Statistica Sinica Preprint No: SS-2017-0211	
Title	Smoothing Spline Mixed-Effects Density Models for Clustered Data
Manuscript ID	SS-2017-0211
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0211
Complete List of Authors	Chi-Yang Chiu Anna Liu and Yuedong Wang
Corresponding Author	Yuedong Wang
E-mail	yuedong@pstat.ucsb.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

Statistica Sinica

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

Abstract: We propose smoothing spline mixed-effects density models for nonparametric estimations of density and conditional density functions with clustered data. The random effects in a density model introduce within-cluster correlation, enabling us to borrow strength across clusters by shrinking cluster-specific density functions to the population average, where the amount of shrinkage is decided by the data. Estimations are carried out using the penalized likelihood and are computed using a Markov chain Monte Carlo stochastic approximation algorithm. We derive an approximate cross-validation estimate of the aggregated Kullback–Leibler loss for the selection of the smoothing parameters. Our simulation study indicates that the proposed estimation method performs well. We apply our methods to investigate the evolution of hemoglobin density functions over time in response to guideline changes on anemia management for dialysis patients.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

Key words and phrases: Markov chain Monte Carlo, penalized likelihood, random effects, smoothing spline ANOVA decomposition, stochastic approximation.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

1. Introduction

Density estimation plays a fundamental role in many areas, including statistics and machine learning. Estimated density functions are useful for model building and diagnostics, inferences, predictions, and clustering. Many nonparametric methods have been developed to estimate the density of independent data (Silverman, 1984; Gu, 2013).

A central problem in statistics is the development of methods to assess the relationship between a dependent variable and one or more independent variables. A regression analysis usually focuses on univariate characteristics, such as the conditional expectation or quantile of the dependent variable, given the independent variables. Typically, the family of conditional distributions is assumed to be known (e.g., Gaussian). In some applications, it is difficult, if not impossible, to specify a specific family of distributions, and the goal is to investigate covariate effects on the whole conditional density function. A conditional density provides the most informative summary of the relationship between independent and dependent variables. For example, it allows us to examine the overall shape and to summarize characteristics such as quantiles and modes. Estimated conditional density functions may be used for further analysis, including inferences, predictions, clustering, and functional data analyses (Petersen and Müller, 2016). Many

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

nonparametric methods have been developed to estimate the conditional density of independent data. See for example, Hall et al. (2004), Fan and Yim (2004), Dunson et al. (2007), Efromovich (2007), Gu (2013), and the references therein.

Clustered data arise in areas such as agriculture, pharmacokinetics, epidemiology, medicine, and social science. Observations from the same cluster are usually correlated, and there is a large body of literature on modeling conditional means using random effects (Wu and Zhang, 2006; Wang, 2011). We are interested in estimating the density or conditional density for a population, as well as for each cluster, in order to investigate the covariate effects on the density functions and the variations between clusters. Despite its importance, there has been little research on density and conditional density estimations for clustered data. One exception is the study by Rodriguez et al. (2009), with interesting applications to DNA damage and repair. Rodriguez et al. (2009) used a finite mixture of Gaussian distributions to approximate the density, and a hierarchical model of mixture weights to assess the heterogeneity across clusters and covariate effects. Note that this is a parametric model because the number of mixtures is finite. However, one needs to specify hyperparameters, which may become difficult when the number of mixtures is not small. To the best of our

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

knowledge, no similar nonparametric method has been developed. Therefore, we propose a general and flexible family of nonparametric mixed-effects models for density and conditional density functions with clustered data.

The remainder of the article is organized as follows. In Section 2, we introduce nonparametric mixed-effects density and conditional density models. Sections 3 and 4 present the procedures used to estimate and select, respectively, the smoothing parameters. Section 5 describes our simulation studies. We apply the proposed methods to investigate the changes in hemoglobin (Hb) distributions over time in Section 6. Section 7 concludes the paper.

2. Nonparametric Mixed-Effects Density Models (NMEDMs)

2.1. Density models for clustered data

Let Y_{ij}, for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, be the jth observation from cluster i, where the domain of Y_{ij} is an arbitrary set $\mathcal{Y}$. Assume that the observed clusters form a random sample from a population of clusters, denoted as Ω, with sampling distribution P. Denote $f(y|\omega)$ as the cluster-specific density function, where ω is a random sample from Ω. Consequently, $f(y|\omega)$ is a random function on the product domain $\mathcal{Y} \times \Omega$. Denote the observed clusters as $\omega_1, \ldots, \omega_m$, which are realizations of the random variable ω. For a given ω_i, we assume that $Y_{ij} \overset{iid}{\sim} f(y|\omega_i)$ and that

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

observations from different clusters are mutually independent. To enforce the conditions of $f > 0$ and $\int_{\mathcal{Y}} f = 1$ for a density function, throughout this article, we use the logistic transformation, $f = \exp(g)/\int_{\mathcal{Y}} \exp(g)$, where g is referred to as the logistic transformation of f (Gu, 2013).

A NMEDM for clustered data assumes

$$g(y, \omega_i) = \eta(y) + b_i(y), \tag{2.1}$$

where $\eta(y)$ is the fixed effect and $b_i(y)$ is the random effect. We assume that $\eta(y) \in \mathcal{H}_\eta$, where $\mathcal{H}_\eta$ is a reproducing kernel Hilbert space (RKHS) and $b_i(y)$ are independent Gaussian processes with mean zero and covariance function $\sigma(s, t)$. Different methods may be used to construct $\mathcal{H}_\eta$ and $\sigma(s, t)$. Here we assume that $\mathcal{H}_\eta$ and $\sigma(s, t)$ are constructed using a smoothing spline ANOVA (SS ANOVA) decomposition (Wang, 2011). Specifically, we assume that $\mathcal{H}_\eta = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \cdots \oplus \mathcal{H}^q$, where $\mathcal{H}^0$ is a finite-dimensional space of all functions that are not penalized, $\mathcal{H}^1, \ldots, \mathcal{H}^q$ are orthogonal RKHSs, and $b_i(y)$ collects some of the random components in the SS ANOVA decomposition.

Details of SS ANOVA decompositions for general tensor products of RKHSs can be found in Chapters 4 and 9 of Wang (2011). We now use an example to illustrate the construction of an NMEDM based on an SS ANOVA decomposition and compare it with the classical one-way random

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

effect model. Suppose $\mathcal{Y} = \mathbb{R}$ and that we want to model g as a function of y using the thin-plate spline model space $\mathcal{H}_y = W_2^3(\mathbb{R}) \ominus \{1\}$, where

$$W_2^3(\mathbb{R}) = \left\{ h : \int_{-\infty}^{\infty} \left(h^{(3)}(y)\right)^2 dy < \infty \right\}. \tag{2.2}$$

The constant functions have been removed from $W_2^3(\mathbb{R})$ for identifiability (Gu, 2013). $\mathcal{H}_y$ is an RKHS, which can be decomposed into $\mathcal{H}_y = \mathcal{H}_{0y} \oplus \mathcal{H}_{1y}$, where $\mathcal{H}_{0y} = \text{span}\{\varphi_2(y),\ \varphi_3(y)\}$, φ_2 and φ_3 are the linear and quadratic basis functions, respectively, and $\mathcal{H}_{1y}$ is the orthogonal complement of $\mathcal{H}_{0y}$. Let P_y be the projection operator onto $\mathcal{H}_{0y}$, defined under a suitably defined inner product (see Gu (2013) for details). Let $P_\omega g = \int_\Omega g(y, \omega) dP$, which computes the population average with respect to the sampling distribution. We have the following SS ANOVA decomposition:

$$g(y,\omega) = [P_y + (I - P_y)][P_\omega + (I - P_\omega)]g \stackrel{\triangle}{=} g_{pf}(y) + g_{sf}(y) + g_{pr}(y,\omega) + g_{sr}(y,\omega), \tag{2.3}$$

where I is the identity map, $g_{pf} \in \mathcal{H}_{0y}$ is a quadratic polynomial corresponding to the parametric fixed main effect of the variable y, g_{sf} is the nonparametric fixed main effect of y, and g_{pr} and g_{sr} are the parametric and nonparametric random effects, respectively. The letters "p" and "s" in the subscripts represent the parametric components in space $\mathcal{H}_{0y}$ and the smooth components in space $\mathcal{H}_{1y} = W_2^3(\mathbb{R}) \ominus \{1, \varphi_2(y), \varphi_3(y)\}$, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

The letters "f" and "r" in the subscripts represent the fixed and random effects, respectively. Compared with the NMEDM (2.1), we have $\eta(y) = g_{pf}(y) + g_{sf}(y)$, $\mathcal{H}^0 = \mathcal{H}_{0y}$, and $\mathcal{H}^1 = \mathcal{H}_{1y}$. Furthermore, assuming that $g_{pr}(y, \omega_i) = u_{1i}\varphi_2(y) + u_{2i}\varphi_3(y)$, $u_{1i} \overset{iid}{\sim} \mathrm{N}(0, \sigma_1^2)$, $u_{2i} \overset{iid}{\sim} \mathrm{N}(0, \sigma_2^2)$, $g_{sr}(y, \omega_i)$ are independent Gaussian processes with mean zero and covariance function $\sigma_3^2 R^1(s,t)$, where $R^1(s,t)$ is the reproducing kernel (RK) of $\mathcal{H}^1$, and u_{1i}, u_{2i}, and $g_{sr}(y, \omega_i)$ are mutually independent, then $b_i(y) = g_{pr}(y, \omega_i) + g_{sr}(y, \omega_i)$ and $\sigma(s,t) = \sigma_1^2\varphi_2(s)\varphi_2(t) + \sigma_2^2\varphi_3(s)\varphi_3(t) + \sigma_3^2 R^1(s,t)$. For simplicity of notation, we assume that u_{1i} and u_{2i} are mutually independent. In practice, a bivariate Gaussian distribution may be assumed for the joint distribution of (u_{1i}, u_{2i}).

The classical one-way random-effect model for clustered data described in this section assumes that

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{2.4}$$

where $\alpha_i \overset{iid}{\sim} \mathrm{N}(0, \sigma_a^2)$, $\epsilon_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, and α_i and ϵ_{ij} are mutually independent. Then, up to a constant independent of y, the logistic density of Y_{ij} conditional on the random effects α_i, has the form $(-y^2/2+\mu y)/\sigma^2+\alpha_i y/\sigma^2$. Compared with the SS ANOVA decomposition (2.3), the one-way random-effect model is a special case, with $g_{pf}(y) \cong (-y^2/2 + \mu y)/\sigma^2$, $g_{pr}(y, \omega_i) \cong \alpha_i y/\sigma^2$, and $g_{sf}(y) = g_{sr}(y, \omega) = 0$, where $\cong$ denotes equality up to a

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

constant.

2.2. Conditional density models for clustered data

Let (X_{ij}, Y_{ij}), for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, be the jth observation from cluster i, where the domains of X_{ij} and Y_{ij} are arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. Again, assume that the observed clusters are a random sample from a population of clusters, denoted as Ω, with sampling distribution P. Denote $f(y|x, \omega)$ as the cluster-specific density function of Y conditional on $X = x$, where ω is a random sample from Ω. Denote the observed clusters as $\omega_1, \ldots, \omega_m$, which are realizations of the random variable ω. For a given ω_i, we assume that $Y_{ij}|X_{ij} = x \overset{iid}{\sim} f(y|x, \omega_i)$ and observations from different clusters are mutually independent. Let $g(y, x, \omega_i)$ be the logistic transformation of $f(y|x, \omega_i)$.

A nonparametric mixed-effects conditional density model (NMECDM) assumes that

$$g(y, x, \omega_i) = \eta(y, x) + b_i(y, x), \tag{2.5}$$

where $\eta(y, x)$ denotes the fixed effects and $b_i(y, x)$ is the random effect. We assume that $\eta(y, x) \in \mathcal{H}_\eta = \mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \cdots \oplus \mathcal{H}^q$, and that $b_i(y, x)$ are independent Gaussian processes with mean zero and covariance function $\sigma(s, t|x)$.

In the Supplementary Material S1, we give an example illustrating the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

construction of an NMECDM based on an SS ANOVA decomposition, and compare the resulting model with the SS ANOVA mixed-effects regression models. In the example, we considered the thin-plate spline space (2.2) because its null space consists of quadratic polynomials that correspond to the logistic transformation of Gaussian density functions. The SS ANOVA decomposition may be derived for tensor products of general RKHSs (Gu, 2013; Wang, 2011). Additional examples of SS ANOVA decompositions logistic transformations of density functions with clustered data can be found in Section 5 and in Chiu (2015). Similarly to the classical ANOVA, the SS ANOVA produces a hierarchical structure that facilitates model selection and interpretation. Note that some of the components (e.g., high-order interactions) in the SS ANOVA decomposition may be dropped to overcome the curse of dimensionality.

3. Estimation

We describe our estimation procedure for NMECDM (2.5) only, because the estimation for NMEDM (2.1) is similar, but simpler.

3.1. Penalized likelihood and its approximate solution

Denote the Gaussian stochastic process of the random effect for cluster i as $B_{ij} = \{b_i(y, X_{ij}), y \in \mathcal{Y}\}$ and $\boldsymbol{B}_i$ as the collection of B_{ij}, for $j = 1, \ldots, n_i$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$. The log likelihood

$$l(\boldsymbol{\zeta}, \eta) = \sum_{i=1}^{m} \log \mathrm{E}_{\boldsymbol{B}_i} f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i), \tag{3.1}$$

where

$$f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i) = \prod_{j=1}^{n_i} \frac{\exp\{\eta(Y_{ij}, X_{ij}) + b_i(Y_{ij}, X_{ij})\}}{\int_{\mathcal{Y}} \exp\{\eta(y, X_{ij}) + b_i(y, X_{ij})\} dy}, \tag{3.2}$$

is the conditional density of $\boldsymbol{Y}_i$, and $\boldsymbol{\zeta}$ collects all parameters related to the random effects $\boldsymbol{B}_i$.

Let $\boldsymbol{z} = (y, x)$ and $\boldsymbol{Z}_{ij} = (Y_{ij}, X_{ij})$. The model space for η is $\mathcal{H}^0 \oplus \mathcal{H}^1 \oplus \cdots \oplus \mathcal{H}^q$, where $\mathcal{H}^0 = \mathrm{span}\{\phi_1(\boldsymbol{z}), \ldots, \phi_p(\boldsymbol{z})\}$ contains those functions that are not penalized, and $\mathcal{H}^j$ for $j = 1, \ldots, q$ are RKHSs with RKs R^j. We estimate $\boldsymbol{\zeta}$ and η by minimizing the penalized likelihood

$$PL = -\frac{1}{N} l(\boldsymbol{\zeta}, \eta) + \frac{1}{2} \sum_{j=1}^{q} \lambda_j ||P_j \eta||^2, \tag{3.3}$$

where $N = \sum_{i=1}^{m} n_i$, P_j is the projector onto the space $\mathcal{H}^j$, and λ_j are smoothing parameters. Denote $\lambda_j = \lambda/\theta_j$. Define $\mathcal{H}_1^* = \oplus_{j=1}^{q} \mathcal{H}^j$ and a new squared norm on $\mathcal{H}_1^*$ as $\sum_{j=1}^{q} \theta_j^{-1} ||P^j f||^2$. Then, the RK of $\mathcal{H}_1^*$ under the new norm is $R_1^* = \sum_{j=1}^{q} \theta_j R^j$ (see Wang (2011) for details). The penalized likelihood (3.3) reduces to

$$PL = -\frac{1}{N} l(\boldsymbol{\zeta}, \eta) + \frac{\lambda}{2} ||P_1^* \eta||^2, \tag{3.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

where $P_1^* = \sum_{j=1}^{q} P_j$. We minimize the PL (3.4) in the following finite-dimensional data-adaptive space:

$$\mathcal{H}_\eta^* = \mathcal{H}^0 \oplus \text{span}\{R_1^*(\boldsymbol{U}_l, \cdot),\ l = 1, \ldots, L\}, \tag{3.5}$$

where $\{\boldsymbol{U}_l,\ l = 1, \ldots, L\}$ is a random subset of $\{\boldsymbol{Z}_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, n_i\}$. Gu and Wang (2003) suggested that an L close to $10N^{2/9}$ is sufficient, in the sense that the estimates in the whole model space $\mathcal{H}_\eta$ and the subspace $\mathcal{H}_\eta^*$ have the same convergence rate. Let $\xi_l(\boldsymbol{z}) = R_1^*(\boldsymbol{U}_l, \boldsymbol{z})$. The minimizer of the PL (3.4) in $\mathcal{H}_\eta^*$ has the form (Gu and Wang, 2003)

$$\hat{\eta}(\boldsymbol{z}) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(\boldsymbol{z}) + \sum_{l=1}^{L} c_l \xi_l(\boldsymbol{z}). \tag{3.6}$$

Let $\boldsymbol{c} = (c_1, \ldots, c_L)^T$ and $\boldsymbol{d} = (d_1, \ldots, d_p)^T$. Substituting (3.6) into (3.4), we have

$$PL(\boldsymbol{\zeta}, \boldsymbol{c}, \boldsymbol{d}) = -\frac{1}{N} l(\boldsymbol{\zeta}, \eta) + \frac{\lambda}{2} \boldsymbol{c}^T Q_\theta \boldsymbol{c}, \tag{3.7}$$

where Q_θ is an $L \times L$ matrix with the (i, j)th entry equal to $R_1^*(\boldsymbol{U}_i, \boldsymbol{U}_j)$.

3.2. Markov chain Monte Carlo (MCMC) stochastic approximation

The log likelihood function (3.1) involves expectations with respect to the random effects that do not have closed forms. We use the MCMC method to approximate the expectations with respect to the random effects. To make our computational procedure converge to the expected fixed points,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

we adopt the stochastic approximation algorithm (SAA), which is described in this section. See Gu and Kong (1998), Gu and Zhu (2001) and, Jiang et al. (2011) for details.

Consider solving the following equation

$$\mathrm{E}_{\boldsymbol{e}}\boldsymbol{H}(\boldsymbol{\beta},\boldsymbol{e})=\mathbf{0}, \tag{3.8}$$

where $\boldsymbol{e}$ is a random vector with a density function $f_{\boldsymbol{e}}$. Let $I(\boldsymbol{\beta},\boldsymbol{e}) = -\partial\boldsymbol{H}(\boldsymbol{\beta},\boldsymbol{e})/\partial\boldsymbol{\beta}$. At iteration k, an MCMC sample of size m_k, with equilibrium distribution $f_{\boldsymbol{e}}$, is drawn and denoted as $\boldsymbol{e}_k^{(1)},\ldots,\boldsymbol{e}_k^{(m_k)}$. Let $\bar{\boldsymbol{H}}_k = \sum_{j=1}^{m_k}\boldsymbol{H}(\boldsymbol{\beta}_{k-1},\boldsymbol{e}_k^{(j)})/m_k$ and $\bar{I}_k = \sum_{j=1}^{m_k} I(\boldsymbol{\beta}_{k-1},\boldsymbol{e}_k^{(j)})/m_k$. Then, the MCMC SAA updates the parameter vector $\boldsymbol{\beta}$ and a matrix Γ, as follows:

$$\Gamma_k=(1-\gamma_k)\Gamma_{k-1}+\gamma_k\bar{I}_k,\quad \boldsymbol{\beta}_k=\boldsymbol{\beta}_{k-1}+\gamma_k\Gamma_k^{-1}\bar{\boldsymbol{H}}_k, \tag{3.9}$$

where Γ_k acts as a proxy for the Hessian matrix, and γ_k is the step-size of the parameter updates. By increasing the MCMC sample size m_k, decreasing the step-size γ_k, or a combination of the two, the variation in $\boldsymbol{\beta}$ decreases as the number of iterations increases. It has been shown that, under some regularity conditions, $\boldsymbol{\beta}_k$ converges to the solution of (3.8) almost surely (Gu and Kong, 1998) when m_k and γ_k satisfy the following conditions: (a) $\gamma_k \le 1$, for all k; (b) $\sum_{k=1}^{\infty}\gamma_k=\infty$; (c) $\sum_{k=1}^{\infty}\gamma_k^{1+\varepsilon}/m_k<\infty$, for some $\varepsilon\in(0,1)$; and (d) $\sum_{k=1}^{\infty}|\gamma_k/m_k-\gamma_{k-1}/m_{k-1}|<\infty$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

3.3. Estimation of η

In this subsection, we apply the MCMC SAA to compute $\boldsymbol{c}$ and $\boldsymbol{d}$ with a fixed $\boldsymbol{\zeta}$. Our goal is to find the values of $\boldsymbol{c}$ and $\boldsymbol{d}$ that minimize the PL (3.7). It is not difficult to show that

$$\frac{\partial PL(\boldsymbol{\zeta},\boldsymbol{c},\boldsymbol{d})}{\partial(\boldsymbol{c}^T,\boldsymbol{d}^T)^T} = -\frac{1}{N}\sum_{i=1}^{m} \mathrm{E}_{\boldsymbol{B}_i|\boldsymbol{Y}_i}\left\{\frac{\partial \log f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i)}{\partial(\boldsymbol{c}^T,\boldsymbol{d}^T)^T}\right\} + \frac{\lambda}{2}\frac{\partial \boldsymbol{c}^T Q_\theta \boldsymbol{c}}{\partial(\boldsymbol{c}^T,\boldsymbol{d}^T)^T}.$$

Denote the above quantity as $\mathrm{E}_{\boldsymbol{e}}\boldsymbol{H}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{B})$, where $\boldsymbol{B}$ and $\boldsymbol{Y}$ collect all $\boldsymbol{B}_i$ and $\boldsymbol{Y}_i$ respectively, $\boldsymbol{e} = \boldsymbol{B}|\boldsymbol{Y}$, and

$$\boldsymbol{H}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{B}) = -\frac{1}{N}\sum_{i=1}^{m} \frac{\partial \log f_{\boldsymbol{Y}_i|\boldsymbol{B}_i}(\boldsymbol{Y}_i)}{\partial(\boldsymbol{c}^T,\boldsymbol{d}^T)^T} + \frac{\lambda}{2}\frac{\partial \boldsymbol{c}^T Q_\theta \boldsymbol{c}}{\partial(\boldsymbol{c}^T,\boldsymbol{d}^T)^T}.$$

Let $I(\boldsymbol{c},\boldsymbol{d},\boldsymbol{B}) = -\partial\boldsymbol{H}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{B})/\partial(\boldsymbol{c}^T,\boldsymbol{d}^T)^T$. It can be shown that (see the Supplementary Material S2)

$$\boldsymbol{H}(\boldsymbol{c},\boldsymbol{d},\boldsymbol{B}) = \begin{bmatrix} -N^{-1}\sum_{i=1}^{m}\Lambda_i\mathbf{1}_{n_i} + \lambda Q_\theta \boldsymbol{c} \\ -N^{-1}\sum_{i=1}^{m} S_i\mathbf{1}_{n_i} \end{bmatrix}, \tag{3.10}$$

and

$$I(\boldsymbol{c},\boldsymbol{d},\boldsymbol{B}) = \begin{bmatrix} N^{-1}\sum_{i=1}^{m} V_{i,\xi\xi} + \lambda Q_\theta & N^{-1}\sum_{i=1}^{m} V_{i,\xi\phi} \\ N^{-1}\sum_{i=1}^{m} V_{i,\phi\xi} & N^{-1}\sum_{i=1}^{m} V_{i,\phi\phi} \end{bmatrix}, \tag{3.11}$$

where Λ_i is an $L \times n_i$ matrix with the $(l,j)th$ entry equal to $\xi_l(Y_{ij},X_{ij}) - \mathrm{E}_{Y|\boldsymbol{B}_i}\xi_l(Y,X_{ij})$, S_i is a $p \times n_i$ matrix with the $(\nu,j)th$ entry equal to $\phi_\nu(Y_{ij},X_{ij}) - \mathrm{E}_{Y|\boldsymbol{B}_i}\phi_\nu(Y,X_{ij})$, $\mathbf{1}_{n_i}$ is an n_i-vector of ones, $V_{i,\xi\xi}$ is an $L \times L$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

matrix with the $(k,l)th$ entry equal to $\sum_{j=1}^{n_i} \mathrm{Cov}_{Y|\boldsymbol{B}_i}(\xi_k(Y, X_{ij}), \xi_l(Y, X_{ij}))$, $V_{i,\phi\phi}$ is a $p \times p$ matrix with the $(\nu,\kappa)th$ entry equal to $\sum_{j=1}^{n_i} \mathrm{Cov}_{Y|\boldsymbol{B}_i}(\phi_\nu(Y, X_{ij}), \phi_\kappa(Y, X_{ij}))$, $V_{i,\xi\phi}$ is an $L \times p$ matrix with the $(l,\nu)th$ entry being $\sum_{j=1}^{n_i} \mathrm{Cov}_{Y|\boldsymbol{B}_i}(\xi_l(Y, X_{ij}), \phi_\nu(Y, X_{ij}))$, and $V_{i,\phi\xi} = V_{i,\xi\phi}^T$.

At iteration k of the MCMC SAA, let $\boldsymbol{B}_k^{(1)}, ..., \boldsymbol{B}_k^{(m_{1k})}$ be an MCMC sample of size m_{1k} generated from $f_{\boldsymbol{B}|\boldsymbol{Y}}$. With fixed $\boldsymbol{\zeta}$ and the current estimates of $\boldsymbol{c}$ and $\boldsymbol{d}$ denoted as $\boldsymbol{c}_{k-1}$ and $\boldsymbol{d}_{k-1}$, respectively, for any $\boldsymbol{B}_k^{(\nu)}$, the conditional distribution of $Y|\boldsymbol{B}_k^{(\nu)}$ is known (more precisely, $Y|X_{ij}, \boldsymbol{B}_k^{(\nu)}$, where the condition on X_{ij} is omitted for simplicity of notation). Therefore, the conditional expectation $\mathrm{E}_{Y|\boldsymbol{B}_k^{(\nu)}} a(Y)$ and conditional covariance $\mathrm{Cov}_{Y|\boldsymbol{B}_k^{(\nu)}}(a(Y), b(Y))$ can be calculated for any functions a and b. We compute the conditional expectations and conditional covariances in $\boldsymbol{H}$ and I, and denote the resulting quantities as $\boldsymbol{H}(\boldsymbol{c}_{k-1}, \boldsymbol{d}_{k-1}, \boldsymbol{B}_k^{(\nu)})$ and $I(\boldsymbol{c}_{k-1}, \boldsymbol{d}_{k-1}, \boldsymbol{B}_k^{(\nu)})$, respectively. We then compute $\bar{\boldsymbol{H}}_k = \sum_{\nu=1}^{m_{1k}} \boldsymbol{H}(\boldsymbol{c}_{k-1}, \boldsymbol{d}_{k-1}, \boldsymbol{B}_k^{(\nu)})/m_{1k}$ and $\bar{I}_k = \sum_{\nu=1}^{m_{1k}} I(\boldsymbol{c}_{k-1}, \boldsymbol{d}_{k-1}, \boldsymbol{B}_k^{(\nu)})/m_{1k}$. Following (3.9), we update $\boldsymbol{c}$, $\boldsymbol{d}$, and Γ as follows:

$$\Gamma_k = (1-\gamma_k)\Gamma_{k-1} + \gamma_k \bar{I}_k, \quad \begin{bmatrix} \mathbf{c}_k \\ \mathbf{d}_k \end{bmatrix} = \begin{bmatrix} \mathbf{c}_{k-1} \\ \mathbf{d}_{k-1} \end{bmatrix} + \gamma_k \Gamma_k^{-1} \bar{\boldsymbol{H}}_k. \tag{3.12}$$

3.4. Estimation of $\boldsymbol{\zeta}$ and the complete algorithm

We now apply the MCMC SAA to compute $\boldsymbol{\zeta}$ with fixed η. Be-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

cause the penalty term in the the PL (3.7) does not depend on $\boldsymbol{\zeta}$, it is not difficult to show that $\partial PL(\boldsymbol{\zeta}, \boldsymbol{c}, \boldsymbol{d})/\partial\boldsymbol{\zeta} = \mathrm{E}_{\boldsymbol{B}|\boldsymbol{Y}}\boldsymbol{H}(\boldsymbol{\zeta}, \boldsymbol{B})$, where $\boldsymbol{H}(\boldsymbol{\zeta}, \boldsymbol{B}) = -N^{-1}\partial \log p_{\boldsymbol{B}}(\boldsymbol{B}; \boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$ and $p_{\boldsymbol{B}}(\boldsymbol{B}; \boldsymbol{\zeta})$ is the joint density function of $\boldsymbol{B}$. It is then straightforward to apply the MCMC SAA to update $\boldsymbol{\zeta}$.

We now have the following complete algorithm:

1. Provide initial values $\boldsymbol{c}^{(0)}$, $\boldsymbol{d}^{(0)}$, and $\boldsymbol{\zeta}^{(0)}$;

2. At iteration k,

   (a) with fixed $\boldsymbol{\zeta}^{(k-1)}$, draw an MCMC sample of size m_{1k} and update $\boldsymbol{c}$ and $\boldsymbol{d}$ using equation (3.12);

   (b) with fixed updated estimates $\boldsymbol{c}^{(k)}$ and $\boldsymbol{d}^{(k)}$, draw another MCMC sample of size m_{2k} and update $\boldsymbol{\zeta}$, as discussed above;

3. Repeat Step 2 until convergence.

Methods for deriving the initial values and a stopping criterion can be found in Chiu (2015). The MCMC procedure is discussed in the Supplementary Material S3. Note that we allow the MCMC sample sizes in (a) and (b) to be different. We considered three MCMC SAA schemes, as in Jiang et al. (2011): (i) $\gamma_k = 1$ and $m_{jk} = m_{j0} + k^2$; (ii) $\gamma_k = 1/k$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

$m_{jk} = m_{j0}$; and (iii) $\gamma_k = 1/\sqrt{k}$ and $m_{jk} = m_{j0} + k$, where $j = 1, 2$, and m_{10} and m_{20} are the starting MCMC sample sizes for steps (a) and (b), respectively. Simulations indicate that the scheme (i) is more stable and efficient. Therefore, we use this scheme in our simulations and application.

The estimate of the population density, $f(y|x) = \exp\{\eta(y,x)\}/\int_{\mathcal{Y}} \exp\{\eta(t,x)\}dt$, is $\hat{f}(y|x) = \exp\{\hat{\eta}(y,x)\}/\int_{\mathcal{Y}} \exp\{\hat{\eta}(t,x)\}dt$. For any fixed $x \in \mathcal{X}$, denote $B_i(x) = \{b_i(y,x),\ y \in \mathcal{Y}\}$. Conditional on $X = x$, let $b_i^{(l)}(y,x)$, for $l = 1, \cdots, M$, be an MCMC sample of size M generated from $f_{B_i(x)|\boldsymbol{Y}}$, with η and $\boldsymbol{\zeta}$ fixed at their estimates, where M is a sufficiently large number. We estimate the random effect $b_i(y,x)$ in (2.5) by $\hat{b}_i(y,x) = \sum_{l=1}^{M} b_i^{(l)}(y,x)/M$, and estimate the cluster-specific density function $f(y|x,\omega_i)$ by $\hat{f}(y|x,\omega_i) = \exp\{\hat{\eta}(y,x) + \hat{b}_i(y,x)\}/\int_{\mathcal{Y}} \exp\{\hat{\eta}(t,x) + \hat{b}_i(t,x)\}dt$.

4. Selection of Smoothing Parameters

The smoothing parameters λ_j, for $j = 1, \ldots, q$, are crucial to the performance of the estimation. In this section, we develop a data-driven approach to choose the smoothing parameters. The Kullback–Leibler (KL) loss is used to evaluate the quality of a density estimate, and is estimated using cross-validation.

Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)$. Denote the estimate (3.6) as $\eta_{\boldsymbol{\lambda}}$, where the dependence on the smoothing parameters is expressed explicitly. Let $B_{\omega x} =$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

$\{b_\omega(y,x),\ y \in \mathcal{Y}\}$ be a Gaussian stochastic process, given ω and x. Denote the true and estimated cluster-specific conditional densities as $f(y|x,\omega) = \exp\{\eta(y,x) + b_\omega(y,x)\}/\int_{\mathcal{Y}} \exp\{\eta(t,x) + b_\omega(t,x)\}dt$ and $f_{\boldsymbol{\lambda}}(y|x,\omega) = \exp\{\eta_{\boldsymbol{\lambda}}(y,x) + b_\omega(y,x)\}/\int_{\mathcal{Y}} \exp\{\eta_{\boldsymbol{\lambda}}(t,x) + b_\omega(t,x)\}dt$, respectively. We define the aggregated KL loss of $f_{\boldsymbol{\lambda}}(y|x,\omega)$ as

$$AKL(f, f_{\boldsymbol{\lambda}}) = \int_\Omega \int_{\mathcal{X}} f(x) \mathrm{E}_{B_{\omega x}} \left\{ \int_{\mathcal{Y}} f(y|x,\omega) \log\left(\frac{f(y|x,\omega)}{f_{\boldsymbol{\lambda}}(y|x,\omega)} \right) dy \right\} dx dP, \tag{4.1}$$

where $f(x)$ is the density function of X, and $\mathrm{E}_{B_{\omega x}}$ is the expectation with respect to the Gaussian process, given ω and x. After removing terms that are independent of the estimate $f_{\boldsymbol{\lambda}}$, the aggregated relative KL loss is

$$\begin{aligned} ARKL(f, f_{\boldsymbol{\lambda}}) &= \int_\Omega \int_{\mathcal{X}} f(x) \mathrm{E}_{B_{\omega x}} \left\{ \int_{\mathcal{Y}} f(y|x,\omega) \log\left(\frac{1}{f_{\boldsymbol{\lambda}}(y|x,\omega)} \right) dy \right\} dx dP \\ &= \int_\Omega \int_{\mathcal{X}} f(x) \mathrm{E}_{B_{\omega x}} \left\{ \log \int_{\mathcal{Y}} \exp[\eta_{\boldsymbol{\lambda}}(t,x) + b_\omega(t,x)]dt \right\} dx dP \\ &\quad - \int_\Omega \int_{\mathcal{X}} f(x) \mathrm{E}_{B_{\omega x}} \left\{ \int_{\mathcal{Y}} f(y|x,\omega) \eta_{\boldsymbol{\lambda}}(y,x) dy \right\} dx dP. \end{aligned} \tag{4.2}$$

Ideally, we want to select the smoothing parameters $\boldsymbol{\lambda}$ that minimize (4.1). This is equivalent to minimizing (4.2), because the aggregated KL loss and the relative aggregated KL loss differ only by a constant, independent of $\boldsymbol{\lambda}$. However, depending on the unknown density functions $f(x)$ and $f(y|x,\omega)$, $ARKL(f, f_{\boldsymbol{\lambda}})$ cannot be calculated directly. Using empirical distributions, the first term of (4.2) can be approximated by $N^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathrm{E}_{B_{ij}} \log \int_{\mathcal{Y}} \exp\{$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

$\eta_{\boldsymbol{\lambda}}(y, X_{ij}) + b_{ij}(y, X_{ij})\}dy$, where $\mathrm{E}_{B_{ij}}$ is the expectation with respect to the Gaussian process, given $\omega = \omega_i$, and $X = X_{ij}$. The second term in (4.2) can be approximated by $N^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\eta_{\boldsymbol{\lambda}}^{[(i,j)]}(Y_{ij}, X_{ij})$, where $\eta_{\boldsymbol{\lambda}}^{[(i,j)]}(Y_{ij}, X_{ij})$ is the estimate that minimizes the delete-one-observation version of (3.7). Hence, we may select the smoothing parameters by minimizing the following cross-validation estimate of (4.2):

$$\begin{aligned} CV(\boldsymbol{\lambda}) &= \tfrac{1}{N}\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathrm{E}_{B_{ij}}\log\int_{\mathcal{Y}}\exp\{\eta_{\boldsymbol{\lambda}}(y, X_{ij}) + b_{ij}(y, X_{ij})\}dy \\ &\quad -\tfrac{1}{N}\textstyle\sum_{i=1}^{m}\sum_{j=1}^{n_i}\eta_{\boldsymbol{\lambda}}^{[(i,j)]}(Y_{ij}, X_{ij}). \end{aligned} \tag{4.3}$$

The computation of $\eta_{\boldsymbol{\lambda}}^{[(i,j)]}(Y_{ij}, X_{ij})$ is costly. Using a quadratic approximation, (4.3) can be approximated by

$$\begin{aligned} CV_{\alpha}(\boldsymbol{\lambda}) &= \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\mathrm{E}_{B_{ij}}\log\int_{\mathcal{Y}}\exp\{\eta_{\boldsymbol{\lambda}}(y, X_{ij}) + b_{ij}(y, X_{ij})\}dy \\ &\quad -\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\eta_{\boldsymbol{\lambda}}(Y_{ij}, X_{ij}) + \alpha\frac{\mathrm{tr}(P_1^{\perp}\breve{R}^T\Pi^{-1}\breve{R}^T P_1^{\perp})}{N(N-1)}, \end{aligned} \tag{4.4}$$

where a constant α is added to avoid potential under-smoothing. Gu (2013, Ch7) suggests using an α-value of about 1.4 for various density estimation problems. The derivation of the approximation (4.4) and the definitions of matrices $P_1^{\perp}$, $\breve{R}$, and Π can be found in the Supplementary Material S4. The optimal smoothing parameters are chosen to minimize of the approximated CV score (4.4).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

5. Simulations

We conduct simulations to evaluate the proposed estimation method and to compare the estimates of cluster-specific conditional densities from an NMECDM with those from separate fits. We present the simulation results for the conditional density models only. The results for the density models are similar, and can be found in Chiu (2015).

We set $\mathcal{Y} = [0, 1]$ and consider two simulation scenarios for x: discrete, with $x = 0.25$ or $x = 0.5$ representing two groups, and continuous, with x taking six equally spaced values in [0.1, 0.9] (i.e., $x = .1 + .16 \times (k - 1)$, for $k = 1, \cdots, 6$).

For the discrete case, we consider the two-dimensional Euclidean space R^2 as the model space for x, and the Soblev space for cubic splines

$$W_2^2[0,1] = \{f : f \text{ and } f' \text{ are absolutely continuous, } \int_0^1 (f^{(2)})^2 dt < \infty\}$$

as the model space for y. The SS ANOVA decomposition leads to the following NMECDM:

$$\begin{aligned} g(y, x, \omega) &= d_2(y - 0.5) + g_{lsf}(y, x) + g_{scf}(y) + g_{ssf}(y, x) \\ &\quad + b_2(\omega)(y - 0.5) + g_{scr}(y, \omega), \end{aligned} \tag{5.1}$$

where the first four terms come from the tensor product of $W_2^2[0, 1] \otimes R^2$, with terms independent of y removed for identifiability. The letters `"c"`,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

"l", and "s" in the subscripts represent the constant, linear, and smooth components, respectively, $b_2(\omega)$ are independent and identical distributed (i.i.d) $N(0, \sigma_1^2)$, $g_{scr}(y, \omega)$ are i.i.d. Gaussian stochastic processes with mean zero and covariance function $\mathrm{Cov}(g_{scr}(y_1, \omega), g_{scr}(y_2, \omega)) = \sigma_2^2 R(y_1, y_2)$, where R is the RK of $W_2^2[0,1] \ominus \{1, y-0.5\}$, and $b_2(\omega)$ and $g_{scr}(y, \omega)$ are mutually independent.

For the continuous case, we consider the following NMECDM:

$$\begin{aligned} g &= d_2(y-0.5) + d_3(x-0.5)(y-0.5) + g_{lsf}(y,x) \\ &\quad + g_{scf}(y) + g_{slf}(y,x) + g_{ssf}(y,x) + b_2(\omega)(y-0.5) + g_{scr}(y,\omega), \end{aligned} \tag{5.2}$$

where the first six terms come from the tensor product of $W_2^2[0,1] \otimes W_2^2[0,1]$, with terms independent of y removed for identifiability, $b_2(\omega)$ are i.i.d. $N(0, \sigma_1^2)$, $g_{scr}(y, \omega)$ are i.i.d. Gaussian stochastic processes with mean zero and covariance function $\mathrm{Cov}(g_{scr}(y_1, \omega), g_{scr}(y_2, \omega)) = \sigma_2^2 R(y_1, y_2)$, and $b_2(\omega)$ and $g_{scr}(y, \omega)$ are mutually independent.

For both cases, we set the fixed effect in the NMECDM (2.5) $\eta(y, x) = -18(y-x)^2$. For the discrete case, this corresponds to setting $d_2 = 0$, $g_{lsf}(y, x) = 36(x-0.5)(y-0.5)$, $g_{scf}(y) = -18y^2+18y-3$, and $g_{ssf}(y, x) = 0$ in model (5.1). For the continuous case, this corresponds to setting $d_2 = 0$, $d_3 = 36$, $g_{slf}(y, x) = g_{lsf}(y, x) = g_{ssf}(y, x) = 0$, and $g_{scf}(y) = -18y^2 +$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

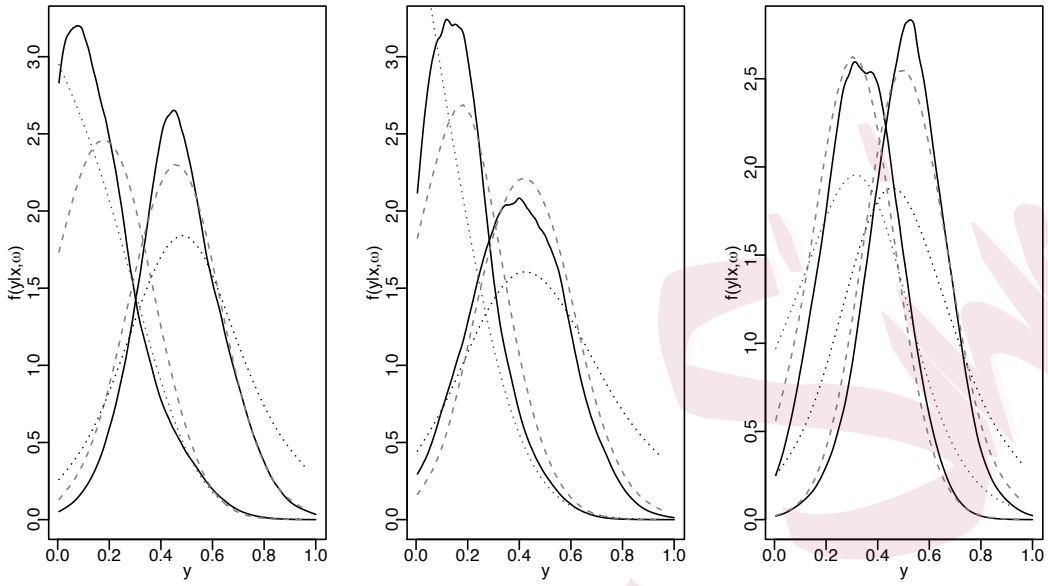

Figure 1: Typical fits of the cluster-specific conditional density functions when x is discrete and $m = 100$. Three clusters are selected randomly. The solid curves are the true cluster-specific conditional densities. The dashed curves are fits based on the NMECDM (5.1). The dotted curves are separate cubic spline estimates based on each cluster's data only.

$18y - 3$ in model (5.2). There are two random effects, $b_2(\omega)$ and $g_{scr}(y, \omega)$, for both cases, generated using $\sigma_1^2 = 5$ and $\sigma_2^2 = 50$.

We consider three sizes of m: $m = 100$, $m = 200$, and $m = 600$. For each cluster i, n_i is the nearest integer of a random number generated from a normal distribution with mean 10 and standard deviation 3. We fit models (5.1) and (5.2) for the discrete and continuous cases, respectively.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

All simulations are replicated 100 times.

We adopt the MCMC SAA scheme (i) with $m_{10} = 50$ to estimate the fixed effects, and $m_{20} = 500$ to estimate the variance components of the random effects. The burn-in phase is chosen as the first 200 MCMC samples, and thinning is performed every 10 MCMC samples. The estimates of the fixed effects converge much faster than those of the variance components. Specifically, they usually converge within 10 iterations. To expedite the computation, we fix the estimates of $\mathbf{c}$ and $\mathbf{d}$ after 10 iterations and let the algorithm run until $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ converge. The stopping criterion is defined as the relative difference between the consecutive estimates $\hat{\sigma}_i^{2(j)}$ at iteration j and $\hat{\sigma}_i^{2(j-1)}$ at iteration $j-1$; specifically,

$$\delta = \frac{\sqrt{(\hat{\sigma}_1^{2(j)} - \hat{\sigma}_1^{2(j-1)})^2 + (\hat{\sigma}_2^{2(j)} - \hat{\sigma}_2^{2(j-1)})^2}}{\sqrt{\hat{\sigma}_1^{4(j-1)} + \hat{\sigma}_2^{4(j-1)}}}.$$

The iterations stop if δ is less then 5e-4.

To evaluate the estimation of variance of the components, we compute the means and MSEs of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. To evaluate the estimation of the population conditional density, we compute the empirical aggregated KL distances

$$\mathrm{AKL}(f(y|x), \hat{f}(y|x)) = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \log(f(Y_{ij}|X_{ij}) / \hat{f}(Y_{ij}|X_{ij})) f(Y_{ij}|X_{ij}).$$

To evaluate the estimation of cluster-specific conditional densities, we com-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

pute the empirical aggregated KL distances

$$\begin{aligned} & \mathrm{AKL}(f(y|x,\omega), \hat{f}(y|x,\omega)) \\ = \; & \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \log(f(Y_{ij}|X_{ij},\omega_i)/\hat{f}(Y_{ij}|X_{ij},\omega_i))f(Y_{ij}|X_{ij},\omega_i). \end{aligned}$$

For comparison, we compute cubic spline conditional density estimates for each cluster separately, using the *sscden* function in the *R* package *gss* (Gu, 2014), and the empirical aggregated KL distances between these cluster-specific conditional densities and the truth.

In Figures A.1 and A.2 of the Supplementary Material, we show the true population density functions and three estimates for discrete and continuous x, respectively, when $m = 200$. We observe that the population density functions are estimated accurately. For three randomly selected clusters, Figure 1 shows the true cluster-specific density functions and estimates based on model (5.1), and the cubic spline estimates based on each cluster's data when x is discrete and $m = 100$. The cluster-specific density estimates from the NMECDM are shrink toward the population conditional density. Table 1 indicates that, by borrowing information across clusters, the cluster-specific density estimates from the NMECDM have about half the AKL losses of the estimates based on individual data. Table 1 also indicates the convergence of the estimates of the variance components as m increases, although the convergence is slow. Overall, the proposed estima-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

Covariate x	Quantities	$m = 100$	$m = 200$	$m = 600$
	AKL$(f, \hat{f})$	0.0027	0.0018	0.0006
	Mean (MSE) of $\hat{\sigma}_1^2$	4.38 (1.49)	4.72 (0.51)	4.83 (0.33)
Discrete	Mean (MSE) of $\hat{\sigma}_2^2$	48.59 (16.22)	49.84 (4.92)	49.70 (2.98)
	AKL_1	0.0628	0.0627	0.0613
	AKL_2	0.1284	0.1299	0.1224
	AKL$(f, \hat{f})$	0.0027	0.0019	0.0011
	Mean (MSE) of $\hat{\sigma}_1^2$	4.73 (1.05)	4.92 (0.39)	5.05(0.14)
Continuous	Mean (MSE) of $\hat{\sigma}_2^2$	49.43 (19.13)	49.74 (4.55)	49.68 (4.39)
	AKL_1	0.0332	0.0327	0.0214
	AKL_2	0.0709	0.0717	0.0711

Table 1: Summary of simulation results. AKL$(f, \hat{f})$ represents the average of the AKL loss between the true population density function f and its estimate $\hat{f}$ based on an NMECDM. AKL_1 represents the average of the AKL losses between each cluster's density and its estimate based on an NMECDM. AKL_2 represents the average of the AKL losses between each cluster's density and its cubic spline estimate using data from this cluster only.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

tion method performs well. The estimation procedure is computationally intensive. When the number of observations is large, as in the real-data analysis in the following section, one may use the divide-and-recombine approach (Cleveland and Hafen, 2014) to reduce the computational burden.

6. Evolution of Hb Distributions Over Time

Anemia is prevalent in the majority of hemodialysis patients and its management is a major challenge. A central aim of anemia management is to maintain patients' Hb levels consistently within a target range. Both low and high Hb are associated with increased risk of mortality and hospitalization. However, the optimal target range has been the subject of much debate, and anemia management guidelines and protocols have changed in recent years (Spiegel et al., 2010; Valliant and Hofmann, 2013). The current optimal range for Hb recommended by the Food and Drug Administration, is 10–12 g/dL (Spiegel et al., 2010). The Centers for Medicare and Medicaid Services (CMS) introduced a Quality Incentive Program (QIP) with anemia management as one of the four outcomes, measured as the percentage of patients in a dialysis facility with Hb greater than 12 g/dL. Facilities that do not meet these standards have their payments reduced by up to 2%. In addition to the optimal range, greater Hb variation is also associated with higher mortality (Yang et al., 2007). Spiegel et al. (2010) noted that

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

the "Hb distribution curve showed a departure from normality in terms of skewness," and that is of interest to investigate how the mean, standard deviation, skewness, and percentage of Hb over/under a certain limit change over time. Therefore, it is important to investigate the whole distribution of Hb and its evolution over time in response to guideline changes without parametric assumptions about the distribution. Fitting a density function at each month, previous approaches tend to ignore the longitudinal nature of the data (e.g., Spiegel et al. (2010)).

Monthly Hb measurements were collected from 200,525 dialysis patients in 811 facilities for the period January 2010 to December 2013. Patients in a given facility may vary from month to month owning to the arrival of new patients and the loss of current patients. Nevertheless, Hb measurements over time from the same facility are likely to be correlated, owning to multiple contributions from the same patients, common practices in Hb management, and patients with a similar demographical background. We are interested in how the distribution of Hb changes over time. Consequently, we fit an NMECDM, with Hb as the dependent variable, time as the independent variable, and facilities as clusters.

We transform both the time and the Hb measurements into [0,1]. Let x be the transformed time variable and y be the transformed Hb measure-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

ments. We consider the model (5.2). Owning to the complexity of the model and the large number of Hb measurements (2,800,430), it is computationally infeasible to fit the NMECDM (5.2) to the whole data set. Therefore, we use the divide-and-recombine approach (Cleveland and Hafen, 2014): we randomly split the 811 facilities into eight subsets and fit model (5.2) to each subset. Rather than combining the estimates to form a single final estimate, we report the estimates from all eight subsets. Because eight subsets may be regarded as random samples from the population, estimates from these subsets will allow us to explore the variation in the estimations of the parameters and functions.

We show the estimates of the variance components from the eight subsets in Table A.1 of the Supplementary Material. It is clear that the estimates of σ_1^2 and σ_2^2 from different subsets are quite close, except for those from subset 2.

The estimates of the population density function from the eight subsets are very close (not shown). The solid lines in Figure 2a show the combined estimates of the population density function as averages of the estimates from the eight subsets in January of each year. The estimated quantiles clearly show that the distribution of Hb shifted downward in response to guideline changes. Figure 2a also shows the cluster-specific density esti-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

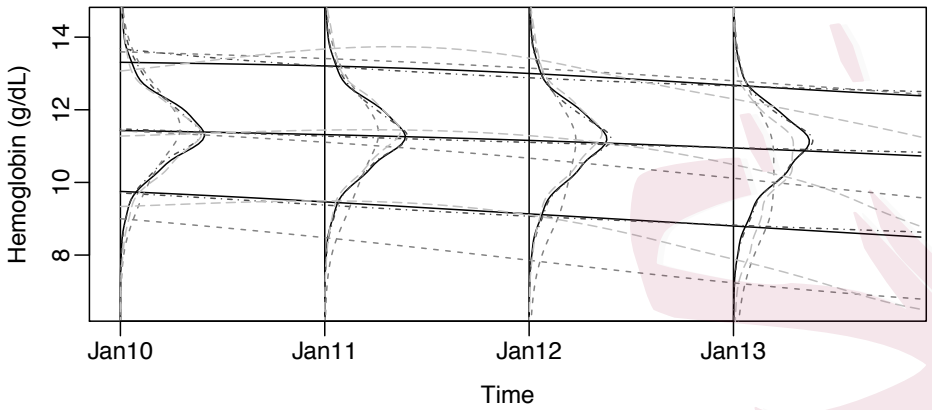

Figure 2a: Population density estimate (solid) in January of each year and density estimates of three facilities corresponding the minimum (dot-dash), median (long dash), and maximum (short dash) AKL from the population density estimate. The horizontal lines are the corresponding estimated 5%, 50%, and 95% quantiles. The population density estimate is obtained by averaging the estimates from the eight subsets.

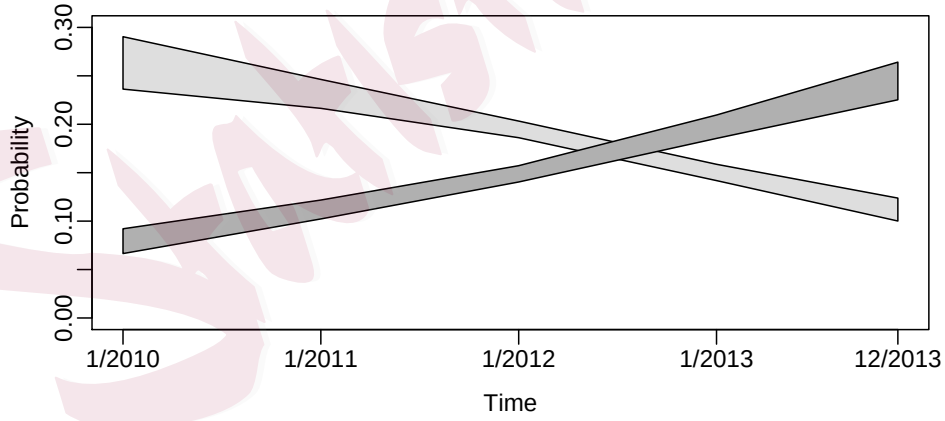

Figure 2b: Envelopes of the probabilities of Hb greater than 12 g/dL (light gray) and the probabilities of Hb smaller than 10 g/dL (darker gray).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

mates based on January of each year from three facilities. The trajectories of facility-specific density functions are useful for identifying facilities with poor management of Hb. For example, the facility with a maximum AKL from the population density estimate (short dashed line in Figure 2a) has heavier tails and became more skewed toward smaller values.

Policymakers are interested in the population probability of Hb outside the target range of 10-12 g/dL. We compute the probabilities of Hb greater than 12 g/dL and those of the probabilities of Hb smaller than 10 g/dL based on the estimated conditional densities from each subset. Figure 2b shows the envelopes (i.e., from the minimum to the maximum from the eight subsets at each time point) of these two probabilities. It is clear that the guideline changes have effectively reduced the probability of Hb over 12 g/dL from .25 in January 2010 to .1 in December 2013. However, they also increased the probability of Hb under 10 g/dL from .08 to over .2 for the same period. This is not surprising because the decrease in the Hb level is a result of the reduction of the erythropoiesis-stimulating agent's dosage. Our conclusions are in good agreement with those of Spiegel et al. (2010), who studied dialysis patients from June 2006 to November 2008. One important new finding is that the probability of Hb under 10 g/dL has been increased significantly in recent years. Consequently, further dialysis

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

patients may suffer from anemia. This unintended consequence should be investigated further.

7. Conclusion

We have introduced general density and conditional density models with random effects for clustered data, and illustrated the construction of these models using SS ANOVA decompositions. Note that other approaches may be used to construct these models. The proposed NMEDMs and NMECDMs are flexible because the domains of both the dependent and the independent variables are arbitrary sets, and different RKHSs and decompositions may be used to construct these models. As illustrated in Section 2, the classical mixed-effects models and SS ANOVA mixed-effects models with Gaussian distributions are special cases of the NMEDM and NMECDM. Therefore, in addition to nonparametric estimations of density and conditional density functions with clustered data, our methods provide potential model building and diagnostic tools for existing mixed-effects models with Gaussian random errors. Model-selection methods for SS ANOVA density models based on the KL projection have been developed by Gu (2013). Further research on model selection and inferences for mixed-effects density models is merited.

Parameters and nonparametric functions are estimated using the penal-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

ized likelihood. We have developed a computation procedure using MCMC SAA and an approximated cross-validation criterion to select the smoothing parameters. Extensive simulations indicate that our estimation procedure is stable. However, the estimates of the variance components may converge slowly. In addition, the estimates of the variance components have a relatively large bias when the sample size is small, which is a common problem with MLEs in mixed-effects models. The adjusted profiled likelihood (McCullagh and Tibshirani, 1990) or bias-reducing penalized likelihood (Kosmidis et al., 2015) may be used to reduce the bias in the estimates of the variance components. Involving integrations with respect to random effects, the marginal likelihood function is not guaranteed to be convex, which makes it very difficult to derive the asymptotic properties. An alternative approach is to use the joint (Henderson) likelihood of observations and random effects, as in Gu and Ma (2005) and Gu (2013). We will explore these topics in future research.

We have applied our methods to investigate the changes in Hb distributions over time. We found that guideline changes have shifted the Hb distributions downward. On the one hand, the probability of Hb over 12 g/dL has been reduced greatly. On the other hand, the probability of Hb under 10 g/dL has been increased substantially, raising concerns that the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

proportion of dialysis patients who suffer from anemia may have increased. The resulting impacts on mortality, hospitalization, cost, and quality of life require further investigation.

Supplementary Material The online Supplementary Material contains derivations not included in the paper.

Acknowledgements

We gratefully acknowledge the support provided by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (Chi-Yang Chiu) and the National Science Foundation (DMS-1507078 for Anna Liu and DMS-1507620 for Yuedong Wang). We thank Fresenius Medical Care North America for the deidentified data and the collaboration in this analysis. We acknowledge support from the Center for Scientific Computing from the CNSI, MRL: an NSF MRSEC (DMR-1720256).

References

Chiu, C. (2015). Nonparametric mixed-eects density regression, Ph.D. Thesis, University of California-Santa Barbara, Dept. of Statistics and Applied Probability.

Cleveland, W. S. and Hafen, R. P. (2014). Divide and recombine (D&R): Data science for large complex data, *Statistical Analysis and Data Mining* **7**: 425–433.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

Dunson, D. B., Pillai, N. and Park, J. H. (2007). Bayesian density regression, *Journal of the Royal Statistical Society B* **69**: 163–183.

Efromovich, S. (2007). Conditional density estimation in a regression setting, *Annals of Statistics* **35**: 2504–2535.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities, *Biometrika* **91**: 819–834.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis*, Chapman and Hall, London.

Gu, C. (2013). *Smoothing Spline ANOVA Models, 2nd ed.*, Springer-Verlag, New York.

Gu, C. (2014). Smoothing spline ANOVA models: R package gss, *Journal of Statistical Software* **58**(1): 1–25.

Gu, C. and Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models, *Annals of Statistics* **33**: 1357–1379.

Gu, C. and Wang, J. (2003). Penalized likelihood density estimation: Direct cross validation and scalable approximation, *Statistica Sinica* **13**: 811–826.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo for incomplete data estimation problems, *Proceedings of the National Academy of Sciences* **95**: 7270–7274.

Gu, M. G. and Zhu, H. T. (2001). Maximum likelihood estimation for spatial models by Markov

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

chain Monte Carlo stochastic approximation, *Journal of the Royal Statistical Society B* **63**: 339–355.

Hall, P., Racine, J. and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities, *Journal of the American Statistical Association* **99**: 1015–1026.

Jiang, Y., Karcher, P. and Wang, Y. (2011). On implementation of the Markov chain Monte Carlo stochastic approximation algorithms, *Advances in Directional and Linear Statistics, A Festschrift for Sreenivasa Rao Jammalamadaka, M Wells and A Sengupta (eds).*

Kosmidis, I., Guolo, A. and Varin, C. (2015). Improving the accuracy of likelihood-based inference in meta-analysis and meta-regression, arXiv:1509.00650v1.

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihood, *Journal of the Royal Statistical Society B* **52**: 325–344.

Petersen, H. and Müller, H. G. (2016). Functional data analysis for density functions by transformation to a Hilbert space, *Annals of Statistics* **44**: 183–218.

Rodriguez, A., Dunson, D. B. and Taylor, J. (2009). Bayesian hierarchically weighted finite mixture models for samples of distributions, *Biostatistics* **10**: 155–171.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method, *Annals of Statistics* **12**: 898–916.

Spiegel, D. M., Khan, I., Krishnan, M. and Mayne, T. J. (2010). Changes in hemoglobin level distribution in us dialysis patients from june 2006 to november 2008, *American Journal of*

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

Kidney Diseases **55**: 113–120.

Valliant, A. and Hofmann, R. M. (2013). Managing dialysis patients who develop anemia caused by chronic kidney disease: focus on peginesatide, *International Journal of Nanomedicine* **8**: 3297–3307.

Wang, Y. (1998). Mixed-effects smoothing spline ANOVA, *Journal of the Royal Statistical Society B* **60**: 159–174.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*, Chapman and Hall, New York.

Wu, H. and Zhang J. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*, John Wiley & Sons.

Yang, W., Israni, R. K., Brunelli, S. M., Joffe, M. M. Fishbane, S. and Feldman, H. I. (2007). Hemoglobin variability and mortality in ESRD, *Journal of the American Society of Nephrology* **18**: 3164–3170.

Chi-Yang Chiu

Division of Biostatistics, Department of Preventive medicine, University of Tennessee Health Science Center, Memphis, TN 38163

E-mail: chiu@uthsc.edu

Anna Liu

Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01002

E-mail: anna@math.umass.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0211

Yuedong Wang

Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106

E-mail: yuedong@pstat.ucsb.edu