List of Tables

1 Construction of the tensor product cubic spline subspace from marginal subspaces $\mathcal{H}_{[1]}, \mathcal{H}_{[2]}$ 9
2 Tensor product cubic spline subspace reproducing kernels and inner products . . . 10

Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake[*] Yoonkyung Lee[†]

March 9, 2018

1 Smoothing Spline Varying-coefficient Models for Covariance Estimation

If we consider the Cholesky decomposition of Σ within such functional context, it is natural to extent the same notion to the elements of T and D. We view the GARPs $\{\phi_{tj}\}$ and innovation variances as the evaluation of the smooth functions $\tilde{\phi}\left(t,s\right)$ and $\sigma^2\left(t\right)$ at observed time points, which we assume are drawn from some distribution having compact domain $\mathcal{T}$. Without loss of generality, we take $\mathcal{T} = [0,1]$. Henceforth, we view $\tilde{\phi}$ and σ^2 as a smooth continuous functions, but for ease of exposition, we let $\tilde{\phi}_{ij}$ denote the varying coefficient function evalutated at (t_i, t_j):

$$\tilde{\phi}_{tj} = \tilde{\phi}\left(t_i, t_j\right).$$

Adopting similar notation for the innovation variance function, denote

$$\sigma_j^2 = \sigma^2\left(t_j\right),$$

where $0 \le t_j < t_i \le 1$ for $j < i$. This leads to varying coefficient model

$$y\left(t_i\right) = \sum_{j=1}^{i-1} \tilde{\phi}\left(t_i, t_j\right) y\left(t_j\right) + \sigma\left(t_j\right) \epsilon\left(t_j\right) \qquad i = 1, \ldots, M, \tag{1}$$

Our goal is now to estimate the above model, utilizing bivariate smoothing to estimate $\tilde{\phi}\left(t,s\right)$ for $0 \le s < t \le 1$, and one-dimensional smoothing to estimate $\sigma\left(t\right)$, $0 \le t \le 1$. Our proposed method for covariance estimation defines a flexible, general framework which makes all of the existing techniques for penalized regression accessible for the seemingly far different task of estimating a covariance matrix.

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

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

Our approach to estimation is constructed to provide a fully data-driven methodology for selecting the optimal covariance model (given some optimization criterion) from a expansive class of estimators ranging in complexity from that of the previously aforementioned parametric models to that of completely unstructured estimators, like the sample covariance matrix. We leverage the collection of regularization techniques that are accessible in the usual function estimation setting. By properly specifying the roughness penalty, our optimization procedure results in null models which correspond to the parametric and semiparametric models for ϕ and σ^2 discussed in **??**. To facilitate the penalty specification that achieves this, we consider modeling the varying coefficient function which takes inputs

$$\begin{aligned} l &= t - s \\ m &= \frac{t+s}{2}, \end{aligned} \tag{2}$$

where l is the continuous analogue of the usual "lag" between time points t and s, and m is simply its orthogonal direction. We have discussed many parsimonious covariance structures which model $y\left(t\right)$ as a stationary process with covariance function which depends on time points t_i and t_j only through the Euclidean distance $||t_i - t_j||$ between them. Covariance functions taking the form $Cov\left(y\left(t_i\right), y\left(t_j\right)\right) = G\left(t_i, t_j\right) = G\left(||t_i - t_j||\right)$ can then be written as

$$Cov\left(y\left(t_i\right), y\left(t_j\right)\right) = G\left(l_{ij}\right)$$

where $l_{ij} = |t_i - t_j|$. Regularizing the functional components of the Cholesky decomposition so that functions incurring large penalty correspond to functions which vary in only l and are constant in m allows us to model nonstationarity in a fully data-driven way. Our goal is to estimate

$$\phi\left(l, m\right) = \phi\left(s - t, \frac{1}{2}\left(s+t\right)\right) = \tilde{\phi}\left(t, s\right). \tag{3}$$

While our framework allows for estimation of the autoregressive coefficient function and the innovation variance function via any nonparametric regression setup, we focus on two primary approaches for representing ϕ and σ. First, we assume that ϕ belongs to a reproducing kernel Hilbert space, $\mathcal{H}$ and employ the smoothing spline methods of Kimeldorf and Wahba (see Kimeldorf and Wahba [1971] and Wahba [1990] for comprehensive presentation.) To enhance the statistical interpretability of model parameters, we decompose ϕ into functional components similar to the notion of the main effect and the interaction terms in classical analysis of variance. We adopt the smoothing spline analogue of the classical ANOVA model proposed by Gu Gu [2013], and estimation is achieved through similar computational strategies.

1.1 Penalized maxiumum likelihood estimation of ϕ, $\log \sigma^2$

Let random vector Y follow a multivariate normal distribution with zero mean vector and covariance Σ. The loglikelihood function $\ell\left(Y, \Sigma\right)$ satisfies

$$-2\ell\left(Y,\Sigma\right)=\log\left|\Sigma\right|+Y'\Sigma Y \tag{4}$$

Using $T\Sigma T'=D$, we can write

$$\left|\Sigma\right|=\left|D\right|=\prod_{i=1}^{m}\sigma_i^2$$

and

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

Writing 4 in terms of the prediction errors and their variances of the non-redundant entries of (T,D), we have

$$\begin{aligned}-2\ell\left(Y,\Sigma\right)&=\log\left|D\right|+Y'T'D^{-1}TY\\&=\sum_{i=1}^{m}\log\sigma_i^2+\sum_{i=1}^{m}\frac{\epsilon_i^2}{\sigma_i^2},\end{aligned} \tag{5}$$

where

$$\epsilon_i=\begin{cases}y\left(t_1\right), & i=1,\\ y\left(t_i\right)-\sum_{j=1}^{i-1}\phi\left(\boldsymbol{v}_{ij}\right)y_j, & i=2,\ldots,M,\end{cases} \tag{6}$$

where $\phi\left(\boldsymbol{v}_{ij}\right)=\phi\left(l_{ij},m_{ij}\right)=\tilde{\phi}\left(t_i,t_j\right)$. Accommodating subject-specific sample sizes and measurement times merely requires appending an additional index to observation times. Let $Y_1,\ldots,Y_N$ denote a sample of N independent mean zero random trajectories from a multivariate normal distribution with common covariance Σ. We associate with each trajectory $Y_i=\left(y_{i1},\ldots,y_{i,m_i}\right)'$ with a vector of potentially subject-specific observation times $\left(t_{i1},\ldots,t_{i,m_i}\right)'$, so that the j^{th} measurement of trajectory i is modeled

$$\begin{aligned}y\left(t_{ij}\right)&=\sum_{k=1}^{j-1}\tilde{\phi}\left(t_{ij},t_{ik}\right)y\left(t_{ik}\right)+\sigma\left(t_{ij}\right)\epsilon\left(t_{ij}\right)\\&=\sum_{k=1}^{j-1}\phi\left(\boldsymbol{v}_{ijk}\right)y\left(t_{ik}\right)+\sigma\left(t_{ij}\right)\epsilon\left(t_{ij}\right)\end{aligned} \tag{7}$$

for $i=1,\ldots,N$, $j=2,\ldots,m_i$. Making similar ammendments to indexing, the joint log likelihood for the sample $Y_1,\ldots,Y_N$ is given by

$$-2\ell\left(Y_1,\ldots,Y_N,\phi,\sigma^2\right)=\sum_{i=1}^{N}\sum_{j=1}^{m_i}\log\sigma_{ij}^2+\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\epsilon_{ij}^2}{\sigma_{ij}^2}, \tag{8}$$

With this, we can estimate ϕ and $\log\sigma^2$ using maximum likelihood or any of its penalized variants by appending a roughness penalty (penalties) to 8. Employing regularization, we take ϕ, σ^2 to minimize

$$-2\ell\left(Y_{1},\ldots,Y_{N},\phi,\sigma^{2}\right)+\lambda J\left(\phi\right)+\breve{\lambda}\breve{J}\left(\sigma^{2}\right), \tag{9}$$

where J and $\breve{J}$ are roughness penalties on ϕ and σ^2, and λ, $\breve{\lambda}$ are non-negative smoothing parameters. To jointly estimate the GARP function and the IV function, we adopt an iterative approach in the spirit of Huang et al. [2006], Huang et al. [2007], and Pourahmadi [2000]. A procedure for minimizing 8 starts with initializing $\left\{\sigma_{ij}^{2}\right\}=1$ for $i=1,\ldots,N$, $j=1,\ldots,m_i$. For fixed σ^2, the penalized likelihood (as a function of ϕ) is given by

$$-2\ell_{\phi}+\lambda J\left(\phi\right)=\sum_{i=1}^{N}\sum_{j=2}^{m_i}\sigma_{ij}^{-2}\left(y_{ij}-\sum_{k<j}\phi\left(\boldsymbol{v}_{ijk}\right)y_{ik}\right)^{2}+\lambda J\left(\phi\right), \tag{10}$$

which corresponds to the usual penalized least squares functional encountered in the nonparametric function estimation literature. The first term, the residual sums of squares, encourages the fitted function's fidelity to the data. The second term penalizes the roughness of ϕ, and λ is a smoothing parameter which controls the tradeoff between the two conflicting concerns. Given ϕ^* the minimizer of 10 and setting $\phi=\phi^*$, we update our estimate of σ^2 by minimizing

$$-2\ell_{\sigma^2}+\breve{\lambda}\breve{J}\left(\sigma^{2}\right)=\sum_{i=1}^{N}\sum_{j=2}^{m_i}\log\sigma_{ij}^{2}+\sum_{i=1}^{N}\sum_{j=1}^{m_i}\sigma_{ij}^{-2}r_{ij}^{*\,2}+\breve{\lambda}\breve{J}\left(\sigma^{2}\right), \tag{11}$$

where the $\left\{r_{ij}^{*\,2}=\left(y_{ij}-\sum_{k<j}\phi^{*}\left(\boldsymbol{v}_{ijk}\right)y_{ik}\right)\right\}$ denote the working residuals based on the current estimate of ϕ. This process of iteratively updating ϕ^* and σ^{2^*} is repeated until convergence is achieved.

The remainder of the chapter is reserved for presenting two functional representations of (ϕ,σ^2). The first leverages the rich theoretical foundation of reproducing kernel Hilbert space techniques for function estimation. This framework has been studied extensively for the problem of estimating a function nonparametrically (see Aronszajn [1950], Wahba [1990], and Berlinet and Thomas-Agnan [2011] for detailed examinations), but to our knowledge has received little attention in the context of covariance models. We use a smoothing spline ANOVA decomposition of the varying coefficient function ϕ to construct a flexible class of covariance models while simultaneously maintaining interpretability. The second approach is based on the penalized B-splines, or P-splines, of Eilers and Marx [1996]; these models exhibit many of the attractive numerical properties of the basis functions on which they are built. The formulation of the penalty is independent of the basis, which provides added modeling flexibility due to the ease with which one can employ various types of regularization.

1.2 Smoothing spline representation of ϕ, σ

1.2.1 An RKHS framework for estimating ϕ

This section presents a method for regularized estimation of the varying coefficient function ϕ using a reproducing kernel Hilbert space (RKHS) framework. To do so, we first must establish some notation and review the relevant mathematical details of the surrounding framework. A Hilbert space $\mathcal{H}$ of functions on a set $\mathcal{V}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is defined as a complete inner product linear space. A Hilbert space is called a reproducing kernel Hilbert space if the evaluation functional $[\boldsymbol{v}]\, f = f\left(\boldsymbol{v}\right)$ is continuous in $\mathcal{H}$ for all $\boldsymbol{v} \in \mathcal{V}$. The Reisz Representation Theorem gives that there exists $Q \in \mathcal{H}$, the representer of the evaluation functional $[\boldsymbol{v}]\left(\cdot\right)$, such that $\langle Q\boldsymbol{v}, \phi \rangle_{\mathcal{H}} = \phi\left(\boldsymbol{v}\right)$ for all $\phi \in \mathcal{H}$. See Gu [2013] Theorem 2.2.

The symmetric, bivariate function $Q\left(\boldsymbol{v}_1, \boldsymbol{v}_2\right) = Q\boldsymbol{v}_2\left(\boldsymbol{v}_1\right) = \langle Q\boldsymbol{v}_1, Q\boldsymbol{v}_2 \rangle_{\mathcal{H}}$ is called the reproducing kernel (RK) of $\mathcal{H}$. The RK satisfies that for every $v \in \mathcal{V}$ and $f \in \mathcal{H}$,

I. $Q\left(\cdot, \boldsymbol{v}\right) \in \mathcal{H}$

II. $f\left(\boldsymbol{v}\right) = \langle f, Q\left(\cdot, v\right) \rangle_{\mathcal{H}}$

The first property is called the reproducing property of Q. Every reproducing kernel uniquely determines the RKHS, and in turn, every RKHS has unique reproducing kernel. See Gu [2013], Theorem 2.3. The kernel satisfies that for any $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n_1}\}$, $\{\breve{\boldsymbol{v}}_1, \ldots, \breve{\boldsymbol{v}}_{n_2}\} \in \mathcal{V}$ and $\{a_1, \ldots, a_{n_1}\}$, $\left\{a_1, \ldots, a'_{n_2}\right\} \in \Re$,

$$\langle \sum_{i=1}^{n_1} a_i Q\left(\cdot, \boldsymbol{v}_i\right), \sum_{j=1}^{n_2} a'_j Q\left(\cdot, \breve{\boldsymbol{v}}_j\right) \rangle_{\mathcal{H}}. \tag{12}$$

Let $\mathcal{N}_J = \{\phi : \; J\left(\phi\right) = 0\}$ denote the null space of J, and consider the decomposition

$$\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J.$$

The space $\mathcal{H}_J$ is a RKHS having $J\left(\phi\right)$ as the squared norm. The minimizer of **??** has form

$$\phi\left(\boldsymbol{v}\right) = \sum_{\nu=1}^{d_0} d_\nu \eta_\nu\left(\boldsymbol{v}\right) + \sum_{i=1}^{n} c_i Q\left(\boldsymbol{v}_i, \boldsymbol{v}\right), \tag{13}$$

where $\{\eta_\nu\}$ is a basis for $\mathcal{N}_J$, and Q_J is the RK in $\mathcal{H}_J$.

The objective function **??** can be rewritten in terms of the squared norm with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}}$:

$$-2\ell_\phi + \lambda J\left(\phi\right) = \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} \left(L_{ijk}\phi\right) y_{ik} \right)^2 + \lambda ||P_J \phi||^2 \tag{14}$$

where P_J is the projection operator which projects ϕ onto the subspace $\mathcal{H}_J$, and L_{ijk} denotes the evalutation functional $[v_{ijk}]\, \phi$. Let ξ_{ijk} denote the representer of L_{ijk}; Kimeldorf and Wahba [1971] established that the minimizer of 14 has form

$$\phi\left(\boldsymbol{v}\right) = \sum_{\nu=1}^{m} d_\nu \eta_\nu\left(v\right) + \sum_{i=1}^{|V|} c_i \left(P_J \xi_i\right) \tag{15}$$

where $V = \bigcup_{i,j,k} \boldsymbol{v}_{ijk}$, and $\{\eta_1, \ldots, \eta_m\}$ span $\mathcal{H}_0$, the null space of P_J. To show this, we start by noting that any $\phi \in \mathcal{H}$ can be written

$$\phi(\boldsymbol{v}) = \sum_{\nu=1}^{m} d_\nu \eta_\nu(v) + \sum_{i=1}^{|V|} c_i (P_J \xi_i) + \rho(\boldsymbol{v}) \tag{16}$$

where $\rho \perp \mathcal{H}_0,\ \text{span}\{(P_1\xi_j)\}_{j=1}^{|V|}$. To establish that the solution has form 15 requires showing that the minimizer of 14 has $\rho = 0$. The proof entails demonstrating that ρ does not improve the residual sums of squares and only adds to the penalty term, $J(\phi)$. Details are similar to those in the proof provided in Wahba [1990] and are left to the appendix **??**.

Convenient construction of a reproducing kernel Hilbert space on a domain

$$\mathcal{V} = \mathcal{V}_1 \otimes \mathcal{V}_2$$

which can be written as a product domain, is available through the tensor product of the RKHS for each of the marginal domains $\mathcal{V}_1$ and $\mathcal{V}_2$. Without loss of generality, we can let $l,\ m \in [0,1] = \mathcal{V}_1 = \mathcal{V}_2$. Given Hilbert space for the domain of l, $\mathcal{H}_{[1]}$ with reproducing kernel Q_1 and Hilbert space on the domain of m, $\mathcal{H}_{[2]}$ with reproducing kernel Q_2, the reproducing kernel $Q = Q_{[1]}Q_{[2]}$ corresponds to that of the tensor product space of $\mathcal{H}_{[1]}$ and $\mathcal{H}_{[2]}$, denoted

$$\mathcal{H} = \mathcal{H}_{[1]} \otimes \mathcal{H}_{[2]}.$$

See **?**, Theorem 2.6. Let $\mathcal{A}_1$, $\mathcal{A}_2$ denote the averaging operators defining ANOVA decompositions on $\mathcal{H}_{[1]}$, $\mathcal{H}_{[2]}$, respectively, where $\mathcal{H}_{0[i]}$ has RK $Q_{0[i]}$, $i = 1, 2$ and $\mathcal{H}_{1[i]}$ has RK $Q_{1[i]}$ satisfying $\mathcal{A}_1 Q_{[1]}(l, \cdot) = \mathcal{A}_2 Q_{[2]}(m, \cdot) = 0$. Then the tensor product space $\mathcal{H}$ has tensor sum decomposition

$$\begin{aligned} \mathcal{H} &= [\mathcal{H}_{0[1]} \oplus \mathcal{H}_{1[1]}] \otimes [\mathcal{H}_{0[2]} \oplus \mathcal{H}_{1[2]}] \\ &= [\mathcal{H}_{0[1]} \otimes \mathcal{H}_{0[2]}] \oplus [\mathcal{H}_{0[1]} \otimes \mathcal{H}_{1[2]}] \oplus [\mathcal{H}_{1[1]} \otimes \mathcal{H}_{0[2]}] \oplus [\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}] \end{aligned} \tag{17}$$

If $Q_{0[i]} \propto 1$ for $i = 1, 2$, then $\mathcal{H}$ can be further simplified:

$$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2, \tag{18}$$

which has reproducing kernel $Q = Q_{[1]}Q_{[2]}$.

Example 1.1. Tensor product cubic spline
Let the marginal domains of l and m correspond to $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, where

$$\mathcal{H}_i = \mathcal{C}^{(m_i)} = \left\{ \phi : \int_0^1 \phi^{(m_i)}\, dv < \infty \right\},$$

which are equipped with inner product

$$\begin{aligned}\langle f,g\rangle &= \langle f,g\rangle_0 + \langle f,g\rangle_1 \\ &= \sum_{\nu=0}^{m_i-1} M_\nu f M_\nu g + \int_0^1 f^{(m_i)}(v)\, g^{(m_i)}(v)\, dv, \quad i=1,2\end{aligned} \tag{19}$$

where the order i differential operator M_ν is defined $M_\nu \phi = \int_0^1 \phi^{(m)}(v)\, dv$, $\nu = 1,\ldots,m_i$, $i=1,2$. Denote the norm corresponding to this inner product by

$$||f||^2 = \langle f,f\rangle = \langle f,f\rangle_0 + \langle f,f\rangle_1 = ||P_0 f||^2 + ||P_1 f||^2$$

The reproducing kernel Q can be expressed in terms of the scaled Bernoulli polynomials $\left\{k_j(v) = \frac{1}{j!}B_j(v)\right\}$, $v \in [0,1]$, where B_j is defined according to:

$$\begin{aligned} B_0(x) &= 1 \\ \frac{d}{dx}B_j(x) &= jB_{j-1}(x),\ j=1,2,\ldots\end{aligned}$$

One can verify that $\int\limits_0^1 k_\mu^\nu dv = \delta_{\mu,\nu}$ for $\nu,\mu = 0,\ldots,m_i-1$, where $\delta_{\mu,\nu}$ is the Kronecker delta. This implies that the k_ν, $\nu=0,\ldots,m_i-1$ for an orthonormal basis for $\mathcal{H}_{0[i]} = \left\{\phi : \phi^{(m_i)} = 0\right\}$ under the inner product

$$\langle f,g\rangle_0 = \sum_{\nu=0}^{m_i-1} M_\nu f M_\nu g, \quad i=1,2,$$

and that

$$Q_{0[i]}(v,v') = \sum_{\nu=0}^{m_i-1} k_\nu(v)\, k_\nu(v')$$

is the reproducing kernel for $\mathcal{H}_{0[i]}$. The subspaces of $\mathcal{H}_{[i]}$ which are orthogonal to $\mathcal{H}_{0[i]}$ are comprised of functions ϕ satisfying

$$\mathcal{H}_{1[i]} = \{\phi : M_\nu f = 0,\ \ \nu = 0,1,\ldots,m_i-1, \int\limits_0^1 \phi^{(m_i)}\, dv < \infty\}, \quad i=1,2.$$

One can show that the representer for the evaluation functional $[v]\,\phi$ in $\mathcal{H}_{1[i]}$ with squared norm $\langle f,g\rangle_1 = \int_0^1 f^{(m_i)} g^{(m_i)}\, dv$ is given by the function

$$Q_{[i]}{}'_v(v) = k_{m_i}(v)\, k_{m_i}(v') + (-1)^{m_i-1} k_{2m_i}(v'-v) \tag{20}$$

See **?** Example 2.3.3 for proof. The tensor product smoothing spline results from letting $m_1 = m_2 = 2$, so that the marginal subspaces can be written

$$\{\phi : \phi'' \in \mathcal{L}_2\,[0,1]\} = \{\phi : \phi \propto 1\} \oplus \{\phi : \phi \propto k_1\} \oplus \left\{\phi : \int_0^1 \phi dv = \int_0^1 \phi' dv = 0,\ \phi'' \in \mathcal{L}_2\,[0,1]\right\} \tag{21}$$

$$= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1, \tag{22}$$

where $\mathcal{H}_{01} \oplus \mathcal{H}_1$ forms the contrast in a one-way ANOVA decomposition with averaging operator $\mathcal{A}\phi = \int_0^1 \phi\, dv$. The corresponding reproducing kernels are

$$Q_{00}\,(v, v') = 1 \tag{23}$$

$$Q_{01}\,(v, v') = k_1\,(v)\,k_1\,(v') \tag{24}$$

$$Q_1\,(v, v') = k_2\,(v)\,k_2\,(v') - k_4\,(v - v')\,. \tag{25}$$

The tensor product space can be constructed with nine tensor sum terms; the construction of the tensor product space from the terms of the tensor sum. The corresponding reproducing kernels and inner products are given in Table 1 and Table 2, respectively.

	$\mathcal{H}_{00[2]}$	$\mathcal{H}_{01[2]}$	$\mathcal{H}_{1[2]}$
$\mathcal{H}_{00[1]}$	$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{00[2]}$	$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{01[2]}$	$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{1[2]}$
$\mathcal{H}_{01[1]}$	$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{00[2]}$	$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{01[2]}$	$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{1[2]}$
$\mathcal{H}_{1[1]}$	$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{00[2]}$	$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{01[2]}$	$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}$

Table 1: Construction of the tensor product cubic spline subspace from marginal subspaces $\mathcal{H}_{[1]}$, $\mathcal{H}_{[2]}$

Table 2: Tensor product cubic spline subspace reproducing kernels and inner products

Subspace	Reproducing kernel	Inner product
$\mathcal{H}_{00[1]} \otimes \mathcal{H}_{00[2]}$	1	$\left(\int_0^1 \int_0^1 f\right) \left(\int_0^1 \int_0^1 g\right)$
$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{00[2]}$	$k_1\left(l\right) k_1\left(l'\right)$	$\left(\int_0^1 \int_0^1 f'_{[1]}\right) \left(\int_0^1 \int_0^1 g'_{[1]}\right)$
$\mathcal{H}_{01[1]} \otimes \mathcal{H}_{01[2]}$	$k_1\left(l\right) k_1\left(l'\right) k_1\left(m\right) k_1\left(m'\right)$	$\left(\int_0^1 \int_0^1 f''_{[12]}\right) \left(\int_0^1 \int_0^1 g''_{[12]}\right)$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{00[2]}$	$k_2\left(l\right) k_2\left(l'\right) - k_4\left(l - l'\right)$	$\int_0^1 \left(\int_0^1 f''_{[12]}\, dl'\right) \left(\int_0^1 g''_{[12]}\, dl'\right)\, dl$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{01[2]}$	$\left[k_2\left(l\right) k_2\left(l'\right) - k_4\left(l - l'\right)\right] k_1\left(m\right) k_1\left(m'\right)$	$\int_0^1 \left(\int_0^1 f^{(3)}_{[112]}\, dl'\right) \left(\int_0^1 g^{(3)}_{[112]}\, dl'\right)\, dl$
$\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}$	$\left[k_2\left(l\right) k_2\left(l'\right) - k_4\left(l - l'\right)\right] \left[k_2\left(m\right) k_2\left(m'\right) - k_4\left(m - m'\right)\right]$	$\int_0^1 \int_0^1 f^{(4)}_{[1122]} g^{(4)}_{[1122]}$

For $\boldsymbol{v} \in V$ where V is a product domain, ANOVA decompositions can be characterized by

$$\mathcal{H} = \bigoplus_{\beta=0}^{g} \mathcal{H}_{\beta} \tag{26}$$

and

$$J(\phi) = \sum_{\beta=0}^{g} \theta_{\beta}^{-1} J_{\beta}(\phi_{\beta}), \tag{27}$$

where $\phi_{\beta} \in \mathcal{H}_{\beta}$, J_{β} is the square norm in $\mathcal{H}_{\beta}$, and $0 < \theta_{\beta} < \infty$. This gives

$$\begin{aligned} \mathcal{H}_0 &= \mathcal{N}_J \\ \mathcal{H}_J &= \bigoplus_{\beta=1}^{g} \mathcal{H}_{\beta}, \text{ and} \\ Q &= \sum_{\beta=1}^{g} \theta_{\beta} Q_{\beta}, \end{aligned}$$

where Q_{β} is the RK in $\mathcal{H}_{\beta}$. The $\{\theta_{\beta}\}$ are additional smoothing parameters, which are implicit in notation to follow for the sake of concise demonstration.

Let Y denote the vector of length $n_y = \sum_i M_i - N$ constructed by stacking the N observed response vectors $Y_1, \ldots, Y_N$ less their first element y_{i1} one on top of each other:

$$\begin{aligned} Y &= (Y_1', Y_2', \ldots, Y_N')' \\ &= \left(y_{12}, y_{13}, \ldots, y_{1,m_1}, \ldots, y_{N2}, y_{N3}, \ldots, y_{N,m_N}\right)' \end{aligned}$$

Define X_i to be the $m_i \times |V|$ matrix containing the covariates necessary for regressing each measurement $y_{i2}, \ldots, y_{i,m_i}$ on its predecessors as in model 7, and stack these on top of one another to obtain

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \tag{28}$$

which has dimension $n_y \times |V|$. Then the solution ϕ minimizing 14 is the solution to the minimization problem

$$||D^{-1/2}\left(Y - X\left(Bd + Qc\right)\right)||^2 + \lambda c' Q c \tag{29}$$

where the (i, j) entry of the $|V| \times |V|$ matrix Q is given by $\langle P_1\xi_i, P_1\xi_j\rangle_{\mathcal{H}}$, and B is the $|V| \times d_0$ matrix with i-ν^{th} element $\eta_\nu(v_i)$, which we assume to be full column rank. The diagonal matrix D holds the $n_y \times n_y$ innovation variances σ^2_{ijk}.

Example 1.2. Construction of X_i with complete data

Straightforward construction of the autoregressive design matrix X_i is straight forward in the case that there are an equal number of measurements on each subject at a common set of measurement times $t_1, \dots, t_M$. When complete data are available for measurement times $t_1, \dots, t_M$,

$$X_i = \begin{bmatrix} y_{i,t_1} & 0 & 0 & 0 & & \dots & 0 \\ 0 & y_{i,t_1} & y_{i,t_2} & 0 & 0 & \dots & 0 \\ \vdots & & & & & & \\ 0 & \dots & 0 & \dots & y_{i,t_1} & \dots & y_{i,t_{M-1}} \end{bmatrix} \tag{30}$$

for all $i = 1, \dots, N$. Note that this design matrix specification does not require that measurement times be regularly spaced.

Example 1.3. Construction of X_i with incomplete data
We demonstrate the construction of the autoregressive design matrices when subjects do not share a universal set of observation times for $N = 2$; the construction extends naturally for an arbitrary number of trajectories. Let subjects have corresponding sample sizes $m_1 = 4$, $m_2 = 4$, with measurements on subject 1 taken at $t_{11} = 0, t_{12} = 0.2, t_{13} = 0.5, t_{14} = 0.9$ and on subject 2 taken at $t_{21} = 0, t_{22} = 0.1, t_{23} = 0.5, t_{24} = 0.7$. Then the unique within-subject pairs of observation times (t, s) such that $0 \leq s < t \leq 1$ are

t	0.1	0.2	0.5	0.5	0.5	0.7	0.7	0.7	0.9	0.9	0.9
s	0.0	0.0	0.0	0.1	0.2	0.0	0.1	0.5	0.0	0.2	0.5

This gives that $V = \{\boldsymbol{v}_{121}, \dots, \boldsymbol{v}_{143}\} \bigcup \{\boldsymbol{v}_{221}, \dots, \boldsymbol{v}_{243}\} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{11}\}$, where the distinct observed $v = (l, m)$ are

l	0.10	0.20	0.50	0.40	0.30	0.70	0.60	0.20	0.90	0.70	0.40
m	0.05	0.10	0.25	0.30	0.35	0.35	0.40	0.60	0.45	0.55	0.70

Then a potential construction of the autoregressive design matrix for subject is given by:

$$X_1 = \begin{bmatrix} 0 & y_{1,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{1,1} & 0 & y_{1,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{1,1} & y_{1,2} & y_{1,3} \end{bmatrix} \tag{31}$$

and similarly, for subject 2:

$$X_2 = \begin{bmatrix} y_{2,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{2,1} & y_{2,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & y_{2,1} & y_{2,2} & y_{2,3} & 0 & 0 & 0 \end{bmatrix} \tag{32}$$

1.2.2 Construction of the solution $\hat{\phi}$

Differentiating $-2\ell_\phi + \lambda J(\phi)$ with respect to c and d and setting equal to zero, we have that

$$\begin{aligned} \frac{\partial}{\partial c}\left[-2\ell_\phi + \lambda J(\phi)\right] &= QX'D^{-1}\left[X(Bd+Qc) - Y\right] + \lambda Qc = 0 \\ &\iff X'D^{-1}X\left[Bd+Qc\right] + \lambda c = X'D^{-1}Y \end{aligned} \tag{33}$$

$$\begin{aligned} \frac{\partial}{\partial d}\left[-2\ell_\phi + \lambda J(\phi)\right] &= B'X'D^{-1}\left[X(Bd+Qc) - Y\right] = 0 \\ &\iff -\lambda B'c = 0 \end{aligned} \tag{34}$$

For fixed smoothing parameter, the solution ϕ is obtained by finding c and d which satisfy

$$Y = X\left[Bd + \left(Q + \lambda\left(X'D^{-1}X\right)^{-1}\right)c\right] \tag{35}$$

$$B'c = 0 \tag{36}$$

Letting $\tilde{Y} = D^{-1/2}Y$, $\tilde{B} = D^{-1/2}XB$, and $\tilde{Q} = D^{-1/2}XQ$, the penalized log likelihood **??** may be written

$$-2\ell_\lambda(c,d) + \lambda J(\phi) = \left[\tilde{Y} - \tilde{B}d - \tilde{Q}c\right]'\left[\tilde{Y} - \tilde{B}d - \tilde{Q}c\right] + \lambda c'Qc. \tag{37}$$

Taking partial derivatives with respect to d and c and setting equal to zero yields normal equations

$$\begin{aligned} \tilde{B}'\tilde{B}d + \tilde{B}'\tilde{Q}c &= \tilde{B}'\tilde{Y} \\ \tilde{Q}'\tilde{B}d + \tilde{Q}'\tilde{Q}c + \lambda Qc &= \tilde{Q}'\tilde{Y}, \end{aligned} \tag{38}$$

Some algebra yields that this is equivalent to solving the system

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q} \\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q} + \lambda Q \end{bmatrix} \begin{bmatrix} d \\ c \end{bmatrix} = \begin{bmatrix} \tilde{B}'\tilde{Y} \\ \tilde{Q}'\tilde{Y} \end{bmatrix} \tag{39}$$

Fixing smoothing parameters λ and θ_β (hidden in Q and $\tilde{Q}$ if present), assuming that $\tilde{Q}$ is full column rank, 39 can be solved by the Cholesky decomposition of the $(n + d_0) \times (n + d_0)$

matrix followed by forward and backward substitution. See **?**. Singularity of $\tilde{Q}$ demands special consideration. Write the Cholesky decomposition

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q} \\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q}+\lambda Q \end{bmatrix} = \begin{bmatrix} C_1' & 0 \\ C_2' & C_3' \end{bmatrix} \begin{bmatrix} C_1 & C_2 \\ 0 & C_3 \end{bmatrix} \tag{40}$$

where $\tilde{B}'\tilde{B} = C_1'C_1$, $C_2 = C_1^{-T}\tilde{B}'\tilde{Q}$, and $C_3'C_3 = \lambda Q + \tilde{Q}'\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{Q}$. Using an exchange of indices known as pivoting, one may write

$$C_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} H \\ 0 \end{bmatrix},$$

where H_1 is nonsingular. Define

$$\tilde{C}_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & \delta I \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} C_1 & C_2 \\ 0 & \tilde{C}_3 \end{bmatrix}; \tag{41}$$

then

$$\tilde{C}^{-1} = \begin{bmatrix} C_1^{-1} & -C_1^{-1}C_2\tilde{C}_3^{-1} \\ 0 & \tilde{C}_3^{-1} \end{bmatrix}. \tag{42}$$

Premultiplying 40 by $\tilde{C}^{-T}$, straightforward algebra gives

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{C}_3^{-T}C_3^TC_3\tilde{C}_3^{-1} \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} C_1^{-T}\tilde{B}'\tilde{Y} \\ \tilde{C}_3^{-T}\tilde{Q}'\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{Y} \end{bmatrix} \tag{43}$$

where $\left(\tilde{d}' \;\; \tilde{c}'\right)' = \tilde{C}'\,(d \;\; c)'$. Partition $\tilde{C}_3 = \begin{bmatrix} K & L \end{bmatrix}$; then $HK = I$ and $HL = 0$. So

$$\begin{aligned} \tilde{C}_3^{-T}C_3^TC_3\tilde{C}_3^{-1} &= \begin{bmatrix} K' \\ L' \end{bmatrix} C_3'C_3 \begin{bmatrix} K & L \end{bmatrix} \\ &= \begin{bmatrix} K' \\ L' \end{bmatrix} H'H \begin{bmatrix} K & L \end{bmatrix} \\ &= \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

If $L'C_3^TC_3L = 0$, then $L'\tilde{Q}'\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{Q}L = 0$, so $L'\tilde{Q}'\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{Y} = 0$. Thus, the linear system has form

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} * \\ * \\ 0 \end{bmatrix}, \tag{44}$$

which can be solved, but with c_2 arbitrary. One may perform the Cholesky decomposition of 39 with pivoting, replace the trailing 0 with δI for appropriate value of δ, and proceed as if $\tilde{Q}$ were of full rank.

It follows that

$$\widehat{\tilde{Y}} = \tilde{B}d + \tilde{Q}c = \begin{bmatrix} \tilde{B} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1}\tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}' \end{bmatrix} \tilde{Y} = \tilde{A}\left(\lambda, \boldsymbol{\theta}\right) \tilde{Y}. \tag{45}$$

where

$$\begin{aligned} \tilde{A}\left(\lambda, \boldsymbol{\theta}\right) &= \begin{bmatrix} \tilde{B} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1}\tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}' \end{bmatrix} \\ &= G + \left(I - G\right)\tilde{Q}\left[\tilde{Q}'\left(I - G\right)\tilde{Q} + \lambda Q\right]^{-1}\tilde{Q}'\left(I - G\right), \end{aligned} \tag{46}$$

for

$$G = \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'.$$

TO DO: discuss efficient approximation of $\mathcal{H}$ using a low rank smoother technique, as discussed in Gu's book at the end of Chapter 3. **?**

1.2.3 Smoothing parameter selection

By varying smoothing parameters λ and θ_β, the minimizer ϕ_λ of 39 defines a family of potential estimates. In practice, we need to choose a specific estimate from the family, which requires effective methods for smoothing parameter selection. We consider two criteria that are commonly used for smoothing parameter selection in the context of smoothing spline models for longitudinal data. The first score is an unbiased estimate of a relative loss and assumes a known variances σ_t^2. The unbiased risk estimate has attractive asymptotic properties; see Gu [2013] for a comprehensive examination. The second score, the leave-one-subject-out cross validation (losoCV) score, provides an estimate of the same loss without assuming a known variance function. We review a computationally convenient approximation of the losoCV score proposed by **?**, who demonstrates the shortcut score's asymptotic optimality. To simplify notation for the initial presentation, we only make explicit the dependence of estimates and their components on λ and conceal any dependence on θ_β.

1.3 Model selection criteria

1.3.1 Unbiased risk estimate

Define $\tilde{Y} = D^{-1/2}Y$, $\tilde{B} = D^{-1/2}XB$, and $\tilde{Q} = D^{-1/2}XQ$ as before. Let $\tilde{\epsilon} = D^{-1/2}\epsilon$ denote the vector of length $\sum_{i=1}^N m_i - N$ containing the standardized prediction errors $\epsilon_{ij} \sim N\left(0, 1\right)$, and write the vector of transformed means

$$\Phi = D^{-1/2} X \left[Bd + Qc \right]. \tag{47}$$

We can assess $\hat{\tilde{Y}}_{\lambda}$, an estimate of the mean of $\tilde{Y}$ based on observed data y_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, m_i$, using the loss function

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{N} \sum_{j=1}^{m_i} \left(\hat{\tilde{y}}_{ij} - E\left[\tilde{y}_{ij} \right] \right)^2 \\ &= ||\tilde{Y} - \tilde{\mu}||^2 \end{aligned} \tag{48}$$

where $\mu = D^{-1/2} W \Phi^*$ denotes the $\left(\sum_i m_i - N \right) \times 1$ with i^{th} element equal to the expected value of the i^{th} element of $\tilde{Y}$. Then straightforward algebra yields that

$$L(\lambda) = \mu' \left(I - \tilde{A} \right)^2 \mu - 2\mu' \left(I - \tilde{A} \right)^2 \tilde{A} \tilde{\epsilon} + \tilde{\epsilon}' \tilde{A}^2 \tilde{\epsilon} \tag{49}$$

Define the unbiased risk estimate

$$U(\lambda) = \frac{1}{N} \tilde{Y}' \left(I - \tilde{A} \right)^2 \tilde{Y} + \frac{2}{N} \text{tr} \tilde{A} \tag{50}$$

Adding and substracting μ to the quadratic terms, one can verify with straightforward algebra that

$$\begin{aligned} U(\lambda) &= \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y} \right)' \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y} \right) + 2\text{tr}\tilde{A} \\ &= \left(\tilde{A}\tilde{Y} - \mu \right)' \left(\tilde{A}\tilde{Y} - \mu \right) + \tilde{\epsilon}'\tilde{\epsilon} + 2\tilde{\epsilon}' \left(I - \tilde{A} \right) \mu - 2 \left(\tilde{\epsilon}' \tilde{A} \tilde{\epsilon} - \text{tr}\tilde{A} \right) \end{aligned} \tag{51}$$

This gives

$$U(\lambda) - L(\lambda) - \tilde{\epsilon}'\tilde{\epsilon} = 2\tilde{\epsilon}' \left(I - \tilde{A} \right) \mu - 2 \left(\tilde{\epsilon}' \tilde{A} \tilde{\epsilon} - \text{tr}\tilde{A} \right), \tag{52}$$

which allows one to easily see that $U(\lambda)$ is unbiased for the relative loss $L(\lambda) + \tilde{\epsilon}'\tilde{\epsilon}$. Under mild conditions on the risk function

$$R(\lambda) = E\left[L(\lambda) \right],$$

one can establish that U is also a consistent estimator. See Gu [2013] Chapter 3 for a formal theorem and proof.

1.4 Leave-one-subject-out cross validation

The conditions under which the the cross validation and generalized cross validation scores traditionally used for smoothing parameter selection yield desirable properties generally do not hold when the data are clustered or longitudinal in nature. Instead, the leave-one-subject-out (LosoCV) cross validation score has been widely used for smoothing parameter selection for semiparametric and nonparametric models for longitudinal or functional data. The LosoCV criterion is defined as

$$V_{loso}\left(\lambda\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right)'\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right) \tag{53}$$

where $\widehat{\tilde{\mu}}_i^{[-i]}$ is the estimate of $E\left[\tilde{Y}_i\right]$ based on the data when $\tilde{Y}_i$ is omitted. Intuitively, the LosoCV score is appealing because it preserves any within-subject dependence by leaving out all observations from the same subject together in the cross-validation. However, despite its prevalent use, theoretical justifications for its use have not been established. In their seminal work, **?** were the first to present a heuristic justification of LosoCV by demonstrating that it mimics the mean squared prediction error: consider new observations $\tilde{Y}_i^* = \left(\tilde{y}_{i1}^*, \tilde{y}_{i1}^*, \ldots, \tilde{y}_{i,m_i}^*\right)$. We may write the mean squared prediction error for the new observations as follows:

$$\begin{aligned} MSPE &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - \widehat{\tilde{\mu}}_i||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - D_i^{-1/2}W_i\Phi^* + D_i^{-1/2}W_i\Phi^* - D_i^{-1/2}W_i\hat{\Phi}^*||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\tilde{\mu}_i - \widehat{\tilde{\mu}}_i^{[-i]}||^2\right]\right\} \end{aligned} \tag{54}$$

where $\tilde{\epsilon}_i = \tilde{Y}_i^* - D_i^{-1/2}W_i\Phi^*$. When $\left\{\sigma^2\left(t\right)\right\}$ is known, $\tilde{\epsilon}_i$ is a mean zero multivariate normal vector with $Cov\left(\tilde{\epsilon}_i\right) = I_{m_i}$, which gives the last equality. Since $\tilde{Y}_i$ and $\widehat{\tilde{\mu}}_i$ are independent, the expected LosoCV score can be written

$$E\left[V_{loso}\left(\lambda\right)\right] = \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\widehat{\tilde{\mu}}_i - \tilde{\mu}_i||^2\right]\right\}. \tag{55}$$

When N is large, we expect that $\widehat{\tilde{\mu}}_i$ should be close to $\widehat{\tilde{\mu}}_i^{[-i]}$, so $E\left[V_{loso}\left(\lambda\right)\right]$ should be a good approximation to the mean-squared prediction error. For a formal proof of consistency, see **?**.

1.4.1 Computation of the LosoCV score

Lemma 1.1 (Shortcut formula for LosoCV). *The LosoCV score satisfies the following identity:*

$$V_{loso}\left(\lambda\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i - \widehat{\tilde{Y}}_i\right)'\left(I_{ii} - \tilde{A}_{ii}\right)^{-T}\left(I_{ii} - \tilde{A}_{ii}\right)^{-1}\left(\tilde{Y}_i - \widehat{\tilde{Y}}_i\right),$$

where $\tilde{A}_{ii}$ is the diagonal block of smoothing matrix $\tilde{A}$ corresponding to the observations on subject i, and I_{ii} is a $m_i \times m_i$ identity matrix.

A detailed presentation and proof can be found in **?** and supplementary materials **?**. The authors additionally proposed an approximation to the LosoCV score to further reduce the computational cost of evaluating V_{loso}, which can be expensive due to the inversion of the $I_{ii} - \tilde{A}_{ii}$. Using the Taylor expansion of $\left(I_{ii} - \tilde{A}_{ii}\right)^{-1} \approx I_{ii} + \tilde{A}_{ii}$, we can use the following to approximate V_{loso}:

$$V^*_{loso}\left(\lambda\right) = \frac{1}{N}||\left(I - \tilde{A}\right)\tilde{Y}||^2 + \frac{2}{N}\sum_{i=1}^{N} \hat{\tilde{e}}_i' \tilde{A}_{ii} \hat{\tilde{e}}_i, \tag{56}$$

where $\hat{\tilde{e}}_i$ is the portion of the vector of prediction errors $\left(I - \tilde{A}\right)\tilde{Y}$ corresponding to subject i. They show that under mild conditions, and for fixed, nonrandom λ, the approximate losoCV score V^* and the true losoCV score V_{loso} are asymptotically equivalent. See Theorem 3.1 of **?**.

1.5 Selection of multiple smoothing parameters

With the definition of the unbiased risk estimate and the leave-one-subject-out criteria, the expression of the smoothing matrix in Equation 46 permits the straightforward evaluation of both scores $U\left(\lambda, \boldsymbol{\theta}\right)$ and $V^*_{loso}\left(\lambda, \boldsymbol{\theta}\right)$, where $\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_g\right)'$ denotes the vector of smoothing parameters associated with each RK. In this section, we discuss a algorithm to minimize the unbiased risk estimate $U\left(\lambda, \boldsymbol{\theta}\right)$ with respect to λ and $\boldsymbol{\theta}$ hidden in $Q = \sum_{\beta=1}^{q} \theta_\beta Q_\beta$, where the (i,j) entry of Q_β is given by $R_\beta\left(\boldsymbol{v}_i, \boldsymbol{v}_j\right)$. We present minimization of the unbiased risk estimate explicitly, but the mechanics of the optimization are very similar to those necessary for optimizing the leave-one-subject-out cross validation criterion. The details of a procedue for explicitly minimizing the alternative criterion are presented in **?**, which is based on the algorithms of **?**, **?** (which is the basis for the algorithm which follows) and **?**. The key difference between the minimization of U and the minimization of V^*_{loso} lies in the calculation of the gradient and the Hessian matrix in the Newton update. To minimize the unbiased risk estimate,

I. Fix $\boldsymbol{\theta}$; minimize $U\left(\lambda|\boldsymbol{\theta}\right)$ with respect to λ.

II. Update $\boldsymbol{\theta}$ using the current estimate of λ.

Executing step 1 follows immediately from the expression for the smoothing matrix. Step 2 requires evaluating the gradient and the Hessian of $U\left(\boldsymbol{\theta}|\lambda\right)$ with respect to $\boldsymbol{\kappa} = \log\left(\boldsymbol{\theta}\right)$. Optimizing with respect to $\boldsymbol{\kappa}$ rather than on the original scale is motivated by two driving factors: first, $\boldsymbol{\kappa}$ is invariant to scale transformations. With examination of U and V^* and **??**, it is immediate that the $\theta_\beta \tilde{Q}_\beta$ are what matter in determining the minimum. Multiplying the $\tilde{Q}_\beta$ by any positive constant leaves the θ_β subject to rescaling, though the problem itself is unchanged by scale transformations. The derivatives of $U(\cdot)$ with respect to $\boldsymbol{\kappa}$ are invariant to such transformations, while the derivatives with respect to $\boldsymbol{\theta}$ are not. In addition, optimizing with respect to $\boldsymbol{\kappa}$ converts a constrained optimization ($\theta_\beta \geq 0$) problem to an unconstrained one.

1.5.1 Algorithms

The following presents the main algorithm for minimizing $U(\lambda, \boldsymbol{\theta})$ and its key components are presented in the section to follow. The minimization of U is done via two nested loops. Fixing tuning parameters, the outer loop minimizes U with respect to smoothing parameters via quasi-Newton iteration of **?**, as implemented in the `nlm` function in `R`. The inner loop then minimizes ℓ_λ with fixed tuning parameters via Newton iteration. Fixing the θ_βs in $J(\phi^*) = \sum_\beta \theta_\beta^{-1} J_\beta\left(\phi_\beta^*\right)$, the outer loop with a single λ is a straightforward task.

Algorithm 1

```
Initialization:
Set $\Delta\boldsymbol{\kappa} := 0$;  $\boldsymbol{\kappa}_- := \boldsymbol{\kappa}_0$;  $V_- = \infty$;  ( or $M_- = \infty$)
Iteration:
while not converged do
  For current value $\boldsymbol{\kappa}^* = \boldsymbol{\kappa}_- + \Delta\boldsymbol{\kappa}$, compute $Q_\theta^* = \sum_{\beta=1}^{g} \theta_\beta^* Q_\beta$ and scale so that tr $(Q_\beta)$ is fixed.
  Compute $\tilde{A}\left(\lambda|\boldsymbol{\theta}^*\right) = \tilde{A}\left(\lambda, \exp\left(\boldsymbol{\kappa}^*\right)\right)$.
  Minimize $U\left(\lambda|\boldsymbol{\kappa}^*\right) = \tilde{Y}'\left(I - \tilde{A}\right)^2 \tilde{Y} + 2\mathrm{tr}\tilde{A}$
  Set $U_* := \min_\lambda Y\left(\lambda|\boldsymbol{\kappa}^*\right)$
  if $U^* > U_-$ then
    Set $\Delta\boldsymbol{\kappa} := \Delta\boldsymbol{\kappa}/2$
    Go to (1).
  else
    Continue
  end if
  Evaluate gradient $\mathbf{g} = \left(\partial/\partial\boldsymbol{\kappa}\right) U\left(\boldsymbol{\kappa}|\lambda\right)$
  Evaluate Hessian $H = \left(\partial^2/\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}'\right) U\left(\boldsymbol{\kappa}|\lambda\right)$.
  Calculate step $\Delta\boldsymbol{\kappa}$:
  if $H$ positive definite then
    $\Delta\boldsymbol{\kappa} := -H^{-1}\mathbf{g}$
  else
    $\Delta\boldsymbol{\kappa} := -\tilde{H}^{-1}\mathbf{g}$, where $\tilde{H} = \mathrm{diag}\left(\boldsymbol{\epsilon}\right)$ is positive definite.
  end if
end while
Calculate optimal model:
if $\Delta\kappa_\beta < -\gamma$, for $\gamma$ large then
  Set $\kappa_{*\beta} := -\infty$
end if
Compute $Q_\theta^* = \sum_{\beta=1}^{g} \theta^* \beta Q_\beta$;
Calculate $\begin{bmatrix} d \\ c \end{bmatrix} = \tilde{C}^{-1}\tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}_*^{\theta\prime} \end{bmatrix} \tilde{Y}$
```

Calculation of the gradient $\boldsymbol{g}$ and Hessian H mirror the details in **?**, replacing the null basis matrix B and representer matrix Q with $D^{-1}XB$ and $D^{-1}XB$, respectively. They also present details on convergence criteria based on those suggested in **?**, who also present detailed discussion of the Newton method based on the Cholesky decomposition necessary for calculating the update direction for $\boldsymbol{\kappa}$. The step in 21 returns a descent direction even when H is not positive definite by adding positive mass to the diagonal elements of H if necessary to produce $\tilde{H} = G'G$ where G is upper triangular. See **?** 4.4.2.2 for details.

The unbiased risk estimate $U\left(\lambda, \boldsymbol{\theta}\right)$ is fully parameterized by

$$\left(\lambda_1, \ldots, \lambda_q\right) = \left(\lambda\theta_1^{-1}, \ldots, \lambda\theta_q^{-1}\right), \tag{57}$$

so the smoothing parameters $\left(\lambda, \theta_1, \ldots, \theta_q\right)$ over-parameterize the score, which is the reason for scaling the trace of Q_β. The starting values for the θ quasi-Newton iteration are obtained with two passes of the fixed-θ outer loop as follows:

I. Set $\breve{\theta}_\beta^{-1} \propto \text{tr}\left(\tilde{Q}_\beta\right)$, minimize $U\left(\lambda\right)$ with respect to λ to obtain $\breve{\phi}$.

II. Set $\breve{\theta}_\beta^{-1} \propto J_\beta\left(\breve{\phi}_\beta\right)$, minimize $U\left(\lambda\right)$ with respect to λ to obtain $\breve{\phi}$.

The first pass allows equal opportunity for each penalty to contribute to the GCV score, allowing for arbitrary scaling of $J_\beta\left(\phi_\beta\right)$. The second pass grants greater allowance to terms exhibiting strength in the first pass. The following θ iteration fixes λ and starts from $\breve{\theta}_\beta$. These are the starting values adopted by **?**; the starting values for the first pass loop are arbitrary, but are invariant to scalings of the θ_β. The starting values in II for the second pass of the outer are based on more involved assumptions derived from the background formulation of the smoothing problem: the penalty is of the form

$$J\left(\right) = \sum_{\beta=1}^{q} \theta_\beta^{-1} \langle \phi, \phi \rangle_\beta$$

After the first pass, the initial fit $\breve{\phi}$ reveals where the structure in the true ϕ lie in terms of the components of the subspaces $\mathcal{H}_\beta$. Less penalty should be applied to terms exhibiting strong signal.

1.5.2 An RKHS framework for estimating $\log \sigma^2$

Once we have an initial estimate of the generalized autoregressive coefficient function, ϕ, we can use the model residuals to estimate the innovation variance function $\sigma^2\left(t\right)$. We use the same estimation approach as outlined in Section 1.2.1. Fixing $\phi = \phi^*$ for given estimate ϕ^*, the negative log likelihood of the data $Y_1, \ldots, Y_N$ is satisfies

$$-\ell\left(Y_1, \ldots, Y_N, \phi, \sigma^2\right) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{m_i} \log \sigma_{ij}^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{\epsilon_{ij}^2}{\sigma_{ij}^2}; \tag{58}$$

where $\epsilon_{ij} = y_{ij} - \sum_{k<j} \phi_{ijk}^* y_{ik}$. Let

$$\text{RSS}\left(t\right) = \sum_{i,j:t_{ij}=t} \left(y_{ij} - \sum_{k<j} \phi_{ijk} y_{ik}\right)^2 \tag{59}$$

denote the squared residuals for the observations y_{ij} having corresponding measurement time $t_{ij} - t$. Then $\text{RSS}\left(t\right)/\sigma^2\left(t\right) \sim \chi_{df_t}^2$, where the degrees of freedom df_t corresponds to the number of observations y_{ij} having corresponding measurement time t. In this light, for fixed ϕ, the penalized likelihood 58 is that of a variance model with the ϵ_{ij}^2 serving as the response. This corresponds to a generalized linear model with gamma errors and known scale parameter equal to 2. Let $z_{ij} = \epsilon_{ij}^2$, and let $Z_i = \left(z_{i1}, z_{i,m_i}\right)'$ denote the vector of residuals for the i^{th} observed trajectory. The Gamma distribution is parameterized by shape parameterα and scale parameter β, where the mean of the distribution given by $\mu = \alpha\beta$. Reparameterizing the Gamma likelihood in terms of (α, μ) and dropping terms that don't involve $\mu(\cdot)$ gives

$$-\ell(z, \mu, \alpha) \propto \alpha \left[\frac{z}{\mu} + \log \mu\right] \tag{60}$$

$$= \alpha \left[ze^{-\eta} + \eta\right], \tag{61}$$

where α^{-1} is the dispersion parameter and $\eta = \log \mu$. Letting μ_{ij} denote $E[z_{ij}] = \sigma_{ij}^2$, the log likelihood of the working residuals becomes

$$-\ell\left(Z_1, \dots, Z_N, \phi, \sigma^2\right) = \sum_{i=1}^{N}\sum_{j=1}^{m_i} \log \mu_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{z_{ij}}{\mu_{ij}}, \tag{62}$$

which we can see coincides with a Gamma dsitribution with scale parameter $\alpha = 2$. Smoothing spline ANOVA models for exponential familes have been studied extensively (**?**, **?**, Gu [2013]). Parallel to the penalized sums of squares for ϕ (14), we can append a smoothness penalty to obtain the penalized likelihood for $\eta(t) = \log \sigma^2(t)$:

$$-\ell\left(Z_1, \dots, Z_N, \phi, \sigma^2\right) + = \sum_{i=1}^{N}\sum_{j=1}^{m_i} \eta_{ij} + \sum_{i=1}^{N}\sum_{j=1}^{m_i} z_{ij}e^{-\eta_{ij}} + \lambda J(\eta), \tag{63}$$

noindent for $\eta \in \mathcal{H} = \oplus_{\beta=0}^{q}\mathcal{H}_\beta$, where the penalty J can be written as a square norm and decomposed as in (27), with

$$J(\kappa) = \langle \eta, \eta \rangle = \sum_{\beta=1}^{q} \theta_\beta^{-1} \langle \eta, \eta \rangle_\beta.$$

The $\langle \cdot, \cdot \rangle_\beta$ are inner products in $\mathcal{H}_\beta$ having reproducing kernels $Q_\beta(t, t')$. The penalty $J(\kappa)$ is an inner product in $\oplus_{\beta=0}^{q}\mathcal{H}_\beta$ with reproducing kernel $\sum_{\beta=1}^{q} \theta_\beta Q_\beta(t, t')$ and null space $\mathcal{N}_J = \mathcal{H}_0$. The first term in (63) serves as a measure of the goodness of fit of κ to the data, and only depends on κ through the evaluation functional $[t_{ij}]\,\kappa$. So the argument justifying the form of the minimizer in (15) applies, and the minimizer of the penalized likelihood has the form

$$\eta(t) = \sum_{\nu=1}^{d_0} d_\nu \kappa_\nu(t) + \sum_{i=1}^{|\mathcal{T}|} c_i Q_J(t, t_i), \tag{64}$$

where $\mathcal{T} = \bigcup_{j=1}^{N}\bigcup_{k=1}^{m_i} t_{jk}$ denotes the unique values of the observations times pooled across subjects, where $\{\kappa_\nu\}_{\nu=1}^{d_0}$ is a basis for the null space $\mathcal{N}_J = \mathcal{H}_0$.

Standard theory for exponential families gives us that the functional

$$L(\eta) = -\sum_{i=1}^{N}\sum_{j=1}^{m_i}\left[z_{ij}\eta\left(t_{ij}\right)\right] \tag{65}$$

$$\tag{66}$$

is continuous and convex in $\eta \in \mathcal{H}$

1.5.3 Smoothing parameter selection for exponential familes

The gamma penalized log likelihood (64) is non-quadratic, so η_λ must be computed using iteration even for fixed smoothing parameter. Performance-oriented iteration and generalized approximate cross validation (GACV) are the most common approaches to selecting the smoothing parameter for penalized regression with exponential families. As in our discussion of model selection for ϕ, we omit dependence of any components on the θ_β and only explicitly express dependence on smoothing parameters through λ.

1.5.4 Performance-oriented iteration

A measure of the discrepancy between distributions belonging to an exponential family with density $p_Z(z) = exp\left\{\left(y\eta(t) - b(\eta(t))\right)/a(\phi) + c(y,\phi)\right\}$ is the Kullback-Leibler distance

$$\begin{aligned} \text{KL}(\eta,\eta_\lambda) &= E_\lambda\left[Z(\eta-\eta_\lambda) - (b(\eta) - b(\eta_\lambda))\right]/a(\phi) \\ &= \left[b'(\eta)(\eta-\eta_\lambda) - (b(\eta) - b(\eta_\lambda))\right]/a(\phi), \end{aligned} \tag{67}$$

which simplifies to

$$-\mu\left(e^{-\eta} - e^{-\tilde{\eta}}\right) - (\eta - \tilde{\eta})$$

for the Gamma distribution. The KL distance is not symmetric, so sometimes people opt for its symmetrized version:

$$\begin{aligned} \text{SKL}(\eta,\eta_\lambda) &= \text{KL}(\eta,\eta_\lambda) + \text{KL}(\eta_\lambda,\eta) \\ &= \left(b'(\eta) - b'(\eta_\lambda)\right)(\nu - \nu_\lambda)/a(\phi), \\ &= (\mu - \mu_\lambda)(\nu - \nu_\lambda)/a(\phi), \end{aligned} \tag{68}$$

A natural choice of loss function for measuring the performance of an estimator $\eta_\lambda(t)$ of $\eta(t)$ is the symmetrized Kullback-Leibler distance averaged over the observed time points $t_{11},\dots,t_{1,m_1},\dots,t_{N1},\dots,t_{N,m_N}$:

$$L(\eta,\eta_\lambda) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_i}\left(\mu(t_{ij}) - \mu_\lambda(t_{ij})\right)\left(\nu(t_{ij}) - \nu_\lambda(t_{ij})\right), \tag{69}$$

which reduces to

$$L\left(\eta,\eta_{\lambda}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_i}\left(\mu\left(t_{ij}\right)-\mu_{\lambda}\left(t_{ij}\right)\right)\left(\nu\left(t_{ij}\right)-\nu_{\lambda}\left(t_{ij}\right)\right), \tag{70}$$

for the Gamma distribution. The ideal smoothing parameters are those which minimize (70). The performance-oriented iteration operates on a alternative expression of the symmetrized Kullback-Leibler loss. The mean value theorem gives us that (70) can be written

$$L_{\omega}\left(\eta,\eta_{\lambda}\right)=L\left(\eta,\eta_{\lambda}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_i}\omega^{*}\left(t_{ij}\right)\left(\nu\left(t_{ij}\right)-\nu_{\lambda}\left(t_{ij}\right)\right)^{2}, \tag{71}$$

where $\omega^{*}\left(t_{ij}\right) = b''\left(\eta^{*}\left(t_{ij}\right)\right)$ and $\eta^{*}\left(t_{ij}\right)$ is a convex combination of $\eta\left(t_{ij}\right)$ and $\eta_{\lambda}\left(t_{ij}\right)$. One can construct an unbiased risk estimate under the weighted loss, L_{ω}, using re-weighted observations. Letting $Z_{i\omega} = W_i Z_i$, where W_i ist he $m_i \times m_i$ diagonal matrix having diagonal entries $\omega^{*}\left(t_{i1}\right),\ldots,\omega^{*}\left(t_{i,m_i}\right)$, an unbiased estimate of relative loss is given by

$$U_{\omega}\left(\lambda\right)=L\left(\eta,\eta_{\lambda}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_i}\omega^{*}\left(t_{ij}\right)\left(\nu\left(t_{ij}\right)-\nu_{\lambda}\left(t_{ij}\right)\right)^{2}, \tag{72}$$

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Wolfgang Dahmen, Charles A Micchelli, and Hans-Peter Seidel. Blossoming begets ??-spline bases built better by ??-patches. *Mathematics of computation*, 59(199):97–115, 1992.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

Paul HC Eilers, Iain D Currie, and Maria Durbán. Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50(1):61–76, 2006.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, pages 85–98, 2006.

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

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Brian D Marx and Paul HC Eilers. Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22, 2005.

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

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

Mohsen Pourahmadi. Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, pages 425–435, 2000.

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

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