A Simple Spline-based Multilevel Modeling Approach for Complex Data

Youngdeok Hwang[1], Eun Ryung Lee*[2], Daegon Cho[3] and Qiong Zhang[4]

[1]Department of Information Systems and Statistics, City University of New York

[2]Department of Statistics, Sungkyunkwan University

[3]KAIST College of Business, Seoul, Korea

[4]School of Mathematical and Statistical Sciences, Clemson University

Abstract

This paper introduces a spline-based multilevel approach for effectively analyzing complex data collected from multiple sources. The proposed method is computationally efficient and easy to implement, and particularly useful for analyzing large scale data that have hierarchical structure with varying complexity at different levels. Also developed is estimation procedure combining Expectation-Maximization algorithm and nonparametric regression approach. This novel estimator is shown to have a good property in that it belongs to a well-known class of linear smoother. Thus, further statistical inference can be easily adopted from the existing literature on linear smoothers. The theoretical properties of the proposed methods are provided. The effectiveness of the approach is illustrated using the music concert data collected from various sources.

*Corresponding author:erlee@skku.edu

KEY WORDS: Expectation-Maximization, hurdle model, nonparametric regression, latent-variable, large-scaled data, linear smoother, smoothing spline.

1 Introduction

With the increased availability of the various data sources, it has become a norm to join and merge different datasets for a study. The examples include joining the areal demographic information from the administrative sources with the spatial observation (Cho et al., 2018), combining the high dimensional numerical weather prediction model outputs with the monitoring station observations (Liu et al., 2016), or merging the point referenced meta data with the automated sensor collected observations (Hwang et al., 2018).

Using the data collected from various sources, however, poses a few challenges. First, the data structure is often complex. Hierarchical structures often exist, and the availability of data set also varies at different levels. Hence the modeling methodologies must be expeditious with minimal computational failures, while accommodating complex relationships. Second, traditional assumptions on the sampling process are difficult to justify. Often, the data is a particular subset chosen by a subjective or practical collection criteria. For example, in our case study in Section 5, we chose the musicians that have made to a top 100 annual chart in North American during the study period, because it is impossible to consider all musicians. Third, the volume of the data is larger than the data collected by traditional methods such as polls and experiments. Automated tools and online platforms, which require a marginal cost for obtaining extra data, collect data for the analysis.

The applications using the data collected from various sources range from the service industry to government platforms. The examples include, recommendation system, and social media analysis using customer transaction records, e-government using citizen engagement and participation using rich textual content, or healthcare decision support using electronic health records (Chen

et al., 2012). However, the analysis conducted by the domain researchers often fail to incorporate this fundamentally different nature of the data that is automatically collected from online platform.

In this work, we propose a statistical framework to analyze the data merged from various sources. The proposed model uses a multilevel model (Gelman, 2006) under nonparametric functional structural model assumption. The multilevel models provide a flexible modeling tool for analyzing complex data. They allow an additional modeling layer for considering the sampling property of data at hand, where the traditional assumptions does not seem to hold. For example, Hwang et al. (2018) use this multilevel modeling approach for making solar power forecasts using a parametric model, to avoid merging and transferring a large scale data that are collected and stored. In contrast, our approach considers a underlying nonparametric regression model under an additional layer for taking the complicated nature of such a data collection process into account. For example, we have adopted the hurdle model (Mullahy, 1986) in our music concert data example of Section 5 to address the sampling process that can adversely affect the estimation.

Nonparametric modeling approach provides a useful extension of parametric models for hierarchical structure, by allowing an additional, flexible smoothness of the parameters. For example a smoothly varying pattern of the coefficients can be imposed in linear regression setting. For an overview of this area, see Wasserman (2006) and references therein. Most literature makes standard assumptions on data structure such as random sample or well-defined time series data. However, these assumptions can be restrictive, especially for complex data because they are often non-standard structure. Considering nonparametric regression approach under multilevel model can be appealing in such applications. To the best of our knowledge, this multilevel model approach has not been considered in the nonparametric literature. Our aim is to develop a statistical procedure for nonparametric modeling approach on the underlying regression function under the multilevel model structure, as well as deriving the theoretical properties of such procedure.

Smoothing spline has gained popularity as a nonparametric technique in various applications (Wahba, 1990; Wang, 2011; Gu, 2013). The smoothing spline has several appealing properties.

For example, it provides an intuitive Bayesian interpretation, and is a linear smoother as in the case for the least square method in linear regression. Those properties allow developing a further statistical inference method, such as tuning parameter selection (Buja et al., 1989) and Bayesian confidence intervals (Wahba, 1983; Nychka, 1988). Asymptotic properties have been studied as well. Those studies mainly rely on the theory of empirical process and reproducing kernel Hilbert space, and have studied the estimation performance, e.g. the global rate of convergence (e.g., Cox, 1983; van de Geer, 2000; Eggermont and LaRiccia, 2006; Eggermont et al., 2010). Further, more advanced technical analyses for asymptotic inference in smoothing spline models have been done (Silverman, 1984; Messer and Goldstein, 1993; Shang and Cheng, 2013).

Under a multilevel model, we have developed a new methodology to estimate the underlying latent smooth function using the Expectation-Maximization (EM) algorithm (e.g., Dempster et al., 1977) and nonparametric smoothing spline regression (Eubank, 1988; Wahba, 1990; Green and Silverman, 1994). The proposed algorithm is computationally efficient and easy to use. We also show that the estimator is a linear smoother (Buja et al., 1989), which allows a further inference method that can be easily developed relying on the existing theory on linear smoothers. For example, adapted to our setting, we suggest generalized cross-validation for tuning parameter selection. We also have derived the theoretical properties of the proposed estimator. Even though we focus on the application on a large scaled music concert count data example, our model and method can be more generally applicable.

The remainder of the paper is organized as follows. Section 2 gives the proposed models and statistical procedures including tuning parameter selection. The theory behind our methodology is given in Section 3. Section 4 presents the simulation study to compare the proposed method with potential comparators. Section 5 presents a case study on a music concert data. Section 6 provides some concluding remarks.

2 Model and Methodology

In this section, we introduce the proposed modeling approach in the context of our application in music concert. Model estimation is described in more general terms as it can be applied more broadly.

2.1 Multilevel Model

In our study, the location and date of each concert event of a group of musicians are observed over the entire United States for a given time period. We count the number of events for every Core Based Statistical Area (CBSA, U.S. Census Bureau, 2012), $i = 1, \ldots, n$ and month $j = 1, \ldots, J$. Denote the number of concert events in the jth month $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})^\top$, with its expectation vector $\boldsymbol{\eta}_j = (\eta(\mathbf{x}_1; \boldsymbol{\beta}_j), \ldots, \eta(\mathbf{x}_n; \boldsymbol{\beta}_j))^\top = (\eta_{1j}, \ldots, \eta_{nj})^\top$ where η is a known link function, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^\top$ is the unknown time-varying parameter to be estimated, and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is the CBSA specific regional covariates.

We assume that the number of concert events follows

$$y_{ij} \sim f(y_{ij}; \mathbf{x}_i, \boldsymbol{\beta}_j) \tag{1}$$

where f is a parametric model. In our case study in Section 5, it is assumed to be hurdle model, but can be chosen flexibly depending on the context.

The time-varying parameter $\boldsymbol{\beta}_j$'s are assumed to be

$$\beta_{jk} = \mu_k(z_j) + \xi_{jk}, \quad \text{for} \quad k = 1, \ldots, p, \tag{2}$$

where the months are re-scaled to $z_1, \ldots, z_J \in [0, 1]$, $\mu_k(z_j)$ is the mean function of β_{jk} that belongs to a class of smooth function with respect to z_j, and ξ_{jk} is a random error following $N(0, \sigma^2)$ with a known $\sigma^2 > 0$. For simplicity of the presentation, we consider $p = 1$ and omit

the subscript k hereinafter, although the idea can be extended to the case of moderate $p > 1$ unless it causes a serious dimensionality problem; for example, thin plate splines (Wahba, 1990) can be considered for $p = 2$. Note that the second level model in (2) concerns the variation of the time-specific parameters deviating from the overall trend. This deviation reflects the natural variation between different months, as well as uncertainty within the month due to the aggregation of the individual events.

Our main interest is to investigate how the relationship between y_{ij} and $\mathbf{x}_i$ changes over time, because it can reveal the shift of overall industry environment during the study period. This can be achieved by finding the second level model, which captures the temporal change of β_j. The main challenge in this approach is that there is no direct measurement of β_j. We address this issue by combining the smoothing splines approach and an EM algorithm.

2.2 Estimation

To estimate the smooth function $\mu(\cdot)$ in (2), we consider the smoothing spline method. The smoothing spline estimate is the minimizer of the penalized sum of squares

$$\sum_{j=1}^{J} (\beta_j - \mu(z_j))^2 + \lambda \int_0^1 (\mu'')^2 , \tag{3}$$

with respect to a function $\mu \in \mathcal{W}^{(2,2)}$, where $\mathcal{W}^{(2,2)} = \{m \in L_2 : \int (m'')^2 < \infty\}$ is the Sobolev class of twice differentiable functions and λ is a penalty constant. The estimation problem in (3) can be rewritten as

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^J} g(\boldsymbol{\theta}; \boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{N\theta})^\top (\boldsymbol{\beta} - \boldsymbol{N\theta}) + \lambda \boldsymbol{\theta}^\top \boldsymbol{W\theta}, \tag{4}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ and $\boldsymbol{N}_{J\times J}{}^\top = (N_{ij})$, $N_{ij} = N_i(z_j)$, and $\boldsymbol{W} = (W_{ij})$, $W_{ij} = \int N_i'' N_j''$ with natural cubic splines N_j having $z_1, \ldots, z_J$ as knots (Eubank, 1988). Because

of this form, (3) is seen as Bayesian estimation with a partially improper prior (Wahba, 1978; Speckman and Sun, 2003; Gu, 2013).

The apparent challenge for this approach to be applied to our problem is that β_j's are not directly observable. To address this challenge, we propose a two-step approach based on the intuition that the information related to β_j can be obtained from the observations $(\boldsymbol{y}_j, \mathbf{X})$, where $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{nj})$ is the collection of the event counts for the jth month, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ the CBSA specific information.

The proposed approach proceeds in two steps. First, we obtain $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_J)^\top$ by individually fitting the model in (1) for each j, $j = 1, \ldots, J$. Since $\boldsymbol{\beta}$ is not observable, it is treated as a latent variable, where $\hat{\boldsymbol{\beta}}$ is treated as an observation for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^\top$. Once all $\hat{\beta}_1, \ldots, \hat{\beta}_J$ are obtained, define $h(\hat{\beta}_j|\beta_j)$, the probability density of the sampling distribution of $\hat{\beta}_j$. The uncertainty from using $\hat{\beta}_j$ instead of β_j is reflected in $h(\hat{\beta}_j|\beta_j)$. An alternative way to view $\hat{\beta}_j$ is that it is a surrogate for the β_j with a measurement error (Berry et al., 2002). We first assume that $h(\hat{\beta}_j|\beta_j)$ is asymptotically normal. It is not only because it is convenient but also it is common in practical situation. If the method of maximum likelihood is used to obtain $\hat{\beta}_j$, for example, $h(\hat{\beta}_j|\beta_j)$ can be assumed to be approximately $N(\beta_j, V_j)$, where V_j is the inverse of Fisher information matrix. Estimates obtained from other methods, such as M-estimators, also belong to this case, only with the different form of V_j (Van der Vaart, 1998).

Second, we estimate $\boldsymbol{\theta}$ using EM algorithm, by minimizing (3) with a consideration of the $h(\hat{\beta}_j|\beta_j)$. Specifically, the smoothing spline criterion in (4) can be viewed as the negative complete loglikelihood with $(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \sigma^2) \sim N(\boldsymbol{N\theta}, \boldsymbol{\Sigma})$ and $(\boldsymbol{\theta} \mid \sigma^2) \sim N(0, \sigma^2\lambda^{-1}\boldsymbol{W}^{-1})$, where $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_J$. Then we minimize

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^J} \int h(\hat{\boldsymbol{\beta}} \mid \boldsymbol{\beta}) \exp(-g(\boldsymbol{\theta}; \boldsymbol{\beta})/2) d\boldsymbol{\beta}, \tag{5}$$

whose minimizer is found by

$$\hat{\boldsymbol{\theta}} = \arg\min E\left[\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\top}\boldsymbol{V}^{-1}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+(\boldsymbol{\beta}-\boldsymbol{N\theta})^{\top}(\boldsymbol{\beta}-\boldsymbol{N\theta})+\lambda\boldsymbol{\theta}^{\top}\boldsymbol{W\theta}\mid\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\theta}}^{(m)},\sigma\right]. \tag{6}$$

It is straightforward to derive

$$(\boldsymbol{\beta}\mid\hat{\boldsymbol{\beta}};\boldsymbol{\theta},\sigma)\sim N\left(\boldsymbol{\beta}^{*},\boldsymbol{V}^{*}\right),$$

where $\boldsymbol{\beta}^{*}=(\boldsymbol{V}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}\left(\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{N\theta}\right)$ and $\boldsymbol{V}^{*}=(\boldsymbol{V}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}$. It is computationally convenient because it allows a closed form. To minimize (6), taking the derivative with respect to $\boldsymbol{\theta}$ gives

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\theta}}E\left[\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{\top}\boldsymbol{V}^{-1}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)+(\boldsymbol{\beta}-\boldsymbol{N\theta})^{\top}(\boldsymbol{\beta}-\boldsymbol{N\theta})+\lambda\boldsymbol{\theta}^{\top}\boldsymbol{W\theta}\mid\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\theta}}^{(m)},\sigma\right]\\&=E\left[\frac{\partial}{\partial\boldsymbol{\theta}}(\boldsymbol{\beta}-\boldsymbol{N\theta})^{\top}(\boldsymbol{\beta}-\boldsymbol{N\theta})+\lambda\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\theta}^{\top}\boldsymbol{W\theta}\mid\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\theta}}^{(m)},\sigma\right].\end{aligned}$$

By setting it to zero, it reduces to solving

$$\begin{aligned}&\boldsymbol{N}^{\top}\boldsymbol{N\theta}^{(m+1)}+\lambda\boldsymbol{W\theta}^{(m+1)}=\boldsymbol{N}^{\top}E\left[\boldsymbol{\beta}|\hat{\boldsymbol{\beta}};\hat{\boldsymbol{\theta}}^{(m)},\sigma\right]\\&=\boldsymbol{N}^{\top}\boldsymbol{\beta}^{*}=\boldsymbol{N}^{\top}\left(\boldsymbol{V}^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\hat{\boldsymbol{\theta}}^{(m)}\right).\end{aligned} \tag{7}$$

Hence M-step updates $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}}^{(m+1)}=\left(\boldsymbol{N}^{\top}\boldsymbol{N}+\lambda\boldsymbol{W}\right)^{-1}\boldsymbol{N}^{\top}\left(\boldsymbol{V}^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}}+\boldsymbol{\Sigma}^{-1}\boldsymbol{N}\hat{\boldsymbol{\theta}}^{(m)}\right).$$

The algorithm iterates until the convergence.

Taking $m\to\infty$ in (7) and rearranging terms, the final estimate at the end of EM algorithm

satisfies

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= [\boldsymbol{N}^\top\boldsymbol{N} + \lambda\boldsymbol{W} - \boldsymbol{N}^\top\left(\boldsymbol{V}^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}]^{-1}\boldsymbol{N}^\top\left(\boldsymbol{V}^{-1}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\boldsymbol{V}^{-1}\hat{\boldsymbol{\beta}} \\ &= [\boldsymbol{N}^\top\boldsymbol{D}\boldsymbol{N} + \lambda\boldsymbol{W}]^{-1}\boldsymbol{N}^\top\boldsymbol{D}\hat{\boldsymbol{\beta}} \qquad (8)\end{aligned}$$

with $\boldsymbol{D} = \boldsymbol{I} - \boldsymbol{V}^*\boldsymbol{\Sigma}^{-1} = \boldsymbol{V}^*\boldsymbol{V}^{-1}$ at the end of E-M iteration. The estimate accordingly can be written as

$$\hat{\mu}(z) = \sum_{j=1}^{J}\hat{\theta}_j N_j(z) \qquad (9)$$

for the smooth mean function. As such, the estimate $\hat{\mu}(\cdot)$ is a linear smoother for the response $\hat{\boldsymbol{\beta}}$, i.e., given z, there exists a vector $\boldsymbol{l}_\lambda(z) = (l_{\lambda,1}(z),\ldots,l_{\lambda,J}(z))$ such that $\hat{\mu}(z) = \sum_{j=1}^{J} l_{\lambda,j}(z)\hat{\beta}_j$. Thus, the fitted response is

$$\hat{\boldsymbol{\mu}} \equiv (\hat{\mu}(z_1),\ldots,\hat{\mu}(z_J))^\top = \boldsymbol{S}_\lambda\hat{\boldsymbol{\beta}}$$

with smoothing matrix $\boldsymbol{S}_\lambda = (\boldsymbol{l}_\lambda(z_1),\ldots,\boldsymbol{l}_\lambda(z_J))^\top$. The estimate $\hat{\mu}(\cdot)$ being a linear smoother provides a convenient way to conduct further inference, e.g., to choose the value of λ (Buja et al., 1989). For any value of λ, one may calculate Generalized Cross Validation (GCV) given by

$$\mathrm{GCV}(\lambda) = J^{-1}\sum_{j=1}^{J}\left(\frac{\hat{\beta}_j - \hat{\mu}(z_j)}{1-\mathrm{tr}(\boldsymbol{S}_\lambda)/J}\right)^2, \qquad (10)$$

because effective degree of freedom is $\mathrm{df}_\lambda = \mathrm{tr}(\boldsymbol{S}_\lambda)$ for a linear smoother estimate.

We have considered the case where $h(\hat{\beta}_j|\beta_j)$ is approximately normal so far. When it is not the case, our approach needs a slight modification because the posterior may not be available in a closed form. Observe that we only need to calculate $E[\boldsymbol{\beta}|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}^{(m)},\sigma]$ in (7) because of the form of the smoothing spline criterion at (4). In such cases, Monte Carlo EM (MCEM, Wei and Tanner, 1990) approach can be used, as summarized below.

With a sample $\boldsymbol{\beta}_1^{(m)},\ldots,\boldsymbol{\beta}_B^{(m)}$ generated from the distribution $(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}};\hat{\boldsymbol{\theta}}^{(m)},\sigma)$, $E[\boldsymbol{\beta}|\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}}^{(m)},\sigma]$

can be estimated by

$$\sum_{b=1}^{B} w_b \boldsymbol{\beta}_b^{(m)} / \sum_{b=1}^{B} w_b,$$

where w_b is the weight for the $\boldsymbol{\beta}_b^{(m)}$. Levine and Casella (2001) provides an extensive investigation of the issues associated with the implementation of MCEM. A straightforward choice of the weight is $w_b = h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_b)$. To see this,

$$\begin{aligned} E[\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\theta}}^{(m)}, \sigma] &= \int \boldsymbol{\beta} \;\; h(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}; \hat{\boldsymbol{\theta}}^{(m)}, \sigma) d\boldsymbol{\beta} \\ &= \int \boldsymbol{\beta} \;\; \frac{h(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)}, \sigma)}{\int h(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)}, \sigma) d\boldsymbol{\beta}} d\boldsymbol{\beta} \\ &= \frac{E_{\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)},\sigma}[\boldsymbol{\beta} h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta})]}{E_{\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)},\sigma}[h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta})]}. \end{aligned} \tag{11}$$

The last equality immediately follows from the facts that

$$\begin{aligned} \int \boldsymbol{\beta} h(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)}, \sigma) d\boldsymbol{\beta} &= \int \boldsymbol{\beta} h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}) h(\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)}, \sigma) d\boldsymbol{\beta} \\ &= E_{\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)},\sigma}[\boldsymbol{\beta} h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta})] \end{aligned}$$

and that $h(\hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}}^{(m)}, \sigma) = \int h(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}|\hat{\boldsymbol{\theta}}^{(m)}, \sigma) d\boldsymbol{\beta} = \int h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}) h(\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)}, \sigma) d\boldsymbol{\beta} = E_{\boldsymbol{\beta}|\hat{\boldsymbol{\theta}}^{(m)},\sigma}[h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta})]$. Thus, once a form of $h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta})$ is given, one can generate Monte Carlo samples $\boldsymbol{\beta}_b \sim (\boldsymbol{\beta}_b; \hat{\boldsymbol{\theta}}^{(m)}, \sigma)$ using the distribution specified by (4) in order to approximate the (11). Specifically,

$$E[\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}^{(m)}, \sigma] \approx \sum_{b=1}^{B} \boldsymbol{\beta}_b h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_b) / \sum_{b=1}^{B} h(\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}_b),$$

with a large B. Since this is a numerical integration, quasi Monte Carlo methods such as Latin hypercube sampling (McKay et al., 1979) or scrambled net quadrature (Owen et al., 1997) can be used to further improve the efficiency when needed.

Discussion so far has assumed σ^2 is known, but it hardly can be the case in practice. The value

of σ^2 is still critical because it is involved in the estimation process of $\hat{\boldsymbol{\theta}}$ via $\boldsymbol{D}$ in (8). We can estimate σ^2 along with the other parameters by iterating

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{(m+1)} &= [\boldsymbol{N}^\top \boldsymbol{D}^{(m)} \boldsymbol{N} + \lambda \boldsymbol{W}]^{-1} \boldsymbol{N}^\top \boldsymbol{D}^{(m)} \hat{\boldsymbol{\beta}}, \\
\hat{\mu}^{(m+1)}(\cdot) &= \sum_{j=1}^{J} \hat{\theta}_j^{(m+1)} N_j(\cdot), \\
\hat{\sigma}^{2\ (m+1)} &= \frac{1}{J} \sum_{j=1}^{J} (\hat{\beta}_j - \hat{\mu}^{(m+1)}(z_j))^2,
\end{aligned}$$

until convergence. Since the final estimates $\hat{\boldsymbol{\theta}}$ and $\hat{\mu}$ are same as in the above but with a replacement of the true σ^2 as the estimate $\hat{\sigma}^2$, the discussions based on linear smoother still hold, for example, GCV criterion is defined using the final $\hat{\mu}$ in the same manner as for (10).

One may consider a full Bayesian framework using the likelihood of $l\left(\beta_j; y_i, \mathbf{x}_i\right)$ and the Gaussian improper prior on β_j (e.g., Carpenter et al., 2017). In this work, our focus is on providing a simple, expeditious alternative.

3 Theoretical properties

In this section, we provide some theoretical properties of our proposed estimator in (9) obtained with (8). Without loss of generality, assume $\mathcal{I} = [0, 1]$, where $\mathcal{I} = \cup_{j=1}^{J} \mathcal{I}_j$. We show that our proposed estimator achieves the same global rate of convergence as in the nonparametric regression model (2), where the latent observation $\boldsymbol{\beta}$ could be directly available. Notice that our proposed estimator uses the noisy observation $\hat{\boldsymbol{\beta}}$, not $\boldsymbol{\beta}$.

For the derivation, we need a few technical assumptions.

(A1) $\max_{j=1,\ldots,J} V_j \ll \sigma^2$ in a sense that $\max_{j=1,\ldots,J} V_j/\sigma^2 \to 0$ as $n \to \infty$ $J \to \infty$.

This assumption indicates that the uncertainties coming from using the estimates $\hat{\beta}_j$ as a proxy for β_j vanishes uniformly in j in an asymptotic sense when sample size increases and the partition gets

finer. As the number of data gets larger, it is desirable and possible to use a finer set of partitions. The assumption implies that observations are obtained uniformly over the entire domain without any data availability problem. It is not very restrictive assumption in our setting, because the number of observations in each partition is much larger than the number of partitions. Within each partition, standard conditions such as well-conditioned design matrix in a linear regression problem (Shao, 2003) can be required to make a sufficient condition for A1. The number J of partitions should increase to infinity in order to approximate the nonparametric function μ on the whole domain $\mathcal{I} = [0, 1]$. And we also assume an additional regularity condition on nonparametric regression model (2):

(A2) The errors $\xi_j, j = 1, 2, \ldots, J$ are i.i.d. with $E(\xi) = 0$ and $\sigma^2 = E(\xi^2) < \infty$; the underlying smooth function μ is twice-differentiable with second derivative μ'' in $L_2(\mathcal{I}) = \{m : \int (m'')^2 < \infty\}$.

The following theorem can be proven applying standard arguments in the literature on nonparametric smoothing spline (e.g., Wahba, 1990; van de Geer, 2000; Eggermont and LaRiccia, 2006; Eggermont et al., 2010) to our setting. The proof of Theorem 1 is deferred to Appendix.

Theorem 1. *Suppose that (A1)-(A2) hold. Then,* $\|\hat{\mu} - \mu\| = O_p((J^{3/4}\lambda^{1/4})^{-1/2} + (\lambda/J)^{1/2})$, *moreover, if* $\lambda \asymp J^{1/5}$ *then* $\|\hat{\mu} - \mu\| = O_p(J^{-2/5})$.

Theorem 1 says that the proposed estimator $\hat{\mu}$ has the convergence rate $J^{-2/5}$ in L_2-sense. This rate is known to be optimal in univariate nonparametric regression (Stone, 1982) such as the model (2) with the unobservables $\boldsymbol{\beta}$. The uniform convergence rate of $\hat{\mu}$ in the following corollary is an immediate result of the theorem.

Corollary 1. *If the assumptions (A1)-(A2) hold, and* $\lambda \asymp J^{1/5}$, *then* $\max_z |\hat{\mu}(z) - \mu(z)| = O_p(J^{-3/10})$.

4 Simulation

In this section, we present numerical examples to corroborate the effectiveness of the proposed method and the associated theoretical results. We compare two other approaches with ours:

Single: Single level model with no multilevel structure, by estimating $\hat{\mu}(z_j) = \hat{\beta}_j$.

Naive: A two-step approach that is simpler than ours. First fit data model (1) for each partition in order to obtain $\hat{\beta}_j$ for $j = 1, \ldots, J$, then treat the estimates $\hat{\beta}_j$ for $j = 1, \ldots, J$ as observations to fit nonparametric regression (2) by minimizing (3) with $\boldsymbol{\beta}$ substituted by $\hat{\boldsymbol{\beta}}$.

Intuitively, single model does not incorporate information from the neighboring partitions. If the structural model in (2) is valid, the proposed method would benefit from using additional data. In particular, when some partitions suffer from data reliability issues, such as data corruption or limited data availability, the estimates can be poor. Similarly, naive approach does not properly incorporate the uncertainty that may differ for different partitions, it may also be sensitive to the data reliability issues. Proposed methodology can leverage the information from the neighboring partition via the structural model structure, while incorporating the uncertainty of the estimated $\hat{\beta}$ by adopting the sampling distribution, hence it can perform better when the multilevel model assumption is plausible.

For the simulation, we assume an interval is divided into equally spaced J partitions, and let $\mu(z_1), \ldots, \mu(z_J)$ be the true structural model values at the center point of J partition. We estimate $\hat{\mu}(z_1), \ldots, \hat{\mu}(z_J)$ and calculate the Root Mean Squared Error (RMSE) of $\hat{\mu_j}$, $\left\{ J^{-1} \sum_{j=1}^{J} \left(\mu(z_j) - \hat{\mu}(z_j) \right)^2 \right\}^{1/2}$ for each method. Two examples were considered to assess the performance of the proposed method in different scenarios. These test functions are depicted in Figure 1, where the first example is smooth over the interval, whereas the second is rugged. The benefit of the proposed method may depend on the data availability, hence we consider different data generation scenarios. Specifically, the sample size for each partition is randomly selected between $N_{\min}$ and 200 with equal probability for different values of $N_{\min}$. GCV in (10) is used as our selection criteria.

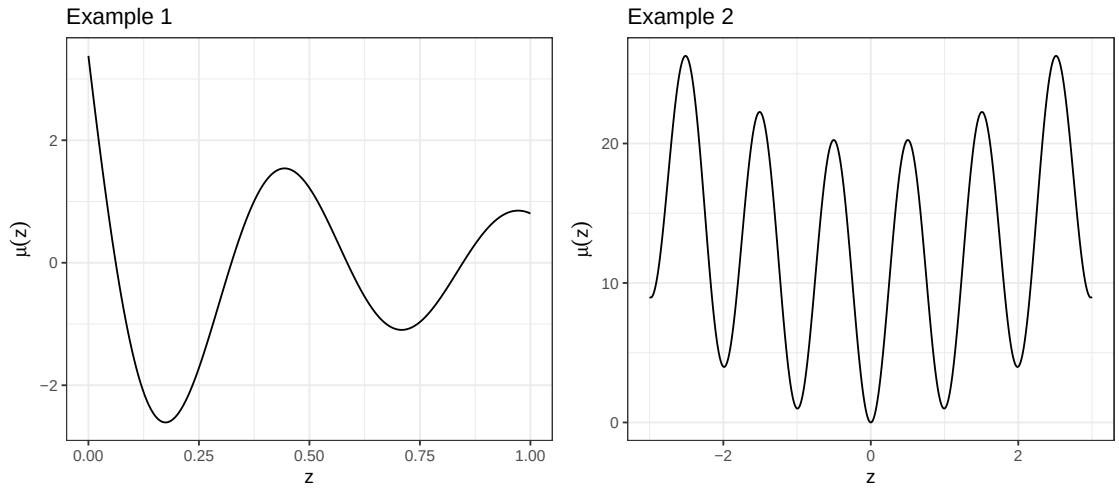

Figure 1: Two test functions used in the simulation study.

Example 1. In this example, we consider structural model

$$\mu(z) = \frac{\sin(12(z+0.2))}{z+0.2}$$

from Section 5.5 of Hastie et al. (2009) with $\beta_j \sim N(\mu(z_j), \tau^2)$ with varying values of τ. As the data model for y_{jk}, we consider two cases: (1) $y_{jk} \sim N(X_k\beta_j, \tau^2)$ with $\tau = 2, 4, 8$, for $k = 1, \ldots, n_j$, and $\beta_j \sim N(\mu(z_j), \sigma^2)$ with $\sigma = 0.5$; (2) $y_{jk} \sim$ Bern(p_{jk}), where $p_{jk} = (1 + \exp(-\eta_{jk}))^{-1}$ and $\eta_{jk} \sim N(X_k\beta_j, \tau^2)$ with $\tau = 2, 4, 8$, for $k = 1, \ldots, n_j$ and $\beta_j \sim N(\mu(z_j), \sigma^2)$ with $\sigma = 0.5$. The design matrix X_k is generated from $N(0, 1)$. The sample size for each partition n_j is chosen between $N_{\min}$ and 200 with the equal probability, where $N_{\min} = 50$ and $N_{\min} = 100$ is considered. When $N_{\min}$ is 50 (100), the sample size for the partition is more (less) imbalanced. We set $J = 50$ and $z_1, \ldots, z_J$ are equally spaced between 0 and 1. Simulations are repeated 200 times.

Table 1 presents the RMSE averaged over the 200 replicates. To calculate standard deviations (SD), RMSEs are calculated for each simulation run, and SD are calculated from 200 RMSEs. Clearly two methods with consideration with the structural model perform better compared to

Model	τ	$N_{\min}$	Single	Naive	Multilevel
LM	2	50	0.53 (0.05)	0.23 (0.05)	0.23 (0.05)
	2	100	0.52 (0.05)	0.23 (0.05)	0.23 (0.05)
	4	50	0.63 (0.06)	0.27 (0.06)	0.27 (0.06)
	4	100	0.60 (0.06)	0.26 (0.06)	0.26 (0.06)
	8	50	0.92 (0.09)	0.38 (0.09)	0.37 (0.08)
	8	100	0.83 (0.08)	0.35 (0.08)	0.35 (0.08)
GLM	2	50	0.62 (0.05)	0.53 (0.06)	0.56 (0.06)
	2	100	0.61 (0.05)	0.53 (0.06)	0.56 (0.06)
	4	50	0.86 (0.04)	0.84 (0.05)	0.86 (0.04)
	4	100	0.86 (0.04)	0.85 (0.04)	0.86 (0.04)
	8	50	1.07 (0.03)	1.08 (0.04)	1.08 (0.03)
	8	100	1.08 (0.03)	1.08 (0.03)	1.09 (0.03)

Table 1: Comparison of RMSE, where Mean and SD are calculated over 200 simulation runs for different sampling setting in Example 1

the single stage method, and the proposed method provides comparable results in all the scenarios. Moreover, the naive method performs considerably well in this simulated model. This is potentially because of the smooth function structure of μ allows a reliable results that are not affected much by data issues in a particular partition.

Example 2. In this example, we consider Rastrigin function

$$\mu(z) = 10 + z^2 - 10\cos(2\pi z),$$

from Mühlenbein et al. (1991). As the data model for y_{jk}, we consider two cases: (1) $y_{jk} \sim N(X_k\beta_j, \tau^2)$ with $\tau = 2, 4, 8$, for $k = 1, \ldots, n_j$, and $\beta_j \sim N(\mu(z_j), \sigma^2)$ with $\sigma = 0.5$; (2) $y_{jk} \sim \text{Bern}(p_{jk})$, where $p_{jk} = (1 + \exp(-\eta_{jk}))^{-1}$ and $\eta_{jk} \sim N(X_k\beta_j, \tau^2)$ with $\tau = 2, 4, 8$, for $k = 1, \ldots, n_j$ and $\beta_j \sim N(\mu(z_j), \sigma^2)$ with $\sigma = 0.5$. The design matrix X_k is generated from $N(0, 1)$. We set $J = 50$, $z_1, \ldots, z_J$ equally spaced between -3 and 3. The other settings for the simulations are set as same as Example 1.

Table 2 presents the RMSE averaged over the 200 replicates. To calculate standard deviations (SD), RMSEs are calculated for each simulation run, and SD are calculated from 200 RMSEs.

Model	τ	$N_{\min}$	Single	Naive	Multilevel
LM	2	50	0.54 (0.03)	0.24 (0.02)	0.24 (0.02)
	2	100	0.53 (0.02)	0.24 (0.03)	0.24 (0.03)
	4	50	0.63 (0.03)	0.28 (0.03)	0.28 (0.03)
	4	100	0.60 (0.03)	0.27 (0.03)	0.27 (0.03)
	8	50	0.92 (0.05)	0.39 (0.04)	0.38 (0.04)
	8	100	0.84 (0.04)	0.35 (0.04)	0.35 (0.04)
GLM	2	50	1453.76 (11826.25)	169.93 (1161.39)	7.78 (1.64)
	2	100	233.40 (648.95)	31.09 (84.32)	6.99 (1.30)
	4	50	156.06 (543.66)	22.74 (72.48)	9.69 (0.64)
	4	100	20.98 (92.79)	9.57 (11.66)	8.96 (0.40)
	8	50	16.64 (47.41)	11.90 (5.43)	11.79 (0.16)
	8	100	11.59 (0.06)	11.60 (0.06)	11.76 (0.05)

Table 2: Comparison of RMSE, where Mean and SD are calculated over 100 simulation runs for different sampling setting in Example 2.

Clearly the proposed method performs better compared to the other two methods. A notable difference of the results in this example compared to Example 1 is that naive approach suffers under GLM. Native approach can be susceptible to the poor estimate in some partitions, whereas multilevel model can help the estimates to be more reliable by simultaneously incorporating the neighboring partitions and considering the uncertainties in each partition.

5 Application

In this section, we apply our method to analyze a data set examining the recent history of music concert events. In recent years, the revenue source of the music industry has shifted from record sales to concert market. Between 2001 and 2010, the concert revenues in the United States have increased from 1.7 to 4.25 billions USD (Pollstar, 2011), while record sales have decreased from 14.32 to 6.55 billion USD (RIAA, 2016). Among many potential factors, studies have suggested that the Internet has contributed to this change, due to easier ticket transaction (Eckard and Smith, 2013), or digital piracy and streaming service (Smith and Telang, 2012).

In this context, it is our interest to examine how the impact of the population size on musi-

cians' decision of the concert locations has shifted. As the concert has become the main source of the revenue, smaller market may be exploited more aggressively, which would bring the concert events to the less populated places that traditionally have not been the interests of the industry. To investigate this, we analyze the concert events history from Songkick.com between 2000 and 2012. The collected data have the main headliner of the concerts, event date, and geographic location as city and state. Among the events, we selected the events associated with those musicians who have made the top 100 North American Annual Chart by Pollstar Magazine (Pollstar, 2011) at least once during this period. The data set accounts for 57511 concert events by 234 musicians. Then each concert location is joined with the CBSA that it belongs to, and its population size. CBSA provides a useful regional grouping reference that shares the similar proximity to the concert venue, and the population of each CBSA can provide a useful information about the size of potential attendees that a concert event is aiming to attract. The definition and population for each CBSA is from the US Census data (U.S. Census Bureau, 2010).

Our modeling approach has been applied to examine the change of the impact of population on the concert events. Multilevel approach provides a flexible model to examine this trend. Some more consideration is needed because of the nature of our data collection, because conventional assumption on data generation cannot be justified. First, the sampled musicians are selected based on the chart performance. Second, the events data are observed over time, so there might be a temporal pattern.

To address these issues, we adopt the hurdle model Mullahy (1986) to take into account the bias caused by preferentially choosing the top musicians. Hurdle model naturally accounts for musicians' economic decision, such that an event occurs only when certain conditions are met. Population of the concert location must be at certain level ("hurdle"), otherwise the location is not considered. Multilevel structure with the smooth splines incorporates the repeated nature of the observations over the years.

For each CBSA, population information x_i, $i = 1, \ldots, n$, and the number of concert events for

each CBSA y_{ij}, $j = 1, \ldots, J$, are available, where $n = 961$ and $J = 156$. Under hurdle model, the probability mass function of Y_{ij} is given by

$$P(Y_{ij} = y) = \begin{cases} \pi_{ij} & y = 0 \\ (1 - \pi_{ij}) \frac{\lambda^{y_{ij}}}{(\exp(\lambda_{ij}) - 1) y_{ij}!} & y = 1, \ldots . \end{cases},$$

where $\pi_{ij} = \exp(-\exp(x_i \beta_j))$ and $\lambda_{ij} = \exp(x_i \eta_j)$ to model the effect of population on both hurdle and count model. We further assume that $\beta_j = \mu(z_j) + \xi_j$ with $\xi_j, \sim N(0, \sigma^2)$.

Denoting $\mathbf{y}_j = y_{1j}, \ldots, y_{nj}$, the log likelihood $l(\beta_j, \eta_j; \mathbf{y}_j, x_i)$ is

$$\sum_{i=1}^{n} \mathbb{I}(y_{ij} = 0)(-e^{x_i \beta_j}) + \sum_{i=1}^{n} \mathbb{I}(y_{ij} \neq 0) \log(1 - e^{-e^{x_i \beta_j}}) + \sum_{i=1}^{n} \mathbb{I}(y_{ij} \neq 0) \left[y_{ij} x_i \eta_j - \log(e^{e^{x_i \eta}} - 1) \right]$$

up to constant. Since $l(\beta_j, \eta_j; \mathbf{y}_j, x_i)$ is the sum of the β_j and η_j parts, the Fisher information is a block diagonal matrix, which makes estimating β_j and η_j is straightforward as they can be estimated separately. Here, our interest is on β_j because it provides the information regarding the change of the musicians' decision. Industry shift as well as some business cycle changes can be reflected in the transition in β_j. Since $\hat{\beta}_j$ is obtained as an MLE, it can be assumed that $\hat{\beta}_j \sim N(\beta_j, V_j)$, where $V_j = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1}$ with $\boldsymbol{W}$ is a diagonal matrix whose jth element $w_j = \frac{\exp(\mathbf{x}_i \boldsymbol{\beta}_j)}{1 + \exp(\mathbf{x}_i \boldsymbol{\beta}_j)}$. One can plug in $\hat{\beta}_j$ to w_j to get the values for $\boldsymbol{W}$.

As discussed in Section 2, once $\hat{\beta}_j$ is obtained for each j, the second level model in (3) can be estimated. Let $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_J \right)$ and $\boldsymbol{V} = (V_1, \ldots, V_J)$. Then we can apply EM algorithm to obtain $\hat{\boldsymbol{\mu}} = (\hat{\mu}(z_1), \ldots, \hat{\mu}(z_J))^\top = \boldsymbol{S}_\lambda \hat{\boldsymbol{\beta}}$, and choose λ based on GCV.

In Figure 2, the points are the estimated $\hat{\beta}_j$'s, the blue dashed curve shows the fitted structural model from naive method, and the red solid curve from our proposed method. Both share the overall global cycle which may reflect the overall cycle of concert sector of music industry; relatively high around 2008, while on the lower side around 2004 and 2012. One interesting observation

Figure 2: Proposed, Smooth Splines, and Observed

is that the trend obtained from the proposed method in solid curve is very periodic, in that it coincides with the annual cycle. One can see that the estimated $\mu(z)$ repeatedly becomes higher in winter time, while low in summer. It can be interpreted that the sensitivity to population is low in summer because the artists tend to have more outdoor events as it is easier to attract more people, and people are willing to travel more in the summer.

6 Summary and Discussion

With the internet being the essential part of people's lives, more data are easily accessible online. In analyzing and studying such data often makes the conventional assumption of the data generation difficult to justify. A simple and efficient but flexible modeling approach to analyze such data allows more insight from diverse sources.

In this paper, we have developed a framework using a multilevel smoothing splines approach, which allows a flexible modeling of data collected from internet. This exploits the both information of uncertainty at the data level and functional smoothness at the underlying level. The computation

is simple and efficient, and can be implemented with various modeling specification. Through the application and simulation, it has been demonstrated that our methodology can handle various measurements. It is presented with music concert record data, in which our proposed method is able to unveil an interesting pattern of people's decision.

There are several potential application for the proposed method. Often, it is not convenient to transfer the entire data due to the high volume or confidential issues, such as patient information from healthcare providers, or credit record from the financial institution. In such case, one can aggregate the data to a set of summary statistics. Provided that underlying smoothness assumption is plausible, it can provide a useful alternative for modeling.

We would like to remark on potential future research topics. First, the proposed method is illustrated with a parametric data model. Extension of the method to more flexible model structure, for example, a non-parametric regression type model can be considered. Second, a multilevel method incorporating the spatial structure with the thin plate splines can be developed for spatial data observed over time. A method that can be quickly used for a complex spatial data will have a room for various application in industry.

Appendix

A.1 Proof of Theorem 1

Notice that $\boldsymbol{D} = \text{diag}(d_1, \ldots, d_J)$ with $d_j = \sigma^2/(\sigma^2 + V_j)$. Hence, by (8), our proposed estimator can be rewritten as

$$\hat{\mu} = \arg \min_{m \in \mathcal{W}^{(2,2)}} L(m),$$

where $L(m) = J^{-1} \sum_{j=1}^{J} d_j \left(\hat{\beta}_j - m(z_j)\right)^2 + \lambda^4 \int (m'')^2$ after rescaling the penalty parameter λ. For any functions f, g in $\mathcal{W}^{(2,2)}$, denote the L_2 norm $\|f\| = [\int f^2]^{1/2}$with usual inner product $< f, g >= \int fg$. And define the Sobolev norm $\|f\|_\lambda = [\int f^2 + \lambda^4 \int (f'')^2]^{1/2}$ with associated inner

product $< f, g >_\lambda = \int fg + \lambda^4 \int f''g''$. Then, from Cauchy-Schwartz inequality and the definitions of $\|\cdot\|$ and $\|\cdot\|_\lambda$, one has

$$< f'', g'' > \leq \|f''\|\|g''\| \leq \lambda^{-2}\|f''\|\|g\|_\lambda. \tag{A.1}$$

We can take a universal constant $0 < c < \infty$, not depending on z, f, such that for $z \in \mathcal{I}$ and $f \in \mathcal{W}^{(2,2)}$,

$$|f(z)| \leq c\lambda^{-1/2}\|f\|_\lambda \tag{A.2}$$

as in Eggermont and LaRiccia (2006). And $\mathcal{W}^{(2,2)}$ equipped with $\|\cdot\|_\lambda$ is a reproducing kernel Hilbert space. Refer to Aronszajn (1950) and Wahba (1990) for more details on the reproducing kernel Hilbert space. Then, there exists the reproducing kernel $K_\lambda(\cdot,\cdot)$ with $K_\lambda(z,\cdot) \in \mathcal{W}^{(2,2)}$ for all $z \in \mathcal{I}$ and

$$f(z) = < K_\lambda(z,\cdot), f >_\lambda, \quad f \in \mathcal{W}^{(2,2)} \tag{A.3}$$

from the reproducing property of $K_\lambda(\cdot,\cdot)$. From (A.2), (A.3) and the assumptions (A1)-(A2), one can easily show that $\|K_\lambda(z,\cdot)\|_\lambda \leq c\lambda^{-1/2}$ and

$$E\|J^{-1}\sum_{j=1}^{J} d_j\xi_j K_\lambda(z,\cdot)\|_\lambda^2 = O((J\lambda)^{-1}), \tag{A.4}$$

where c is the universal constant in (A.2).

Let $\Delta = \hat{\mu} - \mu$. Observe that

$$L(\hat{\mu}) \;=\; L(\mu) + \frac{2}{J}\sum_{j=1}^{J} d_j\xi_j\Delta(z_j) + 2\lambda^4 < \mu'', \Delta'' > + \frac{1}{J}\sum_{j=1}^{J} d_j\left(\Delta(z_j)\right)^2 + \lambda^4\|\Delta''\|^2.$$

This together with the optimality of $\hat{\mu}$, that is, $L(\hat{\mu}) \leq L(\mu)$, gives

$$\begin{aligned} \frac{1}{J}\sum_{j=1}^{J} d_j \left(\Delta(z_j)\right)^2 + \lambda^4 \|\Delta''\|^2 &\leq -\frac{2}{J}\sum_{j=1}^{J} d_j \xi_j \Delta(z_j) - 2\lambda^4 < \mu'', \Delta'' > \\ &\leq (O_p((J\lambda)^{-1/2}) + 2\lambda^2 \|\mu''\|)\|\Delta\|_\lambda. \end{aligned} \tag{A.5}$$

The last inequality directly follows from the facts that (A.3), (A.4) and (A.1). On the other side, one has

$$\frac{1}{J}\sum_{j=1}^{J} d_j \left(\Delta(z_j)\right)^2 + \lambda^4 \|\Delta''\|^2 = \|\Delta\|^2(1 + o_p(1)) + \lambda^4 \|\Delta''\|^2 = \|\Delta\|_\lambda^2 (1 + o_p(1))$$

using standard empirical process techniques (e.g, van de Geer, 2000) and the fact that $d_j = 1 + o_p(1)$ uniformly in j. This and (A.5) complete the proof of Theorem 1 considering the rescaling factor for λ. □

References

U.S. Census Bureau, Population Division (U.S. Census Bureau) (March 2010). Annual estimates of the resident population: April 1, 2000 to july 1, 2009. available at http://www.census.gov/popest/data/metro/totals/2009/.

U.S. Census Bureau, (U.S. Census Bureau) (March 2012). 2010 census summary file 1— technical documentation/prepared by the u.s. census bureau, revised 2012. available at https://www.census.gov/prod/cen2010/doc/sf1.pdf.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*, 337–404.

Berry, S. M., R. J. Carroll, and D. Ruppert (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association 97*, 160–169.

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *The Annals of Statistics 17*, 453–555.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles 76*, 1–32.

Chen, H., R. H. Chiang, and V. C. Storey (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly 36*, 1165–1188.

Cho, D., Y. Hwang, and J. Park (2018). More buzz, more vibes: Impact of social media on concert distribution. *Journal of Economic Behavior & Organization 156*, 103–113.

Cox, D. D. (1983). Asymptotics for m-type smoothing splines. *The Annals of Statistics 11*, 530–551.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B39*, 1–38.

Eckard, E. W. and M. A. Smith (2013). The impact of price discrimination on consumer surplus at popular music concerts. *Economics Letters 118*, 222–224.

Eggermont, P., R. Eubank, and V. LaRiccia (2010). Convergence rates for smoothing spline estimators in varying coefficient models. *Journal of Statistical Planning and Inference 140*, 369–381.

Eggermont, P. P. B. and V. N. LaRiccia (2006). Uniform error bounds for smoothing splines. In *High dimensional probability*, pp. 220–237. Institute of Mathematical Statistics.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, Inc.

Gelman, A. (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics 48*, 432–435.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall.

Gu, C. (2013). *Smoothing spline ANOVA models*, Volume 297. Springer Science & Business Media.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hwang, Y., S. Lu, and J.-K. Kim (2018). Bottom-up estimation and top-down prediction: Solar energy prediction combining information from multiple sources. *Annals of Applied Statistics 12*, 2096–2120.

Levine, R. A. and G. Casella (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics 10*, 422–439.

Liu, X., K. Yeo, Y. Hwang, J. Singh, J. Kalagnanam, et al. (2016). A statistical modeling approach for air quality data based on physical dispersion processes and its application to ozone modeling. *The Annals of Applied Statistics 10*, 756–785.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*, 239–245.

Messer, K. and L. Goldstein (1993). A new class of kernels for nonparametric curve estimation. *The Annals of Statistics 21*, 179–195.

Mühlenbein, H., M. Schomisch, and J. Born (1991). The parallel genetic algorithm as function optimizer. *Parallel computing 17*(6-7), 619–632.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics 33*, 341–365.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American statistical Association 83*, 1134–1143.

Owen, A. B. et al. (1997). Scrambled net variance for integrals of smooth functions. *The Annals of Statistics 25*, 1541–1562.

Pollstar (2011). 2010-Year End Business Analysis. www.pollstarpro.com/files/charts2011/2011BusinessAnalysis.pdf. Accessed: 2018-03-26.

Recording Industry Association of America (2016). RIAA Year-End Industry Shipment and Revenue Statistics. https://www.riaa.com/reports/ri. Accessed: 2018-03-26.

Shang, Z. and G. Cheng (2013). Local and global asymptotic inference in smoothing spline models. *The Annals of Statistics 5*, 2608–2638.

Shao, J. (2003). *Mathematical Statistics* (Second ed.). New York: Springer.

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

Smith, M. D. and R. Telang (2012). Assessing the academic literature regarding the impact of media piracy on sales. http://dx.doi.org/10.2139/ssrn.2132153.

Speckman, P. L. and D. Sun (2003). Fully bayesian spline smoothing and intrinsic autoregressive prior. *Biometrika 90*, 289–302.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*, 1040–1053.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological) 40*, 364–372.

Wahba, G. (1983). Bayesian "confidence interval" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological) 45*, 133–150.

Wahba, G. (1990). *Spline models for observational data*, Volume 59. SIAM.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications. Monographs on Statistics and Applied Probability 121.* Chapman and Hall/CRC Press.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.

Wei, G. C. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association 85*, 699–704.