immediate predecessors, setting the regression coefficient for y_{t-l} to zero for $l > K$. This notion is instrumental in justifying a family of penalties which induces an alternative decomposition of the function space for which the null space of these penalties comprises functions ϕ^* taking nonzero values on a subset of the domain:

$$\mathcal{H}_0(\mathcal{B}) = \{\phi^* : \phi^*(l, m) = 0 \text{ for all } l > l_0\}, \tag{22}$$

which is equivalent to the set of functions having compact support:

$$supp(\phi^*) = (0, l_0].$$

6 *Recap: from the top*

- We propose a general framework for unconstrained covariance estimation.
- Flexibility permits imposing various types of regularization with ease.
- Penalty specification is crucial for performance.

7 What's next?

- "designer" penalties - impose desirable shape constraints (decay in l)
- Additive, ANOVA models for ϕ^*
- P-spline model reparameterized as mixture models

References

P. J. Bickel and E. Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227, 2008.

W. Dahmen, C. A. Micchelli, and H.-P. Seidel. Blossoming begets ??-spline bases built better by ??-patches. *Mathematics of computation*, 59(199):97–115, 1992.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

P. H. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

P. H. Eilers, I. D. Currie, and M. Durbán. Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50(1):61–76, 2006.

J. Fan and J.-T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62 (2):303–322, 2000.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

J. Z. Huang, C. O. Wu, and L. Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128, 2002.

J. Z. Huang, N. Liu, M. Pourahmadi, and L. Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, pages 85–98, 2006.

J. Z. Huang, L. Liu, and N. Liu. Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Computational and Graphical Statistics*, 16(1): 189–209, 2007.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

E. Levina, A. Rothman, and J. Zhu. Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263, 2008.

B. D. Marx and P. H. Eilers. Multidimensional penalized signal regression. *Technometrics*, 47(1): 13–22, 2005.

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518, 1986.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

M. Pourahmadi. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters. *Biometrika*, pages 1006–1013, 2007.

M. Pourahmadi. Covariance estimation: The glm and regularization perspectives. *Statistical Science*, pages 369–387, 2011.

H.-P. Seidel. Symmetric recursive algorithms for surfaces: B-patches and the de boor algorithm for polynomials over triangles. *Constr. Approx*, 7:257–279, 1991.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

G. Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

W. B. Wu and M. Pourahmadi. Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844, 2003.