Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

January 19, 2018

0.1 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

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

An iterative procedure for minimizing **??** starts by first initializing σ_{ij} using, for example, the innovation standard error estimated without the penalty. Then we minimize

$$\sum_{i=1}^{N} \sum_{j=2}^{m_i} \frac{\epsilon_{ij}^2}{\sigma_{ij}^2} \tag{2}$$

to obtain $\tilde{\phi}^*(t, s)$. We then obtain an estimate for $\sigma(t)$ by fixing $\phi^* = \tilde{\phi}^*$ and minimizing 1. We iterate this process until convergence of the estimated coefficient vectors.

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

1 Computation of the smoothing spline estimator

The minimization of the penalized smoothing spline log likelihood

$$\begin{aligned} &-2\ell\left(Y|c,d\right)+\lambda J_m\left(\phi^*\right) \\ &=\left(Y-W\left(Sd+Qc\right)\right)'D^{-1}\left(Y-W\left(Sd+Qc\right)\right)+\lambda c'Qc \end{aligned} \tag{3}$$

lies within a space

$$\mathcal{H}\subseteq\left\{\phi^*:J\left(\phi^*\right)<\infty\right\}$$

in which $J\left(\phi^*\right)$ is a square (semi) norm, or a subspace therein. The evaluation functional $\left[\boldsymbol{v}\right]\phi^*$, which appears in the first term in 3, is assumed to be continuous in $\mathcal{H}$. A space in which the evaluation functional is continuous is called a reproducing kernel Hilbert space (RKHS) endowed with reproducing kernel (RK) $Q\left(\cdot,\cdot\right)$, a non-negative definite function satisfying

$$\left\langle Q\left(\boldsymbol{v},\cdot\right),\phi^*\left(\cdot\right)\right\rangle$$

$\forall\phi^*\in\mathcal{H}$, where $\left\langle\cdot,\cdot\right\rangle$ is an inner product in $\mathcal{H}$. The norm and RK determine each other uniquely.

Let $\mathcal{N}_J=\left\{\phi^*:\ J\left(\phi^*\right)=0\right\}$ denote the null space of J, and consider the tensor sum 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 3 has form

$$\phi^*\left(\boldsymbol{v}\right)=\sum_{\nu=1}^{d_0}d_\nu\eta_\nu\left(\boldsymbol{v}\right)+\sum_{i=1}^{n}c_iQ\left(\boldsymbol{v}_i,\boldsymbol{v}\right), \tag{4}$$

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

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

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

and

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

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

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

where Q_β is the RK in $\mathcal{H}_\beta$. The $\{\theta_\beta\}$ are additional smoothing parameters, which may or may not appear explicitly in notation to follow.

Letting $\tilde{Y} = D^{-1/2}Y$, $\tilde{S} = D^{-1/2}WS$, and $\tilde{Q} = D^{-1/2}WQ$, the penalized log likelihood 3 may be written

$$-2\ell_\lambda(c,d) + \lambda J(\phi^*) = \left[\tilde{Y} - \tilde{S}d - \tilde{Q}c\right]' \left[\tilde{Y} - \tilde{S}d - \tilde{Q}c\right] + \lambda c'Qc. \tag{5}$$

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

$$\begin{aligned} \tilde{S}'\tilde{S}d + \tilde{S}'\tilde{Q}c &= \tilde{S}'\tilde{Y} \\ \tilde{Q}'\tilde{S}d + \tilde{Q}'\tilde{Q}c + \lambda Qc &= \tilde{Q}'\tilde{Y}, \end{aligned} \tag{6}$$

Some algebra yields that this is equivalent to solving the system

$$\begin{bmatrix} \tilde{S}'\tilde{S} & \tilde{S}'\tilde{Q} \\ \tilde{Q}'\tilde{S} & \tilde{Q}'\tilde{Q} + \lambda Q \end{bmatrix} \begin{bmatrix} d \\ c \end{bmatrix} = \begin{bmatrix} \tilde{S}'\tilde{Y} \\ \tilde{Q}'\tilde{Y} \end{bmatrix} \tag{7}$$

Fixing smoothing parameters λ and θ_β (hidden in Q and $\tilde{Q}$ if present), assuming that $\tilde{Q}$ is full column rank, 7 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

$$\begin{bmatrix} \tilde{S}'\tilde{S} & \tilde{S}'\tilde{Q} \\ \tilde{Q}'\tilde{S} & \tilde{Q}'\tilde{Q} + \lambda Q \end{bmatrix} = \begin{bmatrix} C_1' & 0 \\ C_2' & C_3' \end{bmatrix} \begin{bmatrix} C_1 & C_2 \\ 0 & C_3 \end{bmatrix} \tag{8}$$

where $\tilde{S}'\tilde{S} = C_1'C_1$, $C_2 = C_1^{-T}\tilde{S}'\tilde{Q}$, and $C_3'C_3 = \lambda Q + \tilde{Q}'\left(I - \tilde{S}\left(\tilde{S}'\tilde{S}\right)^{-1}\tilde{S}'\right)\tilde{Q}$. 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

$$\tilde{C}_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & \delta I \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} C_1 & C_2 \\ 0 & \tilde{C}_3 \end{bmatrix}; \tag{9}$$

then

$$\tilde{C}^{-1} = \begin{bmatrix} C_1^{-1} & -C_1^{-1}C_2\tilde{C}_3^{-1} \\ 0 & \tilde{C}_3^{-1} \end{bmatrix}. \tag{10}$$

Premultiplying 8 by $\tilde{C}^{-T}$, straightforward algebra gives

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{C}_3^{-T}C_3^TC_3\tilde{C}_3^{-1} \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} C_1^{-T}\tilde{S}'\tilde{Y} \\ \tilde{C}_3^{-T}\tilde{Q}'\left(I - \tilde{S}\left(\tilde{S}'\tilde{S}\right)^{-1}\tilde{S}'\right)\tilde{Y} \end{bmatrix} \tag{11}$$

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

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

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

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} * \\ * \\ 0 \end{bmatrix}, \tag{12}$$

which can be solved, but with c_2 arbitrary. One may perform the Cholesky decomposition of 7 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

$$\widehat{\tilde{Y}} = \tilde{S}d + \tilde{Q}c = \begin{bmatrix} \tilde{S} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1}\tilde{C}^{-T} \begin{bmatrix} \tilde{S}' \\ \tilde{Q}' \end{bmatrix} \tilde{Y} = \tilde{A}\left(\lambda, \boldsymbol{\theta}\right)\tilde{Y}. \tag{13}$$

where

$$\begin{aligned} \tilde{A}\left(\lambda, \boldsymbol{\theta}\right) &= \begin{bmatrix} \tilde{S} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1}\tilde{C}^{-T} \begin{bmatrix} \tilde{S}' \\ \tilde{Q}' \end{bmatrix} \\ &= B + \left(I - B\right)\tilde{Q}\left[\tilde{Q}'\left(I - B\right)\tilde{Q} + \lambda Q\right]^{-1}\tilde{Q}'\left(I - B\right), \end{aligned} \tag{14}$$

for

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

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 5 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

$$\tilde{Y} = D^{-1/2} W \Phi^* + \tilde{\epsilon}, \tag{15}$$

where

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

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

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{N} \sum_{j=1}^{m_i} \left(\hat{\tilde{y}}_{ij} - E\left[\tilde{y}_{ij}\right]\right)^2 \\ &= ||\tilde{Y} - \tilde{\mu}||^2 \end{aligned} \tag{16}$$

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

$$L(\lambda) = \mu' \left(I - \tilde{A}\right)^2 \mu - 2\mu' \left(I - \tilde{A}\right)^2 \tilde{A}\tilde{\epsilon} + \tilde{\epsilon}' \tilde{A}^2 \tilde{\epsilon} \tag{17}$$

Define the unbiased risk estimate

$$U(\lambda) = \tilde{Y}' \left(I - \tilde{A}\right)^2 \tilde{Y} + 2\text{tr}\tilde{A} \tag{18}$$

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

$$\begin{aligned} U(\lambda) &= \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y}\right)' \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y}\right) + 2\text{tr}\tilde{A} \\ &= \left(\tilde{A}\tilde{Y} - \mu\right)' \left(\tilde{A}\tilde{Y} - \mu\right) + \tilde{\epsilon}'\tilde{\epsilon} + 2\tilde{\epsilon}' \left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right) \end{aligned} \tag{19}$$

This gives

$$U(\lambda) - L(\lambda) - \tilde{\epsilon}'\tilde{\epsilon} = 2\tilde{\epsilon}' \left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right), \tag{20}$$

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:

$$V_0(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=2}^{m_i} \frac{1}{m_i} \left(y_{ij} - \hat{y}_{ij}^{[-ij]}\right)^2, \tag{21}$$

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

$$y_k - \hat{y}_k^{[-k]} = (y_k - \hat{y}_k) / (1 - \tilde{a}_{kk}), \tag{22}$$

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, 21 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:

$$\bar{V}(\lambda) = \frac{1}{\sum_i m_i - N} \sum_{k=1}^{\sum_i m_i - N} \tilde{y}_k - \hat{\tilde{y}}_k^{[-k]} = \omega_k \left(\tilde{y}_k - \hat{\tilde{y}}_k\right)^2 / (1 - \tilde{a}_{kk})^2. \tag{23}$$

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 23 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

$$V_{loso}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \widehat{\tilde{\mu}}_i^{[-i]}\right)'\left(Y_i - \widehat{\tilde{\mu}}_i^{[-i]}\right) \tag{24}$$

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.]

$$\begin{aligned} MSPE &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - \widehat{\tilde{\mu}}_i||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - D_i^{-1/2}W_i\Phi^* + D_i^{-1/2}W_i\Phi^* - D_i^{-1/2}W_i\hat{\Phi}^*||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\tilde{\mu}_i - \widehat{\tilde{\mu}}_i^{[-i]}||^2\right]\right\} \end{aligned} \tag{25}$$

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

$$E\left[V_{loso}(\lambda)\right] = \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\widehat{\tilde{\mu}}_i - \tilde{\mu}_i||^2\right]\right\}. \tag{26}$$

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}:

$$V^*_{loso}(\lambda) = \frac{1}{N}||\left(I - \tilde{A}\right)\tilde{Y}||^2 + \frac{2}{N}\sum_{i=1}^{N}\hat{\tilde{e}}_i'\tilde{A}_{ii}\hat{\tilde{e}}_i, \tag{27}$$

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,*

$$V_{loso}(\lambda) - L(\lambda) - \frac{1}{N}\epsilon'\epsilon - o_p(L(\lambda)). \tag{28}$$

as $n \to \infty$

2.4 Selection of multiple smoothing parameters

The expression in 14 permits the straightforward evaluation of the GCV score

$$V(\lambda, \boldsymbol{\theta}) = \frac{(1/n_y) \left\| \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \tilde{Y} \right\|^2}{\left[(1/n_y) \operatorname{tr} \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \right]^2} \tag{29}$$

and the GML score

$$M(\lambda, \boldsymbol{\theta}) = \frac{(1/n_y) \tilde{Y}' \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \tilde{Y}}{\left[\det{}^{+} \left(I - \tilde{A}(\lambda, \boldsymbol{\theta}) \right) \right]^{1/n_y}}. \tag{30}$$

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

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 9. 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 $\breve{\theta}_{\beta}^{-1} \propto J_{\beta}\left(\breve{\phi}_{\beta}^{*}\right)$, minimize $V(\lambda)$ 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 somewhat 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. 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

3 Computation of the P-spline estimator

4 Smoothing parameter selection for tensor product P-splines