Web-based Supplementary Materials for "A Penalized spline approach to functional mixed effects model analysis" by Huaihou Chen and Yuanjia Wang

Semiparametric estimation of the within-subject variation

In this section, we present methods to estimate the population mean and the error variance function in model (1) nonparmetrically by penalized splines. Assume that the mean and the error variance function can be approximated by

$$\mu(t) = B_\mu(t)\beta_\mu,\ \log[\sigma^2(t)] = B_\sigma(t)\eta,$$

where $B_\mu(t)$ and $B_\sigma(t)$ are row vectors of basis functions for the mean and the variance function with possible different order p_μ and p_σ, different number of knots K_μ and K_σ, and β_μ and η are the associated coefficients. The heteroscedastic variance of the residual errors can be expressed as

$$V_i = \text{diag}[\exp(B_\sigma(t_{ij})\eta)]_{j=1,\cdots,m_i}.$$

With the above notation, we can rewrite the model (1) as

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

where $Y_i = (y_{ij})_{j=1,\cdots,m_i}$, $X_i = (x_i, B_\mu^i)$, $B_\mu^i = (B_\mu^T(t_{i1}), \cdots, B_\mu^T(t_{im_i}))^T$, $\beta = (\beta_0^T, \beta_\mu^T)^T$, $x_i = (x_{i1}, \cdots, x_{im_i})^T$, and $Z_i = (z_{i1}, \cdots, z_{im_i})^T$. Denote $Y_i^* = Y_i - X_i\beta - Z_i b_i$, we define the penalized log-likelihood as

$$l_p = \sum_{i=1}^{n} \{ \log |V_i^{\frac{1}{2}} R_i V_i^{\frac{1}{2}}| + Y_i^{*T} (V_i^{\frac{1}{2}} R_i V_i^{\frac{1}{2}})^{-1} Y_i^{*} \} + \lambda_\mu \mu^T P_\mu \mu + \lambda_\sigma \eta^T P_\sigma \eta, \tag{A-1}$$

where λ_μ and λ_σ are smoothing parameters for the mean and the variance function and P_μ and P_σ are penalty matrices depending on the chosen basis. For example, for the p_μ-th order truncated polynomial basis with K_μ knots, $P_\mu = \text{diag}\{\mathbf{0}_{p_\mu+1}, \mathbf{1}_{K_\mu}\}$ which implies that (A-1) only penalizes the spline coefficients. Throughout this section, we use truncated polynomial basis.

For given variance components, we estimate the baseline function by minimizing l_p in (A-1) and the solution takes the form of a ridge estimator as

$$\widehat{\beta} = (\sum_{i=1}^{n} X_i^T \Sigma_i^{-1} X_i + \lambda_\mu \text{diag}\{0_{p_x}, P_\mu\})^{-1} \sum_{i=1}^{n} X_i^T \Sigma_i^{-1} Y_i,$$

where $\Sigma_i = Z_i D Z_i^T + V_i^{\frac{1}{2}} R_i V_i^{\frac{1}{2}}$. To estimate the covariance matrix of the parametric random effects D, we use the EM algorithm. To fit the variance function of the within-subject residual measurement error, since no explicit solution exists for minimizing l_p with respect to η, we use the Newton-Raphson algorithm. To be specific, we obtain $\widehat{\eta}$ iteratively by

$$\widehat{\eta}^{(k+1)} = \widehat{\eta}^{(k)} - \left(\frac{\partial^2 l_p}{\partial \eta \partial \eta^T} | \widehat{\eta}^{(k)} \right)^{-1} \left(\frac{\partial l_p}{\partial \eta} \mid \widehat{\eta}^{(k)} \right),$$

where k index an iteration of the algorithm, and the first and the second derivatives are easily obtained based on (A-1). The correlation parameters θ are obtained by minimizing l_p also through a Newton-Raphson algorithm when no explicit solution exists.

Choosing the smoothing parameters

The smoothing parameters play a crucial role in the estimation procedure. Too small a penalty will lead to wiggly curves, while too large a penalty will result in flat polynomial curves which may lose the characteristic of the functions. Wand (2003) showed that by specifying spline coefficients of truncated polynomial basis functions as random effects in a linear mixed effects model, the penalized spline estimate with the smoothing parameter taken as the ratio of two variance components is identical to the best linear unbiased predictor (BLUP) obtained from a mixed effects model. Krivobokova and Kauermann (2007) showed that using the restricted maximized likelihood (REML) to estimate smoothing parameter outperforms other methods such as (generalized) cross-validation or the Akaike information criterion especially when the error correlation structure is misspecified. Krivobokova et al. (2008) formulated a hierarchical mixed model to estimate local smoothing parameter to achieve adaptive penalized spline smoothing. Kauermann and Wegener (2009) proposed to view the smoothing parameter of a variance function as a parameter and estimate it via maximizing the marginal log-likelihood. Here we use a similar likelihood-based strategy to chose λ_μ and λ_σ.

Denote $X = (X_1^T, \cdots, X_n^T)^T = (X_{(1)}, X_{(2)})$ where $X_{(1)}$ is the first $p_x + p_\mu + 1$ columns of X and $X_{(2)}$ is the remaining K_μ columns, where p_x is the length of the vector x_{ij}. Denote $\beta = (\beta_1^T, \beta_2^T)^T$ as the associated parameter vector. Due to the link of penalized spline likelihood and mixed effect models, we can treat the spline coefficients β_2 as random effects following $N(0, \sigma^2_{\beta_2} I)$ (Wand 2003; Krivobokova and Kauermann 2007). Integrating out the random components $b_i, i = 1, \cdots, n$, and β_2 results in the marginal likelihood. The smoothing parameter can be obtained via maximizing the marginal restricted log-likelihood

$$l_m(\lambda_\mu) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y - X_{(1)}\beta_1)^T\Sigma^{-1}(Y - X_{(1)}\beta_1) - \frac{1}{2}\log|X_{(1)}^T\Sigma^{-1}X_{(1)}|,$$

where Σ is the marginal covariance of Y, i.e.,

$$\begin{aligned}\Sigma &= E\{Var(Y|b,\beta_2)\} + Var\{E(Y|b,\beta_2)\} \\ &= V + Z\text{diag}\{D,\cdots,D\}Z^T + \frac{1}{\lambda_\mu}X_{(2)}X_{(2)}^T,\end{aligned}$$

with $V = \text{diag}\{V_1^{\frac{1}{2}}R_1V_1^{\frac{1}{2}},\cdots,V_n^{\frac{1}{2}}R_nV_n^{\frac{1}{2}}\}$. Note that here the smoothing parameter λ_μ appears as a parameter in the covariance matrix Σ. Applying Newton-Raphson algorithm, we have

$$\lambda_\mu^{\star(k+1)} = \lambda_\mu^{\star(k)} - \left(\frac{\partial^2 l_m(\lambda_\mu)}{\partial\lambda_\mu^{\star 2}}|\lambda_\mu^{*(k)}\right)^{-1}\left(\frac{\partial l_m(\lambda_\mu)}{\partial\lambda_\mu^\star}|\lambda_\mu^{*(k)}\right),$$

where $\lambda_\mu^\star = 1/\lambda_\mu$. The first and the second derivatives are easy to obtain. Finally, we obtain $\widehat{\lambda}_\mu = 1/\widehat{\lambda}_\mu^\star$.

We use a similar strategy to choose the smoothing parameter λ_σ of the variance function. To be specific, regard the spline coefficients in η as random effects and integrate them out to obtain the marginal log-likelihood

$$\begin{aligned}l_m(\lambda_\sigma) &= \log\int\exp\Big\{-\frac{1}{2}\sum_{i=1}^n(\log|V_i^{\frac{1}{2}}R_iV_i^{\frac{1}{2}}| + Y_i^{*T}(V_i^{\frac{1}{2}}R_iV_i^{\frac{1}{2}})^{-1}Y_i^*) \\ &- \frac{1}{2}\lambda_\sigma\eta^TP_\sigma\eta + \frac{1}{2}\log|\lambda_\sigma P_\sigma|_+\Big\}d\eta.\end{aligned}$$

Since there is no explicit solution to such an integration, we apply Laplace approximation to obtain

$$\begin{aligned}l_m(\lambda_\sigma) &\approx -\frac{1}{2}\sum_{i=1}^n(\log|V_i^{\frac{1}{2}}R_iV_i^{\frac{1}{2}}| + Y_i^{*T}(V_i^{\frac{1}{2}}R_iV_i^{\frac{1}{2}})^{-1}Y_i^*) \\ &- \frac{1}{2}\lambda_\sigma\eta^TP_\sigma\eta - \frac{1}{2}\log|H| + \frac{1}{2}\log|\lambda_\sigma P_\sigma|_+,\end{aligned}$$

with $H = -\frac{\partial^2(-\frac{1}{2}l_p)}{\partial\eta\partial\eta^T} = \frac{1}{2}\frac{\partial^2 l_p}{\partial\eta\partial\eta^T}$. Laplace approximation of a likelihood function has been discussed in Wolfinger (1993) and Kauermann and Wegener (2009). Specifically, we can approximate the marginal log-likelihood function

$$\log\int\exp(l(\eta))d\eta \approx l(\widehat{\eta}) - \frac{1}{2}\log|-l^{''}(\widehat{\eta})| + const.$$

The above approximation has an error of order $O(1/n)$. One important condition to achieve this approximation rate is that the number of spline bases functions must be small compared to the sample size n, that is, $K \ll n$ (Severini 2000; Kauermann et al. 2009). This condition is satisfied by penalized spline smoothing since the number of knots is much smaller than the sample size. Denote the right hand side of the above display as $\tilde{l}_m(\lambda_\sigma)$ and set its first derivative with respect to λ_σ to zero, i.e., $\dfrac{\partial \tilde{l}_m(\lambda_\sigma)}{\partial \lambda_\sigma} = -\dfrac{1}{2}\widehat{\eta}^T P_\sigma \widehat{\eta} - \dfrac{1}{2} tr\{H^{-1}P_\sigma\} + \dfrac{K_\sigma}{2\lambda_\sigma} = 0$, yields

$$\widehat{\lambda}_\sigma = \frac{1}{K_\sigma}(\widehat{\eta}^T P_\sigma \widehat{\eta} + tr\{H^{-1}P_\sigma\}).$$

The above formula is used iteratively in conjunction with the estimation of η.

Proofs of the Theorems 1 and 2

In this section, we prove the theorems stated in section 4. We first state the following assumptions for the theorems to hold.

Define $G_{K,n} = \frac{1}{n} N^T \Sigma^{-1} N$ and $H_{K,n} = G_{K,n} + \dfrac{\lambda}{n} D_q$. Applying the Demmler and Reinsch (1975) decomposition, we have

$$(N^T \Sigma^{-1} N)^{-1/2} D_q (N^T \Sigma^{-1} N)^{-1/2} = U^T \mathrm{diag}(S) U, \tag{A-2}$$

where U is an orthogonal matrix.

Lemma 1. *Under the assumption A2 and for the eigenvalues obtained in (A-2),*

$$s_1 = \cdots = s_q = 0, \quad s_j = n^{-1}(j-q)^{2q}\widehat{c}_1 \text{ for } j = q+1, \cdots, K+p+1, \tag{A-3}$$

where $\widehat{c}_1 = c_1(1+o(1))$ with c_1 a constant depending only on q and the design density and o(1) converges to 0 as $n \to \infty$ uniformly for $j_{1n} \le j \le j_{2n}$ for any sequences $j_{1n} \to \infty$ and $j_{2n} = o(n^{\frac{2}{2q+1}})$.

Since that the minimum and maximum eigenvalues of the matrix $(N^T\Sigma^{-1}N)^{-1/2}(N^TN)^{1/2}$ are of the same order, the Theorem 2.2 (2.5d) in Speckman (1985) is applicable.

To prove the main results, we first show the following preliminary results.

Result R1 (Lemma A1 in Zhu et al. 2008)

$$\begin{aligned} \| G_{K,n}^{-1} \|_\infty = \max_{1\le i\le K+p+1} \sum_{j=1}^{K+p+1} |\{G_{K,n}^{-1}\}_{i,j}| &= O(\delta^{-1}), && \text{(A-4)} \\ \sum_{i=1}^{K+p+1}\sum_{j=1}^{K+p+1} |\{G_{K,n} - G\}_{i,j}| &= o(\delta^2). && \text{(A-5)} \end{aligned}$$

Result (A-5) follows from the assumption A2 (A-15) and $\sum_{j=-p}^{K} N_j(t) = 1$.

Result R2

$$\begin{aligned} \| \frac{1}{n} N^T\Sigma^{-1}(\mu - s_\mu) \|_\infty &= o(\delta^{p+2}) && \text{(A-6)} \\ |E(\widehat{\mu}_{reg}(t)) - s_\mu(t)| &= o(\delta^{p+1}). && \text{(A-7)} \end{aligned}$$

Result (A-7) follows from Lemma A3 in Zhu et al. (2008) and $\| G_{K,n}^{-1} \| = O(\delta^{-1})$.

Result R3 (Lemma 6.1 in Cardot 2000)

$$\| D_q \|_\infty = O(\delta^{1-2q}). \tag{A-8}$$

Lemma 2. *Under the assumption A2 (A-15), we have*

$$\max_{1\le i,j\le K+p+1} |\{H_{K,n}^{-1}\}_{i,j}| = O(\delta^{-1}) \tag{A-9}$$

$$\| H_{K,n}^{-1} - H^{-1} \|_\infty = o(\delta^{-1}) \tag{A-10}$$

$$\max_{1\le i,j\le K+p+1} |\{H^{-1}\}_{i,j}| = O(\delta^{-1}). \tag{A-11}$$

Proof. From

$$\begin{aligned} H_{K,n}^{-1} &= G_{K,n}^{-\frac{1}{2}}(I + \frac{\lambda}{n} G_{K,n}^{-\frac{1}{2}} D_q G_{K,n}^{-\frac{1}{2}})^{-1} G_{K,n}^{-\frac{1}{2}} = G_{K,n}^{-\frac{1}{2}} U (I + \lambda \mathrm{diag}(S))^{-1} U^T G_{K,n}^{-\frac{1}{2}} \\ &= G_*(I + \lambda\mathrm{diag}(S))^{-1} G_*^T, \end{aligned}$$

where $G_* = G_{K,n}^{-\frac{1}{2}} U = (g_{ij}^*)_{1\le i,j\le K+p+1}$ and $G_* G_*^T = G_{K,n}^{-1}$, we have

$$
\begin{aligned}
|\{H_{K,n}^{-1}\}_{i,j}| &= |\sum_{l=1}^{K+p+1} \frac{g_{il}^* g_{jl}^*}{1+\lambda s_l}| \le \sqrt{\sum_{l=1}^{K+p+1} \frac{g_{il}^{*2}}{1+\lambda s_l} \sum_{l=1}^{K+p+1} \frac{g_{jl}^{*2}}{1+\lambda s_l}} \\
&\le \max_{1\le i\le K+p+1} \sum_{l=1}^{K+p+1} g_{il}^{*2} \le \| G_{K,n}^{-1} \|_\infty = O(\delta^{-1}). \qquad \text{(A-12)}
\end{aligned}
$$

The first inequality in (A-12) follows from Cauchy-Schwarz inequality, and second inequality follows from $s_l \ge 0$ for $l = 1, \cdots, K+p+1$. Therefore, $\max_{1\le i,j\le K+p+1} |\{H_{K,n}^{-1}\}_{i,j}| = O(\delta^{-1})$. Applying similar arguments as in Lemma A2 of Claeskens et al. (2009), leads to

$$
H^{-1} - H_{K,n}^{-1} = H_{K,n}^{-1}(G_{K,n} - G)\{I - H_{K,n}^{-1}(G_{K,n} - G)\}^{-1} H_{K,n}^{-1}. \qquad \text{(A-13)}
$$

Combing (A-5) with (A-13), yields (A-11). Result (A-11) follows from (A-10) and (A-11).

Note for $K_q = o(1)$,

$$
\| (I + G_{K,n}^{-1} \frac{\lambda}{n} D_q)^{-1} \|_\infty = \| \sum_{i=0}^{\infty} (-G_{K,n}^{-1} \frac{\lambda}{n} D_q)^i \|_\infty \le \sum_{i=0}^{\infty} \| G_{K,n}^{-1} \frac{\lambda}{n} D_q \|_\infty^i = \frac{1}{1+o(1)},
$$

since $\| G_{K,n}^{-1} \frac{\lambda}{n} D_q \|_\infty \le \| G_{K,n}^{-1} \|_\infty \| \frac{\lambda}{n} D_q \|_\infty = O(\delta^{-1}\delta^{1-2q}\frac{\lambda}{n}) = O(K_q) = o(1)$. Following that $\| H_{K,n}^{-1} \|_\infty = \| G_{K,n}^{-1}(I + G_{K,n}^{-1}\frac{\lambda}{n} D_q)^{-1} \|_\infty \le \| G_{K,n}^{-1} \|_\infty \| (I + G_{K,n}^{-1}\frac{\lambda}{n} D_q)^{-1} \|_\infty = O(\delta^{-1})$. Thus we can obtain $\| H_{K,n}^{-1} - H^{-1} \|_\infty = o(\delta^{-1})$ with the assumption A2 (A-14). $\square$

Let $s_\mu(\cdot) = N(\cdot)\beta$ be the best L_∞ approximation to the function μ.

Proof of Theorem 1.

First, we can rewrite

$$
\hat{\mu}(t) = \hat{\mu}_{\text{reg}}(t) - \frac{\lambda}{n} N(t) H_{K,n}^{-1} D_q G_{K,n}^{-1} \frac{1}{n} N^T \Sigma^{-1} Y,
$$

with $\hat{\mu}_{\text{reg}}(t) = \frac{1}{n} N(t) G_{K,n}^{-1} N^T \Sigma^{-1} Y$. Then we have

$$
E\hat{\mu}(t) - \mu(t) = \{s_\mu(t) - \mu(t)\} + \{E\hat{\mu}_{reg}(t) - s_\mu(t)\} - \frac{\lambda}{n} N(t) H_{K,n}^{-1} D_q G_{K,n}^{-1} \frac{1}{n} N^T \Sigma^{-1}(\mu - s_\mu + s_\mu).
$$

Barrow and Smith (1978) showed that $s_\mu(t)-\mu(t)=b_a(x,p+1)+o(\delta^{p+1})$. Here the order of the second term is found in R2 (A-7).

Applying the definition gives $s_\mu^{(q)}(t)=\{N(t)\beta\}^{(q)}=N_q(t)\Delta_q\beta$, with $N_q(t)=\{N_{-p+q,p+1-q}(t),\cdots,N_{K,p+1-q}(t)\}$. Noting $\beta=G_{K,n}^{-1}(\frac{1}{n}N^T\Sigma^{-1}N)\beta=\frac{1}{n}G_{K,n}^{-1}N^T\Sigma^{-1}s_\mu$ and $D_q=\Delta_q^TR\Delta_q$, we can obtain

$$\begin{aligned}
&\frac{\lambda}{n}N(t)H_{K,n}^{-1}D_qG_{K,n}^{-1}N^T\Sigma^{-1}s_\mu/n=\frac{\lambda}{n}N(t)H_{K,n}^{-1}D_q\beta\\
=\;&\frac{\lambda}{n}N(t)H_{K,n}^{-1}\Delta_q^T\int_a^b N_q^T(t)N_q(t)\Delta_q\beta dt=\frac{\lambda}{n}N(t)H_{K,n}^{-1}\Delta_q^T\int_a^b N_q^T(t)s_\mu^{(q)}(t)dt.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&-\frac{\lambda}{n}N(t)H_{K,n}^{-1}\Delta_q^T\int_a^b N_q(t)^Ts_\mu^{(q)}(t)dt\\
=\;&-\frac{\lambda}{n}N(t)H^{-1}\Delta_q^T\int_a^b N_q(t)^Ts_\mu^{(q)}(t)dt-\frac{\lambda}{n}N(t)(H_{K,n}^{-1}-H^{-1})\Delta_q^T\int_a^b N_q(t)^Ts_\mu^{(q)}(t)dt\\
=\;&b_\lambda(t,\Sigma)-\frac{\lambda}{n}N(t)(H_{K,n}^{-1}-H^{-1})\Delta_q^T\int_a^b N_q(t)^Ts_\mu^{(q)}(t)dt.
\end{aligned}$$

Now, we only need to prove that both $-\frac{\lambda}{n}N(t)(H_{K,n}^{-1}-H^{-1})\Delta_q^T\int_a^b N_q(t)^Ts_\mu^{(q)}(t)dt$ and $-\frac{\lambda}{n}N(t)H_{K,n}^{-1}D_qG_{K,n}^{-1}\frac{1}{n}N^T\Sigma^{-1}(\mu-s_\mu)$ are asymptotically ignorable. Note $0\le N_{j,q}(\cdot)\le 1$, it is easy to show that $\max\{\int_a^b N_q(t)dt\}=O(\delta)$. By the characteristic of the function space, $\sup_{t\in[a,b]}|s_\mu^{(q)}(t)|=O(1)$. For the second part of Theorem 4.1, when $\mu\in W^q[a,b]$, we can obtain similar result of $\max\{\int_a^b N_q(t)^Ts_\mu^{(q)}(t)dt\}=O(\delta)$. By definition, $\|\Delta_q\|_\infty=O(\delta^{-q})$ (see also Lemma 6.1 of Cardot 2000). Combing the above results, we have

$$\begin{aligned}
-\frac{\lambda}{n}N(t)(H_{K,n}^{-1}-H^{-1})\Delta_q^T\int_a^b N_q(t)^Ts_\mu^{(q)}(t)dt &= o(\lambda n^{-1}\delta^{-q}),\\
-\frac{\lambda}{n}N(t)H_{K,n}^{-1}D_qG_{K,n}^{-1}\frac{1}{n}N^T\Sigma^{-1}(\mu-s_\mu) &= o(\lambda n^{-1}\delta^{p-2q}).
\end{aligned}$$

Therefore, $E\hat\mu(t)-\mu(t)=b_a(t,p+1)+b_\lambda(t,\Sigma)+o(\delta^{p+1})+o(\lambda n^{-1}\delta^{-q})=O(\delta^{p+1})+O(\lambda n^{-1}\delta^{-q})$.

Next consider the variance, that is,

$$
\begin{aligned}
& Var(\hat{\mu}(t)) = \frac{1}{n} N(t) H_{K,n}^{-1} G_{K,n} H_{K,n}^{-1} N^T(t) \\
= \;& \frac{N(t)}{n} \{ H^{-1} G H^{-1} + H_{K,n}^{-1}(G_{K,n} - G) H_{K,n}^{-1} + H^{-1} G (H_{K,n}^{-1} - H^{-1}) \\
+ \;& (H_{K,n}^{-1} - H^{-1}) G H_{K,n}^{-1} \} N^T(t).
\end{aligned}
$$

Analogous to the bias, we have $\frac{1}{n}N(t)H_{K,n}^{-1}(G_{K,n} - G)H_{K,n}^{-1}N^T(t)$, $\frac{1}{n}N(t)(H_{K,n}^{-1} - H^{-1})GH_{K,n}^{-1}N^T(t)$ and $\frac{1}{n}N(t)H^{-1}G(H_{K,n}^{-1} - H^{-1})N^T(t)$ are of the same order $o(n^{-1}\delta^{-1})$. Finally, note that when $K_q = O(1)$, $o(\lambda n^{-1}\delta^{-q}) = o((\lambda/n)^{1/2})$ and $o(n^{-1}\delta^{-1}) = o(n^{-1}(\lambda/n)^{-1/2q})$. This proves the theorem 1. □

Proof of Theorem 2.

First note from Theorem 1, we have

$$
\frac{E\hat{\mu}(t) - \mu(t) - b_a(t) - b_\lambda(t,\Sigma)}{\sqrt{Var(\hat{\mu}(t))}} = \frac{o(\delta^{p+1}) + o(\lambda n^{-1}\delta^{-q})}{(n\delta)^{-\frac{1}{2}}} = o(\sqrt{n}\delta^{p+3/2}) + o(\lambda n^{-\frac{1}{2}}\delta^{\frac{1}{2}-q}) = o(1).
$$

Therefore, it is sufficient to show that

$$
\frac{\hat{\mu}(t) - E\hat{\mu}(t)}{\sqrt{Var(\hat{\mu}(t))}} \quad \xrightarrow{d} \quad N(0,1).
$$

We can represent

$$
\hat{\mu}(t) - E\hat{\mu}(t) = N(t)(N^T\Sigma^{-1}N + \lambda D_q)^{-1} \sum_{i=1}^n S_i^T V^{-1} \epsilon_i = \sum_{i=1}^n C_{ni}^T \epsilon_i,
$$

where $C_{ni} = N(t)(N^T\Sigma^{-1}N + \lambda D_q)^{-1} S_i^T V^{-1}$ with $S_i = (N^T(t_{i1}), \cdots, N^T(t_{im}))^T$. To check the Lindeberg condition, it suffices to show that

$$
\lim_{n\to\infty} \frac{\max_{1\le i\le n} \| C_{n,i} \|^2}{\sum_{i=1}^n \| C_{n,i} \|^2} = 0.
$$

Rewrite

$$
\begin{aligned}
\| C_{ni} \|^2 &= N^*(t)^T S_i^T V^{-2} S_i N^*(t), \\
\sum_{i=1}^n \| C_{ni} \|^2 &= N^*(t)^T \sum_{i=1}^n S_i^T V^{-2} S_i N^*(t) = N^*(t)^T N^T \Sigma^{-2} N N^*(t),
\end{aligned}
$$

where $N^*(t) = (N^T\Sigma^{-1}N + \lambda D_q)^{-1}N(t)^T$. Since

$$\lambda_{min}(N^T\Sigma^{-2}N) \geq cn\delta, \quad \lambda_{max}(S_iS_i^T) \leq \sum_{j=1}^m N(t_{ij})N(t_{ij})^T \leq \sum_{j=1}^m \sum_{l=-p}^K N_l(t_{ij}) = m$$
$$\lambda_{max}(S_i^TV^{-2}S_i) \leq \lambda_{max}(V^{-2})\lambda_{max}(S_iS_i^T), \quad \max_{1\leq i\leq n}\lambda_{max}(S_i^TV^{-2}S_i) = O(1),$$

where $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote respectively the smallest and largest eigenvalues for A,

$$\frac{\max_{1\leq i\leq n} \parallel C_{n,i} \parallel^2}{\sum_{i=1}^n \parallel C_{n,i} \parallel^2} \leq \frac{\max_{1\leq i\leq n}\lambda_{max}(S_i^TV^{-2}S_i)}{cn\delta} = O(\frac{1}{n\delta}).$$

This proves the theorem. □

Asymptotic properties for P-spline estimator with truncated polynomial basis

Assumption 1. Let $\delta_j = \tau_{j+1} - \tau_j$ and $\delta = \max_{0\leq j\leq K}\delta_j$. There exists a constant $M > 0$, such that $\delta/(\min_{0\leq j\leq K}\delta_j) \leq M$ and $\delta \sim K^{-1}$.

Assumption 2. For any $j, l = 1, \cdots, m$,

$$\sup_{x,y\in[a,b]}|Q_{n,jl}(x,y) - Q_{jl}(x,y)| = o(K^{-2}), \quad \sup_{x\in[a,b]}|Q_{n,j}(x) - Q_j(x)| = o(K^{-2}), \quad \text{(A-14)}$$

$$\sup_{x,y\in[a,b]}|Q_{n,jl}(x,y) - Q_{jl}(x,y)| = o(K^{-4}), \quad \sup_{x\in[a,b]}|Q_{n,j}(x) - Q_j(x)| = o(K^{-3}), \quad \text{(A-15)}$$

where $Q_{n,jl}(x,y) = \frac{1}{n}\sum_{i=1}^n I(t_{i,j} \leq x, t_{i,l} \leq y), Q_{n,j}(x) = \frac{1}{n}\sum_{i=1}^n I(t_{ij} \leq x)$, and $Q_{jl}(x,y)$ and $Q_j(x)$ are certain distribution functions with positive continuous density functions $\rho_{jl}(x,y)$ and $\rho_j(x)$ on $[a,b]\times[a,b]$ and $[a,b]$, respectively.

Assumption 3. The number of knots $K = o(n)$.

We now extend the asymptotic properties in section 4.2 to the truncated polynomial basis. With a slight abuse of notation, let $B(t)$ be the pth order truncated polynomial basis with K knots, let $B = (B(t_{11})^T, \cdots, B(t_{nm})^T)^T$, let $P = \text{diag}(0_{p+1}, 1_K)$

and let λ_* denote the penalty for the truncated polynomial spline estimator. The fitted estimator is

$$\widehat{\mu}_* = B(B^T\Sigma^{-1}B + \lambda_* P)^{-1}B^T Y.$$

Since there exists a square and invertible transition matrix L, such that $N = BL$ (de Boor 2001, Claeskens et al. 2009), we can rewrite the estimator as

$$\widehat{\mu}_* = N(N^T\Sigma^{-1}N + \lambda_* L^T P L)^{-1}N^T Y.$$

Therefore, replacing the penalty term λD_q in a B-spline estimator by $\lambda_* L^T P L$ yields an equivalent estimator, $\widehat{\mu}_*$. Denote $\widehat{\mu}_*(t) = B(t)(B^T\Sigma^{-1}B + \lambda_* P)^{-1}B^T\Sigma^{-1}Y$ and $K_{p+1} = \lambda K^{2p+2}/n$. Applying the asymptotic results obtained in the previous section to the $\widehat{\mu}_*(t)$, we have the following theorems.

Theorem A. 1. *Under the assumptions A1-A3 and $\mu(\cdot) \in C^{p+1}[a,b]$, the following results hold:*

1. *If $K_{p+1} = o(1)$, then*

$$E(\widehat{\mu}_*(t)) - \mu(t) = b_a(t, p+1) + b^*_\lambda(t,\Sigma) + o(\delta^{p+1}) + o(\lambda n^{-1}\delta^{-p}),$$
$$Var(\widehat{\mu}_*(t)) = \frac{1}{n}N(t)(G + \frac{\lambda}{n}D_q)^{-1}G(G + \frac{\lambda}{n}D_q)^{-1}N^T(t) + o((n\delta)^{-1}),$$

and for $K \sim n^{1/(2p+3)}$ and $\lambda = O(n^{2/(2p+3)})$, the optimal rate for MSE $n^{-(2p+2)/(2p+3)}$ is attained by the penalized spline estimator.

2. *If $K_{p+1} = O(1)$, then*

$$E(\widehat{\mu}_*(t)) - \mu(t) = b_a(t, p+1) + b^*_\lambda(t,\Sigma) + o(\delta^{p+1}) + o((\lambda/n)^{(p+1)/(2p+1)}),$$
$$Var(\widehat{\mu}_*(t)) = \frac{1}{n}N(t)(G + \frac{\lambda}{n}D_q)^{-1}G(G + \frac{\lambda}{n}D_q)^{-1}N^T(t) + o(n^{-1}(\lambda/n)^{-1/(2p+1)}),$$

and for $\lambda \sim n^{2/(2p+3)}$ and $K \sim n^{1/(2p+3)}$, the optimal rate for MSE $n^{-(2p+2)/(2p+3)}$ is attained by the penalized spline estimator.

Remark 1. *In contrast to the B-spline basis, the optimal rate of convergence for $\mu(t)$ estimated by truncated polynomial basis is the same for the small and large number of knots case. This result also holds for univariate data (Claeskens et al. 2009).*

Remark 2. *Lin et al. (2004) showed that the asymptotic rate of the MSE of the qth order smoothing spline is $O((\lambda/n)^2) + O(n^{-1+1/2q}\lambda^{-1/(2q)})$. Thus when $\lambda = O(n^{-2q/(4q+1)})$ the optimal rate is achieved at $O(n^{-4q/(4q+1)})$, which corresponds to the second scenario of the Theorem A.1 with $p = 2q - 1$.*

Theorem A. 2. *Assume $K^{2p+3} \sim n$, $\lambda = O(K^2)$ and $h > 0$, $C > 0$, such that $\sup_{i,j} E|\epsilon_{ij}|^{2+h} \leq C$. Then*

$$\frac{\widehat{\mu}_*(t) - \mu(t) - b_a(t, p+1) - b^*_\lambda(t, \Sigma)}{\sqrt{Var(\widehat{\mu}_*(t))}} \longrightarrow N(0, 1)$$

in distribution, as $n \longrightarrow \infty$.

Proofs of Theorems A.1 and A.2.

Note that $\{N(t)\beta\}^{(p)} = \sum_{j=0}^{K} N_{j,1}(t)\beta_j^{(p)} = \sum_{j=1}^{K} I(\tau_j \leq t)(\beta_j^{(p)} - \beta_{j-1}^{(p)}) + \beta_0^{(p)}$, where $\beta^{(p)}$ is the pth difference of β defined in Claeskens et al. (2009). Since the derivative of an indicator function is a Dirac delta function which integrates to one, we have

$$\int_a^b [\{N(t)\beta\}^{(p+1)}]^2 dt = \sum_{j=1}^{K} (\beta_j^{(p)} - \beta_{j-1}^{(p)})^2.$$

The transition matrix L can be obtained from the equation

$$\lambda_* \beta^T L^T P L \beta = \lambda \beta^T D_q \beta = \lambda \sum_{j=1}^{K} (\beta_j^{(p)} - \beta_{j-1}^{(p)})^2.$$

Rewrite $\sum_{j=1}^{K} (\beta_j^{(p)} - \beta_{j-1}^{(p)})^2 = {\beta^{(p)}}^T Q^T Q \beta^{(p)}$, with Q as a $(K+1) \times (K+p+1)$ transition matrix. For equidistant knots, $\beta^{(p)} = \delta^p \nabla_p \beta$ where ∇_p is a difference operator matrix defined in Claeskens et al. (2009). It follows that

$$\lambda_* \beta^T L^T P L \beta = \lambda {\beta^{(p)}}^T Q^T Q \beta^{(p)} = \lambda \delta^{-2p} \beta^T \nabla_p^T Q^T Q \nabla_p \beta.$$

Table A1: Average computing time for the first scenario in Simulation II with 100 replications

$(n,m)^*$	$(100,10)$	$(100,15)$	$(100,20)$	$(100,30)$	$(150,10)$	$(200,10)$	$(300,10)$
Time†	1.42	3.16	7.78	15.72	3.29	7.80	16.31

*: n is the number of subjects and m is the number of observations per subject.
†: The unit of computing time is minute.

Therefore, $\widehat{\mu}_*$ corresponds to a B-spline estimator with equidistant knots which satisfies $\lambda_* L^T PL = \lambda D_q = \lambda\delta^{-2p}\nabla_p^T Q^T Q\nabla_p$. The asymptotic bias, variance and normality can be obtained, following the arguments in the proof of Theorems 1 and 2 via replacing λD_q by $\lambda_*\delta^{-2p}\nabla_p^T Q^T Q\nabla_p$. $\square$

Numerical performance and implementation

The computing time to fit the model by the proposed algorithm depends on the number of subjects and number of observations per subject. We used scenario 1 in the Simulation II to assess computational burden. We present the computing time on a Dell desktop with a 2.67 GHz CPU and 4GB RAM with different configurations of sample size in Table A1 of this appendix. An example of the R source code of the core functions to fit the model can be found at http://www.columbia.edu/ yw2016/SemiCovcode.R.

References

Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics* 12, 503-38.

Demmler, A. and Reinsch, C. (1975). Oscillation matrices with spline smoothing. *Numerische Mathematik* 24, 375-82.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association* **102**, 1328–1337.

Krivobokova, T., Crainiceanu, C.M. and Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, 17, 1–20.

Lin, X. , Wang, N. , Welsh, A. and Carroll, R. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika* 91, 177-193.

Severini, T. A. (2000). Likelihood Methods in Statistics. Oxford: Oxford University Press.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in non-parametric regression models. *Annals of Statistics* 13, 970-83.

Wand, M.P. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* 80, 791-795.

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Annals of Statistics* 26, 1760-82.