Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

January 9, 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. A commonly utilized approach in previous work is to model $\phi_{ijk} = z_{ijk}^T \gamma$ where z_{ijk} is a vector of powers of time differences and γ is a vector of unknown "dependence" parameters to be estimated from the data. **?**, **?**, **?**, and **?** define

$$z_{ijk}^T = \left(1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2, \ldots, (t_{ij} - t_{ik})^{q-1}\right) \tag{1}$$

Modeling the covariance in such a way is reduces a potentially high dimensional problem to something much more computationally feasible; if one models the innovation variances $\sigma^2(t)$ similarly using a d-dimensional vector of covariates, the problem reduces to estimating $q + d$ unconstrained parameters, where much of the dimensionality reduction is a result of characterizing the GARPs in terms of only the difference between pairs of observed time points, and not the time points themselves. Modeling ϕ in such a way is equivalent to specifying a Toeplitz structure for Σ. A $p \times p$ Toeplitz matrix M is a matrix with elements m_{ij} such that $m_{ij} = m_{|i-j|}$ i.e. a matrix of the form

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

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \dots & m_{p-1} \\ m_1 & m_0 & m_1 & \dots & m_{p-2} \\ m_2 & m_1 & m_0 & \dots & m_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p-1} & m_{p-2} & m_{p-3} & \dots & m_0 \end{bmatrix} \tag{2}$$

The estimated covariance matrix may be considerably biased when the specified parametric model is far from the truth. To avoid model misspecification that potentially accompanies parametric analysis, many have alternatively proposed nonparametric and semiparametric techniques approaches to estimation. While these estimators can be very flexible and thus exhibit low bias, this advantage can be offset with high variance. To balance the tradeoff between bias and variance, shrinkage or regularization may be applied to estimates to improve stability of estimators. **?** proposed nonparametric estimation of the covariance matrix of longitudinal data by smoothing raw sample variogram ordinates and squared residuals. [DISCUSS THE NONPARAMETRIC SMOOTHER OF HANS GEORG MULLER HERE] However, neither of these methods ensure that the resulting estimates are positive-definite.

Several others have proposed methods for covariance estimation within the same paradigm of a smooth, continuous function underlying a discretized covariance matrix associated with the observed data. **?** employ the Cholesky decomposition to guarantee positive-definiteness and imposed structure on the elements of the Cholesky decomposition and heuristically argue that $\phi_{t,t-l}$ should be monotonically decreasing in l. That is, the effect of y_{t-l} on y_t through the autoregressive parameterization should decrease as the distance in time between the two measurements increases. In similar spirit, others including **?** and **?** enforce such structure by setting $\phi_{t,t-l}$ equal to zero for l large enough, or equivalently, setting all subdiagonals of T to zero beyond the K^{th} off-diagonal. The tuning parameter K is chosen using a model selection criterion such as Akaike information criterion, Bayesian information criterion, or cross validation or a variant thereof. In terms of the autoregressive model corresponding to the Cholesky decomposition, this form of regularization, known as "banding" the Cholesky factor T, is equivalent to regressing y_t on only its K immediate predecessors, setting $\phi_{tj} = 0$ for $t - j > K$.

From this perspective, it is apparent that the presentation of covariance estimation as a least squares regression problem suggests that the familiar ideas of model regularization for least-squares regression can be used for estimating covariances. Wu and Pourahmadi **?** proposed a two-step estimation procedure using nonparametric smoothing for regularized estimation of large covariance matrices. In the first step, they derive a raw estimate of the covariance matrix and the estimated covariance matrix is subject to the modified Cholesky decomposition. In the second step, they apply local polynomial smoothing to the diagonal elements of D and the subdiagonals of T. The use of the Cholesky parameterization guarantees that their estimate is guaranteed to be positive-definite, however, their procedure is not capable of handling missing data. **?**

however, their two-step method did not utilize the information that many of the subdiagonals of T are essentially zeros at the first step. Inefficient estimation may result because of ignoring

regularization structure in constructing the raw estimator.

Several have applied these approaches to covariance estimation; **?** jointly model the mean and covariance matrix of longitudinal data using basis function expansions. They employ the Cholesky decomposition of the covariance matrix and treat the subdiagonals of T as smooth functions, approximated by B-splines. Estimation is carried out by maximizing the normal likelihood. Their method permits subject-specific observations times, but assumes that observation times lie on some notion of a regular grid. They treat within-subject gaps in measurements as missing data and which they handle using the E-M algorithm.

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.

1.1 Smoothing spline representation of ϕ, σ

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 $\phi^*(\cdot,\cdot) \in \mathcal{H}$ which minimizes

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=2}^{n_i}\sigma_{ij}^{-2}\left(y(t_{ij}) - \sum_{k=1}^{n_i-1}\phi^*\left(l_{jk}^i, m_{jk}^i\right)y(t_{ik})\right)^2 + \lambda J(\phi^*) \tag{3}$$

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

$$\langle f, g\rangle = \langle f, g\rangle_0 + \langle f, g\rangle_1 = \sum_{\nu=0}^{m-1} M_\nu f M_\nu g + \int_0^1 f^{(m)}(x)\, g^{(m)}(x)\, dx \tag{4}$$

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:

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

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

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

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

$$\begin{aligned}
\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m &= W_2 \otimes W_1 \\
&= \left[W_2^0 \oplus W_2^1\right] \otimes \left[W_1^0 \oplus W_1^1\right] \\
&= \left[[\{1\} \oplus \{k_1\}] \oplus W_2^1\right] \otimes \left[\{1\} \oplus W_1^1\right] \\
&= [\{1\} \oplus \{k_1\}] \oplus W_2^1 \oplus W_1^1 \oplus \left[\{k_1\} \otimes W_1^1\right] \oplus \left[W_2^1 \otimes W_1^1\right] \\
&\equiv \left[\mathcal{H}_{\mu^*} \oplus \mathcal{H}_l^0\right] \oplus \left[\mathcal{H}_l^1 \oplus \mathcal{H}_m^1 \oplus \mathcal{H}_{lm}^{01} \oplus \mathcal{H}_{lm}^{11}\right] \\
&= \mathcal{H}_0 \oplus \mathcal{H}_1
\end{aligned} \tag{5}$$

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

$$\phi^*(l, m) = \mu^* + \phi_l^*(l) + \phi_m^*(m) + \phi_{lm}^*(l, m) \tag{6}$$

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

$$\begin{aligned}
R((l, m), (l', m')) &= R_l(l, l') \times R_m(m, m') \\
&= \left[R_l^0(l, l') + R_l^1(l, l')\right] \times \left[R_m^0(l, l') + R_m^1(l, l')\right] \\
&= R_l^0(l, l')\, R_m^0(m, m') + R_l^0(l, l')\, R_m^1(m, m') \\
&\qquad + R_l^1(l, l')\, R_m^0(m, m') + R_l^1(l, l')\, R_m^1(m, m') \\
&= [k_1(l)\, k_1(l')] + \left[R_l^1(l, l') + k_1(l, l')\, R_m^1(m, m') + R_l^1(l, l')\, R_m^1(m, m')\right] \\
&= R^0((l, m), (l', m')) + R^1((l, m), (l', m'))
\end{aligned} \tag{7}$$

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

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

Then we may verify that any $\phi^* \in \mathcal{H}$ can be written

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

where $\rho \perp \mathcal{H}_0 = \{1\} \oplus \{k_1\},\ span\{R_1((l_i, m_i), \cdot)\}$. We do so by demonstrating that ρ does not improve the first term in (**??**) (the data fit functional) and only adds to the penalty term, $J(\phi^*)$. Consequently, if $\hat{\phi}^*$ is the minimizer of (**??**), then $\rho = 0$. Using the properties of reproducing kernels, we can rewrite ϕ^* as an inner product of itself with R:

$$
\begin{aligned}
\phi^{*}\left(l_j, m_j\right) &= \left\langle R\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \phi^{*}(\cdot,\cdot)\right\rangle \\
&= \Big\langle R_0\left(\left(l_j, m_j\right),(\cdot,\cdot)\right)+R_1\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), d_0+d_1 k_1(\cdot) \\
&\qquad\qquad +\sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right),(\cdot,\cdot)\right)+\rho\left((\cdot,\cdot)\right)\Big\rangle \\
&= \left\langle R_0\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), d_0+d_1 k_1(\cdot)\right\rangle+\left\langle R_0\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right),(\cdot,\cdot)\right)\right\rangle \\
&\quad +\left\langle R_0\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \rho\left((\cdot,\cdot)\right)\right\rangle+\left\langle R_1\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), d_0+d_1 k_1(\cdot)\right\rangle \\
&\quad +\left\langle R_1\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right),(\cdot,\cdot)\right)\right\rangle+\left\langle R_1\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \rho\left((\cdot,\cdot)\right)\right\rangle \\
&= \left\langle R_0\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), d_0+d_1 k_1(\cdot)\right\rangle+\left\langle R_1\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right),(\cdot,\cdot)\right)\right\rangle \\
&\quad +\underbrace{\left\langle R_0\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \rho(\cdot,\cdot)\right\rangle}_{0}+\underbrace{\left\langle R_1\left(\left(l_j, m_j\right),(\cdot,\cdot)\right), \rho(\cdot,\cdot)\right\rangle}_{0} \\
&= d_0+d_1 k_1(\cdot)+\sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right),\left(l_j, m_j\right)\right)
\end{aligned}
$$

Rewriting the data fit functional, we have that

$$
\begin{aligned}
&\sum_{i=1}^{N} \sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y\left(t_{ij}\right)-\sum_{k=1}^{j-1} \phi^{*}\left(t_{ij}, t_{ik}\right) y\left(t_{ik}\right)\right)^2 \\
= &\sum_{i=1}^{N} \sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y\left(t_{ij}\right)-\sum_{k=1}^{j-1}\left\langle R\left(\left(l_{jk}^{i}, m_{jk}^{i}\right),(\cdot,\cdot)\right), \phi^{*}(\cdot,\cdot)\right\rangle y\left(t_{ik}\right)\right)^2
\end{aligned}
$$

which is free of ρ. Consider the contribution of any nonzero ρ to $J\left(\phi^{*}\right)$:

$$
\begin{aligned}
J\left(\phi^{*}\right) &= \left|\left|P_1 \phi^{*}\right|\right|^2 \\
&= \left\langle \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right),(\cdot,\cdot)\right)+\rho(\cdot,\cdot), \sum_{j=1}^{N_{\phi^*}} c_j R_1\left(\left(l_j, m_j\right),(\cdot,\cdot)\right)+\rho(\cdot,\cdot)\right\rangle \\
&= \left|\left| \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right),(\cdot,\cdot)\right)\right|\right|^2+||\rho||^2
\end{aligned}
$$

Thus, including ρ in ϕ^* only increases the penalty without improving (decreasing) the data fit functional, so we indeed have that the minimizer of (**??**) has the form

$$\phi^*(l,m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{N_{\phi^*}} c_i R_1((l,m),(l_i,m_i)) \tag{8}$$

1.2 Penalized likelihood estimation

Let Y hold the N observed response vectors $y_1, \ldots, y_N$ less their first element y_{i1} stacked into a single vector of dimension $n_y = \left(\sum_i M_i\right) - N$. Let M denote the total number of distinct observation times across all subjects. For ease of exposition, let $\sigma_{ij} = \sigma(t_{ij})$ and $\phi_{ijk} = \phi(t_{ijk})$. The loglikelihood **??** becomes

$$\begin{aligned}
-2\ell(Y,\Sigma) &= \sum_{t=1}^{M} \log\sigma_t^2 + \sum_{i=1}^{N}\sum_{j=1}^{m_i} \frac{\epsilon_{ijk}^2}{\sigma_{ij}^2} \\
&= \sum_{t=1}^{M} \log\sigma_t^2 + \sum_{i=1}^{N} \frac{\epsilon_{i1}^2}{\sigma_{i1}^2} + \sum_{i=1}^{N}\sum_{j=2}^{m_i} \frac{\epsilon_{ij}^2}{\sigma_{ij}^2} \qquad (9) \\
&= \sum_{t=1}^{M} \log\sigma_t^2 + \sum_{i=1}^{N} \frac{y_{i1}^2}{\sigma_{i1}^2} + \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma_{ij}^{-2}\left(y_{ij} - \sum_{k<j}\phi_{ijk}y_{ik}\right)^2.
\end{aligned}$$

$$\sum_{t=1}^{M} \log\sigma_t^2 + \sum_{i=1}^{N} \frac{y_{i1}^2}{\sigma_{i1}^2} + \sum_{i=1}^{N}\sum_{j=2}^{m_i} \frac{\epsilon_{ij}^2}{\sigma_{ij}^2} \tag{10}$$

For ease of exposition, we assume that $\sigma^2(t)$ is fixed and known.