Smoothing Spline Mixed-Effects Density Models for Clustered Data

Chi-Yang Chiu, Anna Liu and Yuedong Wang

University of Tennessee Health Science Center,

University of Massachusetts at Amherst, and

University of California at Santa Barbara

Supplementary Material

S1 An example on construction of an NMECDM based on an SS ANOVA decomposition

We assume $\mathcal{Y} = \mathbb{R}$ and model g as a function of y using the thin-plate spline model space $\mathcal{H}_y = W_2^3(\mathbb{R}) \ominus \{1\}$. Suppose that we want to model g as a function of x using an RKHS $\mathcal{H}_x = \mathcal{H}_{0x} \oplus \mathcal{H}_{1x}$ where $\mathcal{H}_{0x}$ is a finite dimensional space collecting functions which are not penalized. Let P_y and P_x be the projection operators onto $\mathcal{H}_{0y}$ and $\mathcal{H}_{0x}$ respectively. Let

$P_\omega g = \int_\Omega g(y,x,\omega)dP$. We have the following SS ANOVA decomposition

$$\begin{aligned} g &= [P_y + (I-P_y)][P_x + (I-P_x)][P_\omega + (I-P_\omega)]g \\ &\triangleq g_{ppf}(y,x) + g_{psf}(y,x) + g_{spf}(y,x) + g_{ssf}(y,x) \\ &\quad + g_{ppr}(y,x,\omega) + g_{psr}(y,x,\omega) + g_{spr}(y,x,\omega) + g_{ssr}(y,x,\omega) \end{aligned} \tag{S1.1}$$

where g_{ppf} is the parametric fixed effect, g_{psf}, g_{spf} and g_{ssf} are nonparametric fixed effects, g_{ppr} is the parametric random effect, and g_{psr}, g_{spr} and g_{ssr} are nonparametric random effects. Comparing to the NMECDM (2.5), we have $\mathcal{H}_\eta = (W_2^3(\mathbb{R}) \ominus \{1\}) \otimes \mathcal{H}_x$, $\eta(y,x) = g_{ppf}(y,x) + g_{psf}(y,x) + g_{spf}(y,x) + g_{ssf}(y,x)$ and $b_i(y) = g_{ppr}(y,x,\omega_i) + g_{psr}(y,x,\omega_i) + g_{spr}(y,x,\omega_i) + g_{ssr}(y,x,\omega_i)$.

Rather than the conditional density, regression models focus on the conditional expectation $\mu(x,\omega) = \mathrm{E}(Y|X=x,\omega)$. Suppose that we want to model μ as a function of x using the RKHS $\mathcal{H}_x = \mathcal{H}_{0x} \oplus \mathcal{H}_{1x}$. Applying a similar SS ANOVA decomposition to the random function $\mu(x,\omega)$ we have the following SS ANOVA mixed effects model (Wang, 1998)

$$Y_{ij} = \mu_{pf}(X_{ij}) + \mu_{sf}(X_{ij}) + \mu_{pr}(X_{ij},\omega_i) + \mu_{sr}(X_{ij},\omega_i) + \epsilon_{ij}, \tag{S1.2}$$

where μ_{pf} and μ_{sf} are parametric and nonparametric fixed effects, μ_{pr} and μ_{sr} are parametric and nonparametric random effects, $\epsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$, and random effects and ϵ_{ij} are mutually independent. Then, up to a con-

stant independent of y, the logistic density of Y_{ij} conditional on $X_{ij} = x$ and cluster ω has the form $\{(-y^2/2 + \mu_{pf}y) + \mu_{sf}(x)y + \mu_{pr}(x,\omega)y + \mu_{sr}(x,\omega)y\}/\sigma^2$. Comparing to the SS ANOVA decomposition (S1.1), it is obvious that the SS ANOVA mixed effects model (S1.2) is a special case with $g_{ppf}(y,x) \cong (-y^2/2 + \mu_{pf}(x)y)/\sigma^2$, $g_{psf}(y,x) \cong \mu_{sf}(x)y/\sigma^2$, $g_{ppr}(y,x,\omega) \cong \mu_{pr}(x,\omega)y/\sigma^2$, $g_{psr}(y,x,\omega) \cong \mu_{sr}(x,\omega)y/\sigma^2$, and $g_{spf}(y,x,\omega) = g_{ssf}(y,x,\omega) = g_{spr}(y,x,\omega) = g_{ssr}(y,x,\omega) = 0$.

S2 Derivatives of log-likelihood

Note that

$$\log f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i) \cong \sum_{j=1}^{n_i} \{\eta(Y_{ij},X_{ij}) + b_i(Y_{ij},X_{ij}) - \log\int_{\mathcal{Y}} \exp\{\eta(y,X_{ij}) + b_i(y,X_{ij})\}dy\Big\},$$

where

$$\eta(y,x) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(y,x) + \sum_{l=1}^{L} c_l R_1^*(\boldsymbol{U}_l,(y,x)).$$

Hence

$$\begin{aligned}
&\frac{\partial \log f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i)}{\partial c_l} \\
=\ & \sum_{j=1}^{n_i}\left\{R_1^*(\boldsymbol{U}_l,(Y_{ij},X_{ij})) - \frac{\int_{\mathcal{Y}} R_1^*(\boldsymbol{U}_l,(y,X_{ij}))\exp\{\eta(y,X_{ij}) + b_i(y,X_{ij})\}dy}{\int_{\mathcal{Y}}\exp\{\eta(y,X_{ij}) + b_i(y,X_{ij})\}dy}\right\} \\
=\ & \sum_{j=1}^{n_i}\Big\{R_1^*(\boldsymbol{U}_l,(Y_{ij},X_{ij})) - \mathrm{E}_{Y|\boldsymbol{B}_i} R_1^*(\boldsymbol{U}_l,(Y,X_{ij}))\Big\},
\end{aligned}$$

and

$$
\begin{aligned}
& \frac{\partial^2 \log f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i)}{\partial c_l \partial c_k} \\
= & -\sum_{j=1}^{n_i} \left\{ \frac{\int_{\mathcal{Y}} R_1^*(\boldsymbol{U}_l,(y,X_{ij}))R_1^*(\boldsymbol{U}_k,(y,X_{ij}))\exp\{\eta(y,X_{ij})+b_i(y,X_{ij})\}dy}{\int_{\mathcal{Y}}\exp\{\eta(y,X_{ij})+b_i(y,X_{ij})\}dy} \right. \\
& - \left(\frac{\int_{\mathcal{Y}} R_1^*(\boldsymbol{U}_l,(y,X_{ij}))\exp\{\eta(y,X_{ij})+b_i(y,X_{ij})\}dy}{\int_{\mathcal{Y}}\exp\{\eta(y,X_{ij})+b_i(y,X_{ij})\}dy} \right) \\
& \left. \left(\frac{\int_{\mathcal{Y}} R_1^*(\boldsymbol{U}_k,(y,X_{ij}))\exp\{\eta(y,X_{ij})+b_i(y,X_{ij})\}dy}{\int_{\mathcal{Y}}\exp\{\eta(y,X_{ij})+b_i(y,X_{ij})\}dy} \right) \right\} \\
= & \sum_{j=1}^{n_i} \mathrm{Cov}_{Y|\boldsymbol{B}_i}(R_1^*(\boldsymbol{U}_k,(Y,X_{ij})), R_1^*(\boldsymbol{U}_l,(Y,X_{ij}))).
\end{aligned}
$$

Other first and second derivatives can be calculated similarly.

S3 Markov Chain Monte Carlo

We need to generate MCMC samples for the computation of conditional expectations with respect to $\boldsymbol{B}_i|\boldsymbol{Y}_i$. We first divide the domain $\mathcal{Y}$ into a finite number of disjoint subsets $\mathcal{Y} = \cup_{k=1}^{K}\mathcal{Y}_k$ and select points $\tilde{y}_k \in \mathcal{Y}_k$ for $k = 1,\ldots,K$. We then apply the Metropolis-Hastings (MH) procedure to generate MCMC samples for random vectors $\tilde{\boldsymbol{B}}_{ij} = (b_i(\tilde{y}_1,X_{ij}),\ldots,b_i(\tilde{y}_K,X_{ij}))$ conditional on $\boldsymbol{Y}_i$ for $i = 1,\ldots,m$ and $j = 1,\ldots,n_i$. Finally we approximate the conditional expectations with respect to $\boldsymbol{B}_i|\boldsymbol{Y}_i$ using these MCMC samples. Details of the MH procedure can be found in Gelman et al. (2003).We use a multivariate normal centered at the current MCMC

sample as the proposal distribution with a scaled covariance matrix such that the acceptance rate is near 23% as suggested by Gelman et al. (2003, Ch11). Convergence and mixing are assessed by visualization tools such as trace plot and autocorrelation.

S4 Quadratic Approximation and Cross-Validation

The log marginal likelihood for cluster i is $l_i = \log \mathrm{E}_{\boldsymbol{B}_i} f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i)$ where $f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i)$ is given in (3.2). For $f, g \in \mathcal{H}_\eta$ and $\alpha \in \mathbb{R}$, consider

$$L_{f,g}(\alpha) = \log \mathrm{E}_{\boldsymbol{B}_i} \frac{\exp\{\sum_{j=1}^{n_i}[f(Y_{ij}, X_{ij}) + \alpha g(Y_{ij}, X_{ij}) + b_i(Y_{ij}, X_{ij})]\}}{\int_{\mathcal{Y}} \exp\{f(y, X_{ij}) + \alpha g(y, X_{ij}) + b_i(y, X_{ij})\} dy} \tag{S4.1}$$

as a function of α. It can be shown that the derivatives of $L_{f,g}(\alpha)$ with respect to α evaluated at zero are given as follows:

$$\begin{aligned} L'_{f,g}(0) &= \sum_{j=1}^{n_i} g(Y_{ij}, X_{ij}) - \sum_{j=1}^{n_i} \mathrm{E}^f_{\boldsymbol{B}_i|\boldsymbol{Y}_i} E^f_{Y|\boldsymbol{B}_i} g(Y, X_{ij}), \\ L''_{f,g}(0) &= \sum_{j=1}^{n_i} \mathrm{E}^f_{\boldsymbol{B}_i|\boldsymbol{Y}_i} \mathrm{V}^f_{Y|\boldsymbol{B}_i} g(Y, X_{ij}) + \mathrm{V}^f_{\boldsymbol{B}_i|\boldsymbol{Y}_i} \left\{ \sum_{j=1}^{n_i} E^f_{Y|\boldsymbol{B}_i} g(Y, X_{ij}) \right\}, \end{aligned}$$

where $\mathrm{E}^f_{\cdot|\cdot}$ and $\mathrm{V}^f_{\cdot|\cdot}$ respectively represent conditional expectation and variance under model (2.5) with $\eta(y, x) = f(y, x)$.

For any fixed $\tilde{\eta}$, setting $f = \tilde{\eta}$ and $g = \eta - \tilde{\eta}$, we have the quadratic

approximation of the log-likelihood l_i at $\tilde{\eta}$

$$\begin{aligned}
l_i &= L_{f,g}(1) \\
&\approx L_{f,g}(0) + L'_{f,g}(0) + \frac{1}{2}L''_{f,g}(0) \\
&= \log \mathrm{E}_{\boldsymbol{B}_i} \frac{\exp\{\sum_{j=1}^{n_i}[\tilde{\eta}(Y_{ij}, X_{ij}) + b_i(Y_{ij}, X_{ij})]\}}{\int_{\mathcal{Y}} \exp\{\tilde{\eta}(y, X_{ij}) + b_i(y, X_{ij})\}dy} + \\
&\quad \sum_{j=1}^{n_i} \{\eta(Y_{ij}, X_{ij}) - \tilde{\eta}(Y_{ij}, X_{ij})\} \\
&\quad - \sum_{j=1}^{n_i} \mathrm{E}^{\tilde{\eta}}_{\boldsymbol{B}_i|\boldsymbol{Y}_i} E^{\tilde{\eta}}_{Y|\boldsymbol{B}_i} \{\eta(Y, X_{ij}) - \tilde{\eta}(Y, X_{ij})\} \\
&\quad + \frac{1}{2} \sum_{j=1}^{n_i} \mathrm{E}^{\tilde{\eta}}_{\boldsymbol{B}_i|\boldsymbol{Y}_i} \mathrm{V}^{\tilde{\eta}}_{Y|\boldsymbol{B}_i} \{\eta(Y, X_{ij}) - \tilde{\eta}(Y, X_{ij})\} \\
&\quad + \frac{1}{2} \mathrm{V}^{\tilde{\eta}}_{\boldsymbol{B}_i|\boldsymbol{Y}_i} \left\{ \sum_{j=1}^{n_i} E^{\tilde{\eta}}_{Y|\boldsymbol{B}_i} (\eta(Y, X_{ij}) - \tilde{\eta}(Y, X_{ij})) \right\}.
\end{aligned}$$

Dropping terms do not involve η and the last term for computational stability, and summing log-likelihoods over all clusters, we have the quadratic approximation to log marginal likelihood at $\tilde{\eta}$

$$\begin{aligned}
\tilde{l}(\boldsymbol{\zeta}, \eta) &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \Big\{ \eta(Y_{ij}, X_{ij}) - \mathrm{E}^{\tilde{\eta}}_{\boldsymbol{B}_i|\boldsymbol{Y}_i} E^{\tilde{\eta}}_{Y|\boldsymbol{B}_i} \eta(Y, X_{ij}) \\
&\quad + \frac{1}{2} \mathrm{E}^{\tilde{\eta}}_{\boldsymbol{B}_i|\boldsymbol{Y}_i} \mathrm{V}^{\tilde{\eta}}_{Y|\boldsymbol{B}_i} (\eta(Y, X_{ij}) - \tilde{\eta}(Y, X_{ij})) \Big\}.
\end{aligned}$$

Consider the approximated penalized likelihood

$$-\frac{1}{N}\tilde{l}(\boldsymbol{\zeta}, \eta) + \frac{1}{2}\sum_{j=1}^{q} \lambda_j ||P_j \eta||^2, \tag{S4.2}$$

and denote $\eta^{[i,j]}_{\boldsymbol{\lambda}}$ as the solution to (S4.2) with the jth observation from cluster i begin removed. Let $\boldsymbol{\psi} = (\phi_1, \ldots, \phi_p, \xi_1, \ldots, \xi_L)^T$, $\breve{R} = (\boldsymbol{\psi}(\boldsymbol{Z}_{11}), \ldots,$

$\boldsymbol{\psi}(\boldsymbol{Z}_{mn_m}))^T$ where $\boldsymbol{Z}_{ij} = (Y_{ij}, X_{ij})$, $P_{\mathbf{1}}^{\perp} = I_N - \mathbf{1}_N\mathbf{1}_N^T/N$ where I_N is an $N \times N$ identity matrix and $\mathbf{1}_N$ is an N-vector of all ones, and $\Pi = \mathrm{E}_{\boldsymbol{B}|\boldsymbol{Y}} I(\tilde{\boldsymbol{c}}, \tilde{\boldsymbol{d}}, \boldsymbol{B})$ where I is a proxy of the Hessian matrix defined in Section 3.2, and $\tilde{\boldsymbol{c}}$ and $\tilde{\boldsymbol{d}}$ are estimates at the current iteration. Following similar arguments as in Gu (2013), it can be shown that

$$\begin{aligned}\eta_{\boldsymbol{\lambda}}^{[i,j]}(Y_{ij}, X_{ij}) &= \eta_{\boldsymbol{\lambda}}(Y_{ij}, X_{ij}) - \frac{1}{N-1}\{\boldsymbol{\psi}(\boldsymbol{Z}_{ij})\}^T\Pi^{-1}\{\boldsymbol{\psi}(\boldsymbol{Z}_{ij}) \\ &\quad -N^{-1}\breve{R}^T\mathbf{1}\}.\end{aligned}$$

Noting that $\breve{R}^T\mathbf{1} = \sum_{i=1}^m \sum_{j=1}^{n_i} \boldsymbol{\psi}(\boldsymbol{Z}_{ij})$, we have

$$\frac{1}{N}\sum_{i=1}^m\sum_{j=1}^{n_i}\eta_{\boldsymbol{\lambda}}^{[i,j]}(Y_{ij}, X_{ij}) = \frac{1}{N}\sum_{i=1}^m\sum_{j=1}^{n_i}\eta_{\boldsymbol{\lambda}}(Y_{ij}, X_{ij}) - \frac{tr(P_1^{\perp}\breve{R}\Pi^{-1}\breve{R}^T P_1^{\perp})}{N(N-1)}.$$

The cross-validation score (4.4) is derived by plugging in the above into (4.3).

S5 Figures and A Table

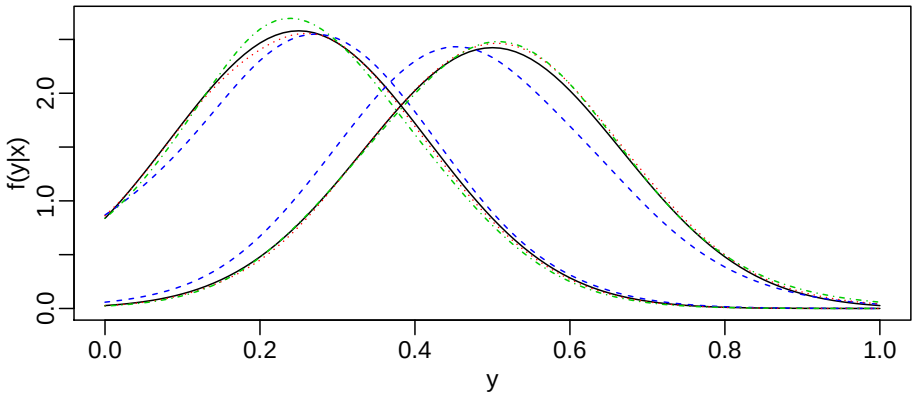

Figure A.1: True and estimates of the population conditional density function when x is discrete and $m = 200$. The solid black curves are the true population conditional density functions of the two groups, the dashed blue curves are the estimates with the largest AKL loss, the dotted red curves are the estimates with the smallest AKL loss, and the dash-dot green curves are the estimates with the median AKL loss.

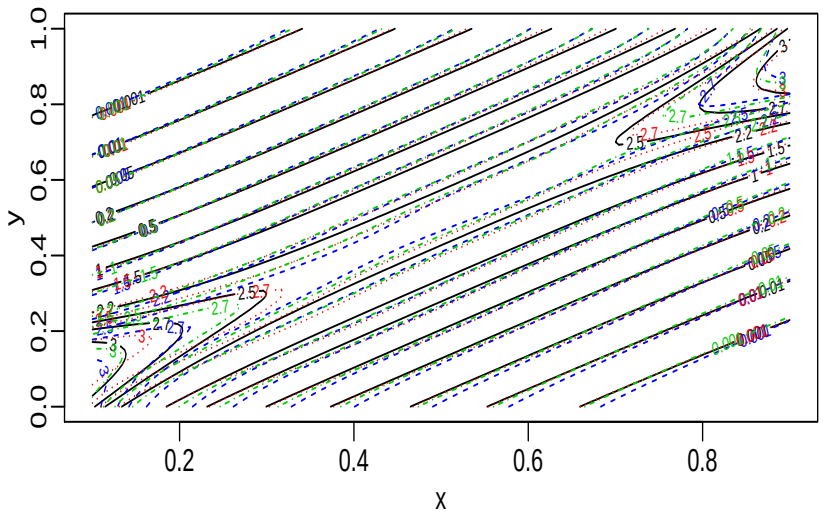

Figure A.2: Contour plots of the true population conditional density function (solid black lines) and three estimates corresponding to the largest AKL loss (dashed blue lines), median AKL loss (dash-dot green lines), and the smallest AKL loss (dotted red lines) when x is continuous and $m = 200$.

Parameter	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8
σ_1^2	0.75	0.31	0.82	0.71	0.83	0.84	0.74	0.71
σ_2^2	142.27	192.23	144.71	142.38	144.11	144.97	143.37	143.21

Table A.1: Estimates of the variance components from the eight subsets of the Hb measurements.