Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

February 28, 2018

1 Smoothing Spline Varying-coefficient Models for Covariance Estimation

A predominant difficulty in the estimation of covariance matrices is the potentially high dimensionality of the problem, as the number of unknown elements in the covariance matrix grows quadratically with the size of the matrix. It is well-known that the sample covariance matrix can be unstable in high dimensions; ways for controlling the complexity of estimates is highly desirable for improving stability of estimates. In the longitudinal-data literature, it is a common practice to use parametric models for the covariance structure. Many have specified parsimonious parametric models for ϕ_{ijk} to overcome the issue of dimensionality.

We naturally accommodate irregularly spaced data and unequal sample sizes between subjects by defining the autoregressive parameters as the values of a smooth function evaluated at within-subject pairs of observed time points. Furthermore, by viewing $\phi(t, s)$ as a smooth *bivariate* function, we can utilize the information across the subdiagonals of T to inform the fit, rather than treating each subdiagonal separately. As in the classical nonparametric function estimation setting, we assume ϕ to vary in a high-dimensional (possibly infinite) function space. We propose two representations of $\phi(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$: approximation by smoothing splines and approximation by B-spline basis expansion.

We assume $Y(t)$ has covariance function $G(t, s)$ and that $\epsilon(t)$ follows a zero mean Gaussian white noise process with unit variance. Under mild assumptions regarding the behaviour of Y, then $G(t, s)$ satisfies some smoothness conditions, where smoothness is defined in terms of square integrability of certain derivatives. We view the entries of Σ as values of G evaluated at the distinct pairs of within-subject observed time points.

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

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

variances as the evaluation of the smooth functions $\tilde{\phi}(t,s)$ and $\sigma^2(t)$ 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):

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

Adopting similar notation for the innovation variance function, denote

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

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

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

Our goal is now to estimate the above model, utilizing bivariate smoothing to estimate $\tilde{\phi}(t,s)$ for $0 \leq s < t \leq 1$, and one-dimensional smoothing to estimate $\sigma(t)$, $0 \leq t \leq 1$. 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.

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 $y(t)$ 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 $Cov(y(t_i), y(t_j)) = G(t_i, t_j) = G(||t_i - t_j||)$ can then be written as

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

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

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

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 **?** and **?** 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 **?**, 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 $\ell(Y, \Sigma)$ satisfies

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

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

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

and

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

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

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

where

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

where $\phi(v_{ij}) = \phi(l_{ij}, m_{ij}) = \tilde{\phi}(t_i, t_j)$. 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

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

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 **??**. 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 **?**, **?**, and **?**. A procedure for minimizing **??** 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

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

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 **??** 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 $\left\{ {r_{ij}^*}^2 = \left(y_{ij} - \sum_{k<j} \phi^*\left(v_{ijk}\right) y_{ik}\right) \right\}$ denote the working residuals based on the current estimate of ϕ. This process of iteratively updating ϕ^* and ${\sigma^2}^*$ 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 **?**, **?**, and **?** 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 **?**; 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 $[v] f = f(v)$ is continuous in $\mathcal{H}$ for all $v \in \mathcal{V}$. The Reisz Representation Theorem gives that there exists $Q \in \mathcal{H}$, the representer of the evaluation functional $[v](\cdot)$, such that $\langle Q_v, f \rangle_{\mathcal{H}} = f(v)$ for all $f \in \mathcal{H}$. See **?** Theorem 2.2.

The symmetric, bivariate function $Q(v_1, v_2) = Q_{v_2}(v_1) = \langle Q_{v_1}, Q_{v_2} \rangle_{\mathcal{H}}$ 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. $Q(\cdot, v) \in \mathcal{H}$

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

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 **?**, Theorem 2.3. The kernel satisfies that for any $\{v_1, \ldots, v_{n_1}\}$, $\{\breve{v}_1, \ldots, \breve{v}_{n_2}\} \in \mathcal{V}$ and $\{a_1, \ldots, a_{n_1}\}$, $\{a_1, \ldots, a'_{n_2}\} \in \Re$,

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

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

$$-2\ell_\phi + \lambda J(\phi) = \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} (L_{ijk}\phi) y_{ik} \right)^2 + \lambda ||P_1 \phi||^2 \tag{13}$$

where P_1 is the projection operator which projects ϕ onto the subspace $\mathcal{H}_1$, and L_{ijk} denotes the evalutation functional $[v_{ijk}]\,\phi$. Let ξ_{ijk} denote the representer of L_{ijk}; **?** established that the minimizer of **??** has form

$$\phi(v) = \sum_{\nu=1}^{m} d_\nu \eta_\nu(v) + \sum_{i=1}^{|V|} c_i (P_1 \xi_i) \tag{14}$$

where $V = \bigcup_{i,j,k} v_{ijk}$, and $\{\eta_1, \ldots, \eta_m\}$ span $\mathcal{H}_0$, the null space of P_1,

$$\mathcal{H}_0 = \{\phi \in \mathcal{H} :\ J(\phi) = 0\}.$$

To show this, we start by noting that any $\phi \in \mathcal{H}$ can be written

$$\phi(v) = \sum_{\nu=1}^{m} d_\nu \eta_\nu(v) + \sum_{i=1}^{|V|} c_i (P_1 \xi_i) + \rho(v) \tag{15}$$

where $\rho \perp \mathcal{H}_0,\ \text{span}\{(P_1\xi_j)\}_{j=1}^{|V|}$. To establish that the solution has form **??** requires showing that the minimizer of **??** 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 left to the appendix **??**.

Let Y denote the vector

$$Y = \left(y_{12}, y_{13}, \ldots, y_{1,m_1}, \ldots, y_{N2}, y_{N3}, \ldots, y_{N,m_N}\right)'$$

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. 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 **??**, and stack these on top of one another to obtain

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \tag{16}$$

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

$$||D^{-1/2}(Y - X(Bd + Qc))||^2 + \lambda c' Q c \tag{17}$$

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}.

¡¡¡¡¡¡¡ HEAD
======= ¿¿¿¿¿¿¿ cc10e8225503967265891c7b98fd982e18d01ca5

Example 1.1. 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 $t_1, \ldots, t_M$. When complete data are available for measurement times $t_1, \ldots, t_M$,

$$X_i = \begin{bmatrix} y_{i,t_1} & 0 & 0 & 0 & & \ldots & 0 \\ 0 & y_{i,t_1} & y_{i,t_2} & 0 & 0 & \ldots & 0 \\ \vdots & & & & & & \\ 0 & \ldots & 0 & \ldots & y_{i,t_1} & \ldots & y_{i,t_{M-1}} \end{bmatrix} \tag{18}$$

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

Example 1.2. 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 $V = \{v_{121}, \ldots, v_{143}\} \bigcup \{v_{221}, \ldots, v_{243}\} = \{v_1, \ldots, v_{11}\}$, 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:

$$X_1 = \begin{bmatrix} 0 & y_{1,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{1,1} & 0 & y_{1,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{1,1} & y_{1,2} & y_{1,3} \end{bmatrix} \tag{19}$$

and similarly, for subject 2:

$$X_2 = \begin{bmatrix} y_{2,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{2,1} & y_{2,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & y_{2,1} & y_{2,2} & y_{2,3} & 0 & 0 & 0 \end{bmatrix} \tag{20}$$

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

$$\frac{\partial}{\partial c}\left[-2\ell_\phi + \lambda J\left(\phi\right)\right] = QX'D^{-1}\left[X\left(Bd+Qc\right)-Y\right] + \lambda Qc = 0$$
$$\iff X'D^{-1}X\Big[Bd+Qc\Big] + \lambda c = X'D^{-1}Y \tag{21}$$

$$\frac{\partial}{\partial d}\left[-2\ell_\phi + \lambda J\left(\phi\right)\right] = B'X'D^{-1}\left[X\left(Bd+Qc\right)-Y\right] = 0$$
$$\iff -\lambda B'c = 0 \tag{22}$$

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

$$Y = X\Big[Bd + \Big(Q + \lambda\left(X'D^{-1}X\right)^{-1}\Big)c\Big] \tag{23}$$
$$B'c = 0 \tag{24}$$

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

$$-2\ell_\lambda\left(c,d\right) + \lambda J\left(\phi\right) = \Big[\tilde{Y} - \tilde{B}d - \tilde{Q}c\Big]'\Big[\tilde{Y} - \tilde{B}d - \tilde{Q}c\Big] + \lambda c'Qc. \tag{25}$$

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

$$\begin{aligned}\tilde{B}'\tilde{B}d + \tilde{B}'\tilde{Q}c &= \tilde{B}'\tilde{Y}\\ \tilde{Q}'\tilde{B}d + \tilde{Q}'\tilde{Q}c + \lambda Qc &= \tilde{Q}'\tilde{Y},\end{aligned} \tag{26}$$

Some algebra yields that this is equivalent to solving the system

$$\begin{bmatrix}\tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q}\\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q}+\lambda Q\end{bmatrix}\begin{bmatrix}d\\c\end{bmatrix} = \begin{bmatrix}\tilde{B}'\tilde{Y}\\ \tilde{Q}'\tilde{Y}\end{bmatrix} \tag{27}$$

Fixing smoothing parameters λ and θ_β (hidden in Q and $\tilde{Q}$ if present), assuming that $\tilde{Q}$ is full column rank, **??** 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

$$\begin{bmatrix}\tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q}\\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q}+\lambda Q\end{bmatrix} = \begin{bmatrix}C_1' & 0\\ C_2' & C_3'\end{bmatrix}\begin{bmatrix}C_1 & C_2\\ 0 & C_3\end{bmatrix} \tag{28}$$

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

$$\tilde{C}_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & \delta I \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} C_1 & C_2 \\ 0 & \tilde{C}_3 \end{bmatrix}; \tag{29}$$

then

$$\tilde{C}^{-1} = \begin{bmatrix} C_1^{-1} & -C_1^{-1} C_2 \tilde{C}_3^{-1} \\ 0 & \tilde{C}_3^{-1} \end{bmatrix}. \tag{30}$$

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

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{C}_3^{-T} C_3^T C_3 \tilde{C}_3^{-1} \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} C_1^{-T} \tilde{B}' \tilde{Y} \\ \tilde{C}_3^{-T} \tilde{Q}' \left(I - \tilde{B} \left(\tilde{B}' \tilde{B} \right)^{-1} \tilde{B}' \right) \tilde{Y} \end{bmatrix} \tag{31}$$

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

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

If $L' C_3^T C_3 L = 0$, 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

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} * \\ * \\ 0 \end{bmatrix}, \tag{32}$$

which can be solved, but with c_2 arbitrary. One may perform the Cholesky decomposition of **??** 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

$$\widehat{\tilde{Y}} = \tilde{B} d + \tilde{Q} c = \begin{bmatrix} \tilde{B} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1} \tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}' \end{bmatrix} \tilde{Y} = \tilde{A} \left(\lambda, \boldsymbol{\theta} \right) \tilde{Y}. \tag{33}$$

where

$$\begin{aligned}\tilde{A}(\lambda,\boldsymbol{\theta}) &= \begin{bmatrix}\tilde{B} & \tilde{Q}\end{bmatrix}\tilde{C}^{-1}\tilde{C}^{-T}\begin{bmatrix}\tilde{B}'\\ \tilde{Q}'\end{bmatrix} \\ &= G + (I-G)\,\tilde{Q}\left[\tilde{Q}'(I-G)\,\tilde{Q}+\lambda Q\right]^{-1}\tilde{Q}'(I-G),\end{aligned} \tag{34}$$

for

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

1.2.2 Example: m^{th} order Sobolev space, $W_m(0,1)$

We first consider Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ be the reproducing kernel Hilbert space (r.k.h.s) corresponding to the tensor product of the first-order and second-order Sobolev spaces:

$$\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m, \quad \mathcal{H}_l = W_2(0,1), \quad \mathcal{H}_m = W_1(0,1) \text{ where}$$

$$W_m(0,1) \equiv \{f: \quad f', \ldots, f^{(m-1)} \text{absolutely continuous}, \int_0^1 \left(f^{(m)}\right)^2 dt < \infty\}$$

We seek $(\cdot,\cdot) \in \mathcal{H}$ which minimizes

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=2}^{n_i}\sigma_{ij}^{-2}\left(y(t_{ij}) - \sum_{k=1}^{n_i-1}\phi\left(l_{jk}^i, m_{jk}^i\right)y(t_{ik})\right)^2 + \lambda J(\phi) \tag{35}$$

where $P_1\phi$ is the projection of ϕ onto $\mathcal{H}_1$, $J(\phi) = ||P_1\phi||^2$. Define the differential operator $M_\nu f = \int_0^1 f^{(m)}(x)\,dx$, $\nu = 1, \ldots, m$ and endow $W_m(0,1)$ with inner product

$$\langle f, g\rangle = \langle f, g\rangle_0 + \langle f, g\rangle_1 = \sum_{\nu=0}^{m-1} M_\nu f M_\nu g + \int_0^1 f^{(m)}(x)\,g^{(m)}(x)\,dx \tag{36}$$

which induces norm

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

Let $k_j(x) = B_j(x)/j!$ for $x \in [0,1]$, where $B_j(x)$ is the j^{th} Bernoulli polynomial which can be defined according to the recursive relationship:

$$B_0(x) = 1, \qquad \frac{d}{dx}B_r(x) = rB_{r-1}(x)$$

Noting that $M_\nu B_r = \delta_{\nu-r}$, W_m can be written as a direct sum of the m orthogonal subspaces: $\{k_r\}_{r=0}^{m-1}$ and W_m^1. Here, $\{k_r\}$ is the subspace spanned by k_r and W_m^1 is the space orthogonal to $W_m^0 \equiv \{1\} \oplus \{k_1\} \oplus \cdots \oplus \{k_{m-1}\}$ which satisfies

$$W_m^1 = \{f : M_\nu f = 0, \quad \nu = 0, 1, \ldots, m-1\}$$

Writing $\mathcal{H}$ as the tensor product of the two decomposed Sobolev spaces, we have

$$\begin{aligned}
\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m &= W_2 \otimes W_1 \\
&= \left[W_2^0 \oplus W_2^1\right] \otimes \left[W_1^0 \oplus W_1^1\right] \\
&= \left[\left[\{1\} \oplus \{k_1\}\right] \oplus W_2^1\right] \otimes \left[\{1\} \oplus W_1^1\right] \\
&= \left[\{1\} \oplus \{k_1\}\right] \oplus W_2^1 \oplus W_1^1 \oplus \left[\{k_1\} \otimes W_1^1\right] \oplus \left[W_2^1 \otimes W_1^1\right] \\
&\equiv \left[\mathcal{H}_{\mu^*} \oplus \mathcal{H}_l^0\right] \oplus \left[\mathcal{H}_l^1 \oplus \mathcal{H}_m^1 \oplus \mathcal{H}_{lm}^{01} \oplus \mathcal{H}_{lm}^{11}\right] \\
&= \mathcal{H}_0 \oplus \mathcal{H}_1
\end{aligned} \tag{37}$$

where the functional components corresponding to $\mathcal{H}_\mu$, $\mathcal{H}_l^0$, $\mathcal{H}_l^1$, $\mathcal{H}_m^1$, and $[\mathcal{H}_{lm}^{01} \oplus \mathcal{H}_{lm}^{11}]$ are the overall mean, the nonparametric main effect of l, the parametric main effect of l, the parametric main effect of m, the nonparametric-parametric interaction, and the parametric-parametric interaction (between l and m). Given this decomposition of the function space, any $\phi \in \mathcal{H}$ may be written as a sum of components from each of the

$$\phi(l,m) = \mu^* + \phi_l^*(l) + \phi_m^*(m) + \phi_{lm}^*(l,m) \tag{38}$$

where $\int_0^1 \phi_l(l)\, dl = \int_0^1 \phi_m(m)\, dm = 0$, $\int_0^1 \phi_{lm}(l,m)\, dl = \int_0^1 \phi_{lm}(l,m)\, dm = 0$. The reproducing kernel (r.k.) for $\{k_r\}$ is $k_r(x)\, k_r(x')$. It can be verified that the r.k. for W_m^1 (Craven and Wahba 1979) is given by $R^1(x,x') = k_m(x)\, k_m(x') + (-1)^{m-1} k_{2m}([x-x'])$ where $[\alpha]$ is the fractional part of α. The r.k. for W_m is given by

$$\begin{aligned}
R(x,x') &= R^0(x,x') + R^1(x,x') \\
&= \left[\sum_{\nu=1}^{m-1} k_\nu(x)\, k_\nu(x')\right] + \left[k_m(x)\, k_m(x') + (-1)^{m-1} k_{2m}([x-x'])\right]
\end{aligned}$$

Using the fact that the r.k. for a tensor product space is the product of the corresponding reproducing kernels, the r.k. for $\mathcal{H}$ is given by

$$\begin{aligned}
R((l,m),(l',m')) &= R_l(l,l') \times R_m(m,m') \\
&= \left[R_l^0(l,l') + R_l^1(l,l')\right] \times \left[R_m^0(l,l') + R_m^1(l,l')\right] \\
&= R_l^0(l,l')\, R_m^0(m,m') + R_l^0(l,l')\, R_m^1(m,m') \\
&\qquad + R_l^1(l,l')\, R_m^0(m,m') + R_l^1(l,l')\, R_m^1(m,m') \\
&= \left[k_1(l)\, k_1(l')\right] + \left[R_l^1(l,l') + k_1(l,l')\, R_m^1(m,m') + R_l^1(l,l')\, R_m^1(m,m')\right] \\
&= R^0((l,m),(l',m')) + R^1((l,m),(l',m'))
\end{aligned} \tag{39}$$

We must introduce some notation to simplify the following expression of the form of the elements in $\mathcal{H}$. Denote the set of unique pairs of observed within-subject time points and the corresponding set of unique transformed coordinates by $\mathcal{W}$ and $\mathcal{W}^*$, respectively:

$$\begin{aligned}\mathcal{W} &= \bigcup_{i=1}^{N}\bigcup_{j>k}(t_{ij}, t_{ik}) \\ \mathcal{W}^* &= \bigcup_{i=1}^{N}\bigcup_{j>k}\left(t_{ij} - t_{ik}, \frac{1}{2}(t_{ij} + t_{ik})\right) = \bigcup_{i=1}^{N}\bigcup_{j>k}(l_{jk}^i, m_{jk}^i)\end{aligned}$$

with $|\mathcal{W}| = |\mathcal{W}^*| = N_\phi$. For simplicity of presentation, relabel the elements of $\mathcal{W}^*$ so that

$$\mathcal{W}^* = \{(l_1, m_1), (l_2, m_2), \ldots, \left(l_{N_\phi}, m_{N_\phi}\right)\}$$

One may verify that any $\phi \in \mathcal{H}$ can be written

$$\phi(l, m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{n} c_i R_1((l, m), (l_i, m_i)) + \rho(l, m) \tag{40}$$

where $\rho \perp \mathcal{H}_0 = \{1\} \oplus \{k_1\}$, span$\{R_1((l_i, m_i), \cdot)\}$. It can be shown that the minimizer of **??** has $\rho = 0$, so that the $\phi \in \mathcal{H}$ minimizing **??** can be written as a (finite) linear combination of inner products:

$$\phi(l, m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{n} c_i R_1((l, m), (l_i, m_i)) \tag{41}$$

The proof entails demonstrating that ρ does not improve the first term in (**??**) (the data fit functional) and only adds to the penalty term, $J(\phi)$. Details are left to the appendix **??**. Let Φ be the $N_\phi \times 1$ vector of regression coefficients given by (**??**) corresponding to ϕ evaluated at the elements of $\mathcal{W}^*$, $\phi = \left(\phi_1, \phi_2, \ldots, \phi_{N_\phi}\right)^T$. Let $d = (d_0, d_1)^T$, $c = \left(c_1, \ldots, c_{N_\phi}\right)^T$, and $b = (b_1, \ldots, b_{N_m})^T$. Define K_{11}, K_{12}, K_{22}, B_1, and B_2 as follows:

$$\begin{array}{ll}
K_{11}[i, j] = R_1((l_i, m_i), (l_j, m_j)) & i, j = 1, \ldots, N_\phi \\
K_{12}[i, j] = R_1((l_i, m_i), (0, m_j)) & i = 1, \ldots, N_\phi,\ j = 1, \ldots, N_m \\
K_{22}[i, j] = R_1((0, m_i), (0, m_j)) & i, j = 1, \ldots, N_m \\
B_1[i, j] = k_j(l_i) & i = 1, \ldots, N_\phi,\ j = 1, 2 \\
B_2[i, j] = k_j(0) & i = 1, \ldots, N_m,\ j = 1, 2
\end{array}$$

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{12}^T & K_{22} \end{bmatrix} = \begin{bmatrix} K_1 \\ K_2 \end{bmatrix}; \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

In matrix notation:

$$\phi = Sd + Rc$$

1.2.3 Smoothing parameter selection

The smoothing matrix $\tilde{A}$ plays an integral role of calculating all of the model selection criteria we will discuss in the sections to follow; the diagonal elements of $\tilde{A}$, $\tilde{a}_{kk}$ are of particular importance in quantifying model complexity. In classical regression theory, the degrees of freedom are clearly defined as the number of variables included in the model. **?** and**?** refer to this measure of model complexity as the model's *effective dimension* ED; they follow **?**, who discuss the effective dimensions of linear smoother and propose to use the trace of the smoother matrix as an approximation.

2 Model selection criteria

By varying smoothing parameters λ and θ_β, the minimizer ϕ^*_λ of **??** 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 three criteria that are commonly used for smoothing parameter selection in the context of smoothing spline models. The first score is an unbiased estimate of a relative loss and assumes a known variances σ_t^2. The second score, the generalized cross validation (GCV) score of **?**, provides an estimate of the same loss without assuming a known variance function. These scores have attractive asymptotic properties; see **?** for a comprehensive examination. To simplify presentation for the initial presentation, we only make explicit the dependence of estimates and their components on λ and conceal any dependence on θ_β.

2.1 Unbiased risk estimate

We can write

$$\tilde{Y} = D^{-1/2} W \Phi + \tilde{\epsilon}, \tag{42}$$

where

$$\Phi^* = \left(\phi^*\left(v_{121}\right), \phi^*\left(v_{131}\right), \ldots, \phi^*\left(v_{N,m_N,m_n-1}\right)\right)'$$

denotes the vector holding the values of ϕ^* evaluated at the observed within-subject pairs of time points, and $\tilde{\epsilon} = D^{-1/2}\epsilon$ where ϵ is the vector of $\sum_{i=1}^N m_i - N$ associated prediction errors. 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

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{N} \sum_{j=1}^{m_i} \left(\hat{\tilde{y}}_{ij} - E\left[\tilde{y}_{ij}\right]\right)^2 \\ &= ||\tilde{Y} - \tilde{\mu}||^2 \end{aligned} \tag{43}$$

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

$$L(\lambda) = \mu' \left(I - \tilde{A}\right)^2 \mu - 2\mu' \left(I - \tilde{A}\right)^2 \tilde{A}\tilde{\epsilon} + \tilde{\epsilon}' \tilde{A}^2 \tilde{\epsilon} \tag{44}$$

Define the unbiased risk estimate

$$U(\lambda) = \tilde{Y}' \left(I - \tilde{A}\right)^2 \tilde{Y} + 2\text{tr}\tilde{A} \tag{45}$$

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

$$\begin{aligned} U(\lambda) &= \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y}\right)' \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y}\right) + 2\text{tr}\tilde{A} \\ &= \left(\tilde{A}\tilde{Y} - \mu\right)' \left(\tilde{A}\tilde{Y} - \mu\right) + \tilde{\epsilon}'\tilde{\epsilon} + 2\tilde{\epsilon}' \left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right) \end{aligned} \tag{46}$$

This gives

$$U(\lambda) - L(\lambda) - \tilde{\epsilon}'\tilde{\epsilon} = 2\tilde{\epsilon}' \left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right), \tag{47}$$

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 **?** Chapter 3 for a formal theorem and proof.

2.2 Leave-one-out and generalized cross validation

The use of the unbiased risk estimate $U(\lambda)$ to select the optimal smoothing parameter requires knowledge of the innovation variance $\sigma(t)^2$, which is, in practice, unknown and we can at best approximate with an estimate. An alternative for selecting λ is cross validation; it and its variants have long been utilized for smoothing parameter selection in spline models, and their properties have been studied extensively. A short list of supplemental references include **?**, **?**, **?**, **?**, and **?**. There are a number of ways to calculate a measure of cross validated prediction error; we first focus on the leave-one-out method. Let $\hat{\tilde{y}}_{ij}^{[-ij]}$ denote the predicted value for the observation $\tilde{y}_{ij}$ when $\tilde{y}_{ij}$ itself is removed from the data used for fitting the model. We can calculate these predictions for each observation in the data set to obtain the leave-one-out (LOO) cross validation score:

$$V_0(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=2}^{m_i} \frac{1}{m_i} \left(y_{ij} - \hat{y}_{ij}^{[-ij]}\right)^2, \tag{48}$$

Brute force calculation of **??** is generally impractical, especially if the number of observations is large. However, this labor can be sidestepped using the following fact:

$$\widehat{\tilde{Y}} = \tilde{A}\tilde{Y}$$

With some abuse of notation, let $\tilde{y}_k$ denote the k^{th} element of the full vector of responses $\tilde{Y}$, for $k = 1, \ldots, \sum_i m_i - N$. One can show that

$$y_k - \hat{y}_k^{[-k]} = (y_k - \hat{y}_k) / (1 - \tilde{a}_{kk}), \tag{49}$$

where $\{\tilde{a}_{kk}\}$ denote the diagonal elements of the smoothing matrix $\tilde{A}$, which can be calculated quickly. An informal proof of is as follows: suppose that we change the i^{th} element of $\tilde{Y}$, obtaining a new response vector $\tilde{Y}^*$. Then $\hat{\tilde{Y}}^* = \tilde{A}\tilde{Y}^*$. Since

$$\hat{\tilde{y}}_k = \sum_l \tilde{a}_{kl}\tilde{y}_l$$

and

$$\hat{\tilde{y}}_{-k} = \sum_l \tilde{a}_{ij}\tilde{y}_j^*,$$

we have that

$$\hat{\tilde{y}}_k - \hat{\tilde{y}}_{-k} = \sum_l \tilde{a}_{kl}\left(\tilde{y}_k - \tilde{y}_k^*\right) = \tilde{a}_{kk}\left(\tilde{y}_k - \tilde{y}_k^*\right).$$

With this, **??** can be rewritten as

$$\begin{aligned} V_0(\lambda) &= \frac{1}{\sum_i m_i - N} \sum_{k=1}^{\sum_i m_i - N} y_k - \hat{y}_k^{[-k]} \\ &= \frac{1}{\sum_i m_i - N} \sum_{k=1}^{\sum_i m_i - N} \left(\tilde{y}_k - \hat{\tilde{y}}_k\right)^2 / (1 - \tilde{a}_{kk})^2, \end{aligned}$$

The best λ is the value that minimizes the cross validation score. The leave-one-out cross validation score weighs all data points equally in the estimate of prediction error. However, one may wish to adjust for any imbalance in the contribution of the t_{ij} to the estimate of ϕ, which can be done by simply weighting observations when averaging the prediction errors:

$$\bar{V}(\lambda) = \frac{1}{\sum_i m_i - N} \sum_{k=1}^{\sum_i m_i - N} \tilde{y}_k - \hat{\tilde{y}}_k^{[-k]} = \omega_k \left(\tilde{y}_k - \hat{\tilde{y}}_k\right)^2 / (1 - \tilde{a}_{kk})^2. \tag{50}$$

We obtain Craven and Wahba's generalized cross validation score (GCV) if we take

$$\omega_i = (1 - \tilde{a}_{kk})^2 \Big/ \left[\frac{\text{tr}\left(I - \tilde{A}(\lambda)\right)}{\sum_i m_i - N}\right]^2,$$

which is equivalent to substituting $\tilde{a}_{kk}$ in **??** for the average $(\sum_i m_i - N)^{-1} \sum_{k=1}^{\sum_i m_i - N} \tilde{a}_{kk}$. Under mild conditions, the GCV score is a consistent estimator of relative loss **??**, see **?** for detailed discussion.

2.3 Leave-one-subject-out cross validation

The conditions under which the the cross validation and GCV scores 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

$$V_{loso}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{\tilde{\mu}}_i^{[-i]}\right)'\left(Y_i - \widehat{\tilde{\mu}}_i^{[-i]}\right) \tag{51}$$

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 $Y_i^* = \left(y_{i1}^*, y_{i1}^*, \ldots, y_{i,m_i}^*\right)$

[Present heuristic argument here. See Xu, G., Huang, J. Z. (2012). Asymptotic optimality and efficient computation of the leave-subject-out cross-validation. The Annals of Statistics, 40(6), 3003-3030.]

$$\begin{aligned} MSPE &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - \widehat{\tilde{\mu}}_i||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - D_i^{-1/2}W_i\Phi^* + D_i^{-1/2}W_i\Phi^* - D_i^{-1/2}W_i\hat{\Phi}^*||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\tilde{\mu}_i - \widehat{\tilde{\mu}}_i^{[-i]}||^2\right]\right\} \end{aligned} \tag{52}$$

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

$$E\left[V_{loso}(\lambda)\right] = \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\widehat{\tilde{\mu}}_i - \tilde{\mu}_i||^2\right]\right\}. \tag{53}$$

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

2.3.1 Formal justification for LosoCV by showing that it is asymptotically equivalent to loss function when appropriately defined

2.3.2 Regularity conditions necessary for asymptotic properties of LosoCV score to hold

2.3.3 Optimality of LosoCV score

2.3.4 Computation of the LosoCV score

Lemma 2.1 (Shortcut formula for LosoCV)**.** *The LosoCV score satisfies the following identity:*

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

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.

See **?** and supplementary materials **?** for a detailed presentation and proof.

Proof. For fixed λ, $\boldsymbol{\theta}$, $\sigma^2(t)$, let $\hat{a}^{[-i]} = \left(\hat{d}^{[-i]}, \hat{c}^{[-i]}\right)$ denote the minimizer of the penalized log likelihood □

2.3.5 Approximation of leave-one-subject-out cross validation

? 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}:

$$V^*_{loso}(\lambda) = \frac{1}{N}||\left(I - \tilde{A}\right)\tilde{Y}||^2 + \frac{2}{N}\sum_{i=1}^{N}\hat{\tilde{e}}_i'\tilde{A}_{ii}\hat{\tilde{e}}_i, \tag{54}$$

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.

Theorem 2.2. *Under conditions 1-5, for predetermined $\sigma^2(t)$ and nonrandom λ,*

$$V_{loso}(\lambda) - L(\lambda) - \frac{1}{N}\epsilon'\epsilon - o_p(L(\lambda)). \tag{55}$$

as $n \to \infty$

2.4 Selection of multiple smoothing parameters

The expression in **??** permits the straightforward evaluation of the GCV score

$$V(\lambda, \boldsymbol{\theta}) = \frac{(1/n_y) \left\| \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \tilde{Y} \right\|^2}{\left[(1/n_y) \operatorname{tr} \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \right]^2} \tag{56}$$

and the GML score

$$M(\lambda, \boldsymbol{\theta}) = \frac{(1/n_y) \tilde{Y}' \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \tilde{Y}}{\left[\det{}^{+} \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \right]^{1/n_y}}. \tag{57}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_g)'$ denotes the vector of smoothing parameters associated with each RK. To minimize the functions $V(\lambda, \boldsymbol{\theta})$ and $M(\lambda, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and λ, we iterate as follows:

I. Fix $\boldsymbol{\theta}$; minimize $V(\lambda|\boldsymbol{\theta})$ or $M(\lambda|\boldsymbol{\theta})$ 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 $V(\boldsymbol{\theta}|\lambda)$ or $M(\boldsymbol{\theta}|\lambda)$ with respect to $\boldsymbol{\kappa} = \log(\boldsymbol{\theta})$. 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 V and M 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 $V(\cdot)$ and $M(\cdot)$ 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.

2.5 Algorithms

The main algorithm and discussion of its key components are presented in the section to follow. The minimization of the model selection criterion is done via two nested loops. Fixing tuning parameters, the outer loop minimizes V (or M) 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 with step-halving as safeguards. Fixing the θ_βs in $J(\phi^*) = \sum_\beta \theta_\beta^{-1} J_\beta(\phi_\beta^*)$, the outer loop with a single λ is a straightforward task.

Algorithm 1

Initialization:

Set $\Delta\boldsymbol{\kappa} := 0;\ \boldsymbol{\kappa}_{-} := \boldsymbol{\kappa}_0;\ V_{-} = \infty;\ (\text{ or } M_{-} = \infty)$

Iteration:

while not converged **do**

For current value $\boldsymbol{\kappa}_{*} = \boldsymbol{\kappa}_{-} + \Delta\boldsymbol{\kappa}$, compute $Q_{*}^{\theta} = \sum_{\beta=1}^{g} \theta_{\beta} Q_{\beta}$.

Compute $\tilde{A}\left(\lambda|\boldsymbol{\theta}_{*}\right) = \tilde{A}\left(\lambda, \exp\left(\boldsymbol{\kappa}_{*}\right)\right)$.

Minimize

$$V\left(\lambda|\boldsymbol{\kappa}_{*}\right) = \frac{\left(1/n_y\right)\left\|\left(I - \tilde{A}\left(\lambda|\boldsymbol{\theta}_{*}\right)\right)\tilde{Y}\right\|^2}{\left[\left(1/n_y\right)\text{tr}\left(I - \tilde{A}\left(\lambda|\boldsymbol{\theta}_{*}\right)\right)\right]^2}$$

or

$$M\left(\lambda|\boldsymbol{\kappa}_{*}\right) = \frac{\left(1/n_y\right)\tilde{Y}'\left(I - \tilde{A}\left(\lambda|\boldsymbol{\theta}_{*}\right)\right)\tilde{Y}}{\left[\det{}^{+}\left(I - \tilde{A}\left(\lambda|\boldsymbol{\theta}_{*}\right)\right)\right]^{1/n_y}}.$$

Set

$$V_{*} := \min_{\lambda} V\left(\lambda|\boldsymbol{\kappa}_{*}\right)$$

$$\left(M_{*} := \min_{\lambda} M\left(\lambda|\boldsymbol{\kappa}_{*}\right)\right)$$

if $V_{*} > V_{-}$ (or $M_{*} > M_{-}$) **then**

Set $\Delta\boldsymbol{\kappa} := \Delta\boldsymbol{\kappa}/2$

Go to (1).

else

Continue

end if

Evaluate gradient $\mathbf{g} = \left(\partial/\partial\boldsymbol{\kappa}\right)V\left(\boldsymbol{\kappa}|\lambda\right)$ (or $\left(\partial/\partial\boldsymbol{\kappa}\right)M\left(\boldsymbol{\kappa}|\lambda\right)$)

Evaluate Hessian $H = \left(\partial^2/\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}'\right)V\left(\boldsymbol{\kappa}|\lambda\right)$ (or $\left(\partial^2/\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}'\right)M\left(\boldsymbol{\kappa}|\lambda\right)$).

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 $\tilde{H} = \text{diag}\left(\boldsymbol{\epsilon}\right)$ is positive definite.

end if

end while

Calculate optimal model:

if $\Delta\kappa_{\beta} < -\gamma$, for γ large **then**

Set $\kappa_{*\beta} := -\infty$

end if

Compute $Q_{*}^{\theta} = \sum_{\beta=1}^{g} \theta_{*\beta} Q_{\beta}$;

Calculate $\begin{bmatrix} d \\ c \end{bmatrix} = \tilde{C}^{-1}\tilde{C}^{-T}\begin{bmatrix} \tilde{B}' \\ \tilde{Q}_{*}^{\theta'} \end{bmatrix}\tilde{Y}$

The update direction $\Delta\boldsymbol{\kappa} = -\tilde{H}^{-1}\mathbf{g}$ is calculated via the modified Newton method on the modified Cholesky decomposition given in **??**. Detailed discussion can be found in **?**.

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 $V(\lambda)$ with respect to λ to obtain $\breve{\phi}^*$.

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

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 $\check{\theta}_\beta$. These are the starting values adopted by **?**; the starting values for the first pass loop are somewhat arbitrary, but are invariant to scalings of the θ_β. The starting values in **??** for the second pass of the outer are based on more involved assumptions derived from the background formulation of the smoothing problem. See **?** for a detailed discussion.

TO DO: Outline the argument for using the starting values $\breve{\theta}_\beta$

2.6 Algorithm

2.6.1 Computation of the gradient and the Hessian of $V(\lambda)$

2.6.2 Starting values for the Newton iteration

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

Recall that the joint likelihood of the data $Y_1, \ldots, Y_N$ is satisfies

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

Let

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

denote the squared residuals for the observations y_{ij} having corresponding measurement time $t_{ij} - t$. Then $\text{RSS}(t)/\sigma^2(t) \sim \chi^2_{df_t}$, 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 **??** 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.

¡¡¡¡¡¡¡ HEAD =======

¿¿¿¿¿¿¿ cc10e8225503967265891c7b98fd982e18d01ca5