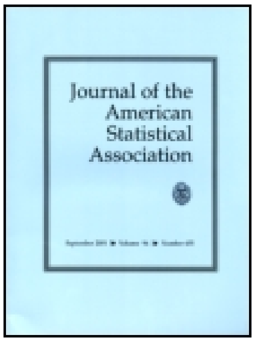

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

# Additive Functional Regression for Densities as Responses

Kyunghee Han, Hans-Georg Müller & Byeong U. Park

**To cite this article:** Kyunghee Han, Hans-Georg Müller & Byeong U. Park (2019): Additive Functional Regression for Densities as Responses, Journal of the American Statistical Association, DOI: 10.1080/01621459.2019.1604365

**To link to this article:** https://doi.org/10.1080/01621459.2019.1604365

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2019.1604365

# Additive Functional Regression for Densities as Responses

Kyunghee Han[a], Hans-Georg Müller[a], and Byeong U. Park[b]

[a]Department of Statistics, University of California, Davis; [b]Department of Statistics, Seoul National University, Seoul, Republic of Korea

**ABSTRACT**

We propose and investigate additive density regression, a novel additive functional regression model for situations where the responses are random distributions that can be viewed as random densities and the predictors are vectors. Data in the form of samples of densities or distributions are increasingly encountered in statistical analysis and there is a need for flexible regression models that accommodate random densities as responses. Such models are of special interest for multivariate continuous predictors, where unrestricted nonparametric regression approaches are subject to the curse of dimensionality. Additive models can be expected to maintain one-dimensional rates of convergence while permitting a substantial degree of flexibility. This motivates the development of additive regression models for situations where multivariate continuous predictors are coupled with densities as responses. To overcome the problem that distributions do not form a vector space, we utilize a class of transformations that map densities to unrestricted square integrable functions and then deploy an additive functional regression model to fit the responses in the unrestricted space, finally transforming back to density space. We implement the proposed additive model with an extended version of smooth backfitting and establish the consistency of this approach, including rates of convergence. The proposed method is illustrated with an application to the distributions of baby names in the United States.

**ARTICLE HISTORY**
Received May 2018
Revised January 2019

**KEYWORDS**
Additive models; Functional data analysis; Random densities; Smooth backfitting.

## 1. Introduction

We consider regression models where $d$-vectors of continuous predictors $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d})$ are coupled with responses that can be viewed as random densities $f_i$ for $i = 1, \ldots, n$ subjects, and one is interested to infer the regression relation $E(f|\mathbf{X}) = g(\mathbf{X})$. This situation arises in many situations, however, there are only very few methods available at this time to deal with such data. For example in neuroimaging, intra-hub connectivity can be quantified as a density of correlations (Petersen and Müller 2016; Petersen, Chen, and Müller 2018) and it is then of interest how this connectivity density changes with the age of the subject. Similarly, predicting distributions is of interest in finance (Sen and Ma 2015). We will present a specific example for a regression problem with vector predictors and distribution response in Section 5, where we study how the proposed additive density regression can be applied to investigate the dependency of the temporal distribution of the popularity of baby names in the United States over calendar years on the initial popularity of a name.

Systematic studies of densities as random objects in regression models have used simplicial models within the Aitchison geometry (Talská et al. 2018), with applications to the modeling of distributions of particle sizes in relation to predictors in soil science (Menafoglio, Guadagnini, and Secchi 2014; Menafoglio, Secchi, and Guadagnini 2016). General metric approaches, including Fréchet regression with the Wasserstein metric were studied recently (Petersen and Müller 2018). Various tools for the analysis of samples of densities as data objects have been proposed over the last decade (Egozcue, Diaz-Barrero, and Pawlowsky-Glahn 2006; Delicado 2007, 2011; Zhang and Müller 2011; Panaretos and Zemel 2016; Petersen and Müller 2016; Bigot et al. 2017), extending the pioneering work of Kneip and Utikal (2001). Many of these methods draw on concepts that were developed in functional data analysis, such as functional principal component analysis (Ramsay and Silverman 2005; Wang, Chiou, and Müller 2016).

Specifically, the data we consider are of the type $(X_i, f_i)$, $i = 1, \ldots, n$, where $X_i$ are predictor vectors and the responses $f_i$ are density functions. While the regression problem to infer $E(f|X = x)$ is relevant for many applications, it has not yet been much investigated, and this provides the motivation to introduce an additive approach in this paper. The problem we consider here stands in contrast to a different and well-studied problem, where one has data $(X_i, Y_i)$, $i = 1, \ldots, n$, with scalars $Y_i$, and is interested in inferring the conditional density of $Y$ given $X$, $f_{Y|X}$, which is derived from the conditional distribution $P(Y \leq y|X = x)$, which is not parametrically specified. This latter problem is usually considered in the general framework of nonparametric regression and has a rich literature (see, e.g., Roussas 1969; Bhattacharya and Gangopadhyay 1990; Koenker, Ng, and Portnoy 1994; Hall, Wolff, and Yao 1999; Hall and Müller 2003; Dunson, Pillai, and Park 2007; Li, Lin, and Racine 2013). This problem has also been sometimes referred to as density regression, but it is best characterized as conditional density estimation within the framework of nonparametric regression.

**CONTACT** Byeong Park ✉ bupark@stats.snu.ac.kr Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Apart from the complication that the responses we consider are density functions, which do not form a linear space, any unrestricted nonparametric approach would be subject to the curse of dimensionality for larger dimensions $d$ of the predictor vectors, especially if $d > 3$. Additive modeling is attractive for such situations as it provides a structured regression approach that avoids the curse of dimensionality, while maintaining a large degree of flexibility of the fitted regression functions. For this reason, our goal in this paper is to develop additive density regression, which is a version of additive modeling that is suitable for density responses. In order to deal with the restrictions that are inherent to density functions, and which are not convenient for additive modeling, we adopt a transformation approach as in Petersen and Müller (2016), whereby density functions are transformed by one-to-one maps to unrestricted square integrable functions.

The models we consider feature a random density $f$ as response, coupled with predictors $\mathbf{X}_i$ that are $d$-vectors. We mainly use the log-quantile transformation

$$\Psi_1 : \mathcal{F} \mapsto \mathrm{L}_2, \quad \Psi_1(f) = \log(F^{-1})' \tag{1.1}$$

for a collection of density or distribution functions $\mathcal{F}$, but other transformations that satisfy certain structural constraints and map a density or distribution into an unrestricted square integrable random function could also be used. Such alternative transformations satisfying the required criteria are discussed in Petersen and Müller (2016), where specifically the log hazard transformation is investigated. This provides an alternative transformation, mapping densities to $\mathrm{L}_2$ by

$$\Psi_2 : \mathcal{F} \mapsto \mathrm{L}_2, \qquad \Psi_2(f) = \log(f/\bar{F}), \quad \bar{F} = 1 - F. \tag{1.2}$$

Given a transformation that satisfies the requirements, the goal is to estimate $\mathrm{E}\left(\Psi(f)|\mathbf{X}\right)$.

A challenge is that the responses $f_i$ in the random copies $(\mathbf{X}_1,f_1), \dots, (\mathbf{X}_n,f_n)$ of $(\mathbf{X},f)$ are not actually observed, but must be estimated from random samples that are generated by these densities, $Y_{i,1}, \dots, Y_{i,N_i} \overset{\text{iid}}{\sim} f_i$. A natural approach is to substitute estimates $\hat{f}_i$ for the unknown response densities $f_i$, obtaining these estimates from these random samples, which we may view as noisy observations of the true responses $f_i$. Petersen and Müller (2016) employed special boundary-corrected kernel estimators $\hat{f}_i$ that converge to $f$ uniformly,

$$\sup_{f_i \in \mathcal{F}} \|\hat{f}_i - f_i\|_\infty \to 0$$

as $N_i \to \infty$; conventional kernel density estimators do not have this property due to boundary effects. When fitting the model $\mathrm{E}(\Psi(f)|\mathbf{X} = \mathbf{x})$, we take the estimated densities $\hat{f}_i$ as the observations of the response $f$. Thus, there are two sources of errors in the estimation that need to be taken care of. The first is the typical error in the response as in the standard regression problem, $\Delta_{i,1} = f_i - \Psi^{-1}(\mathrm{E}(\Psi(f)|\mathbf{X}_i))$. The second is the error resulting from the estimation of the $f_i$, $\Delta_{i,2} = \hat{f}_i - f_i$. We then write

$$\hat{f}_i = \mu(\cdot|\mathbf{X}_i) + \Delta_{i,1} + \Delta_{i,2},$$

where $\mu(\cdot|\mathbf{x}) := \Psi^{-1}(\mathrm{E}(\Psi(f)|\mathbf{X} = \mathbf{x}))$.

The article is organized as follows. We introduce the proposed additive density regression model and corresponding estimates in Section 2 and present asymptotic theory in Section 3. Simulation results are reported in Section 4 and we provide a data illustration on the popularity of baby names in the United States in Section 5. Detailed derivations of the theoretical results and proofs are in the online supplement.

## 2. Model and Estimation

### 2.1. Additive Density Regression Model

Our starting point is the *conditional Fréchet mean* of the density response given the predictor vector, which defines Fréchet regression (Petersen and Müller 2018). Let $\mathcal{F}$ denote the space of density functions defined on a common support, say $\mathcal{Y}$. We take $\mathcal{Y} = [0,1]$ for simplicity. Given a 1:1 transformation $\Psi : \mathcal{F} \to \mathrm{L}_2$, where $\mathrm{L}_2$ is equipped with the usual metric $d(r_1, r_2) = \|r_1 - r_2\|_2 = (\int_0^1 (r_1(u) - r_2(u))^2\, du)^{1/2}$, we use the metric $d_f(f_1, f_2) = d(\Psi(f_1), \Psi(f_2))$ for $f_1, f_2 \in \mathcal{F}$. Then for a given $\mathbf{x} \in \mathbb{R}^d$, the conditional Fréchet mean of the random density $f$ is defined by

$$\mu(\cdot|\mathbf{x}) = \underset{\phi\in\mathcal{F}}{\arg\min}\, \mathrm{E}\left(\|\Psi(f) - \Psi(\phi)\|_2^2 \,|\, \mathbf{X} = \mathbf{x}\right),$$

where the expectation ‘E’ refers to the joint distribution of $(\mathbf{X}, f)$. Thus,

$$\Psi(\mu(\cdot|\mathbf{x}))(u) = \mathrm{E}(\Psi(f)(u) \,|\, \mathbf{X} = \mathbf{x}), \quad u \in [0,1],$$

so that the proposed additive density regression model leads to the following model for the conditional densities at predictor levels $\mathbf{X} = \mathbf{x}$, corresponding to the mean regression,

$$\mu(v|\mathbf{x}) = \Psi^{-1}\left(\mathrm{E}(\Psi(f) \,|\, \mathbf{X} = \mathbf{x})\right)(v), \quad v \in [0,1]. \tag{2.1}$$

We consider the random densities $f_i$ that serve as responses in the proposed additive density regression model to be realizations of $\mathrm{L}_2$-processes on $[0,1]$ that take values in $\mathcal{F}$, with $f$ as a generic element. Let $\varepsilon$ denote an error process that is an element of $\mathrm{L}_2$, with iid realizations $\varepsilon_i$ that are independent from all other random elements in the regression model

$$\Psi(f_i) = g_0 + \sum_{j=1}^{d} g_j(\cdot, X_{i,j}) + \varepsilon_i, \tag{2.2}$$

where $g_0$ is an unknown univariate satisfying $g_0(u) = \mathrm{E}\Psi(f)(u)$ and $g_j$, for $1 \le j \le d$, are unknown bivariate component functions. We assume that $(\mathbf{X}_i, \varepsilon_i)$ are independent copies of $(\mathbf{X}, \varepsilon)$ satisfying $E(\varepsilon_i|\mathbf{X}_i) = 0$. Then

$$\mathrm{E}\left(\Psi(f)(u)|\mathbf{X}\right) = g_0(u) + \sum_{j=1}^{d} g_j(u, X_j), \quad u \in [0,1]. \tag{2.3}$$

Without loss of generality, assume that the bivariate functions $g_j$ satisfy

$$\int_0^1 g_j(u, x_j) p_j(x_j)\, dx_j = 0, \quad 1 \le j \le d, \tag{2.4}$$

and $g_0$ is determined according to the constraints (2.4); note that this is a standard approach in additive modeling. We assume that each predictor $X_j$ is supported on a compact set $I_j$, and without loss of generality take $I_j = [0,1]$ for all $1 \le j \le d$.

## 2.2. Estimation

Based on the representation in (2.3), we estimate the component functions $g_j$ for $0 \le j \le d$. Given estimates $\hat{g}_j$ for the $g_j$, our estimator of the conditional Fréchet mean $\mu(\cdot|\mathbf{x})$ is

$$\hat{\mu}(v|\mathbf{x}) = \Psi^{-1}\left(\hat{g}_0 + \sum_{j=1}^{d} \hat{g}_j(\cdot, x_j)\right)(v), \tag{2.5}$$

which is a density function with argument $v$. We employ a kernel smoothing technique to estimate the unknown functions $g_j$ in model (2.3). For a symmetric probability density defined on $[-1, 1]$ that serves as kernel function $K$, define a normalized kernel

$$K_h(x, z) = c_h(z)\frac{1}{h}K\left(\frac{x-z}{h}\right), \tag{2.6}$$

where $c_h(z)$ is defined by $c_h(z)^{-1} = \int_0^1 h^{-1}K(h^{-1}(x-z))\,dx$ so that $\int_0^1 K_h(x, z)\,dx = 1$ for all $z \in [0, 1]$.

The latter normalization property is important for the theoretical development of our kernel method for the additive model. For example, if we estimate the joint density $p$ of $\mathbf{X}$ by

$$\hat{p}(\mathbf{x}) = n^{-1}\sum_{i=1}^{n} K_{h_1}(x_1, X_{i,1}) \times \cdots \times K_{h_d}(x_d, X_{i,d}),$$

all marginal densities $p_S$ of $\mathbf{X}_S \equiv (X_j : j \in S)$, for any index set $S \subset \{1, 2, \dots, d\}$, are obtained by simply integrating $\hat{p}$ over the variables with indices not in $S$. The typical kernel estimators of the marginal densities $p_j$ of $X_j$ and $p_{jk}$ of $(X_j, X_k)$ are obtained by

$$\begin{aligned}\hat{p}_j(x_j) &:= n^{-1}\sum_{i=1}^{n} K_{h_j}(x_j, X_{i,j}) = \int_{[0,1]^{d-1}} \hat{p}(\mathbf{x})\,d\mathbf{x}_{-j},\\ \hat{p}_{jk}(x_j, x_k) &:= n^{-1}\sum_{i=1}^{n} K_{h_j}(x_j, X_{i,j})K_{h_k}(x_k, X_{i,k})\\ &= \int_{[0,1]^{d-2}} \hat{p}(\mathbf{x})\,d\mathbf{x}_{-j,k},\end{aligned} \tag{2.7}$$

where here and below $\mathbf{x}_{-j} = (x_l : 1 \le l \le d, l \ne j)$ and $\mathbf{x}_{-j,k} = (x_l : 1 \le l \le d, l \ne j, k)$ for a $d$-vector $\mathbf{x}$.

We now consider the realistic case with regard to implementation of our method, where one observes random copies $\mathbf{X}_i$ of $\mathbf{X}$, but does not directly observe the associated response densities $f_i$. Instead, the information available about each $f_i$ is the sample $Y_{i,1}, \dots, Y_{i,N_i}$, which is iid generated from the unknown density $f_i$, coupled with $\mathbf{X}_i$. Our method of estimating $g_j$ is then based on the observations $(\mathbf{X}_i, Y_{i,1}, \dots, Y_{i,N_i})$, $1 \le i \le n$. With $\mathbf{h} = (h_1, \dots, h_d)$ we write $K_{\mathbf{h}}(\mathbf{x}, \mathbf{z}) = K_{h_1}(x_1, z_1) \times \cdots \times K_{h_d}(x_d, z_d)$.

If the density responses $f_i$ were available, we would minimize

$$\int_{[0,1]^{d+1}} n^{-1}\sum_{i=1}^{n}\left(\Psi(f_i)(u) - \eta_0(u) - \sum_{j=1}^{d}\eta_j(u, x_j)\right)^2 \times K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}_i)\,d\mathbf{x}\,du \tag{2.8}$$

over the space of tuples of square integrable functions $\eta_j, 0 \le j \le d$. The objective functional at (2.8) can indeed be interpreted as an estimator of

$$\begin{aligned}&\mathrm{E}\int_0^1 \left(\Psi(f)(u) - \eta_0(u) - \sum_{j=1}^{d}\eta_j(u, X_j)\right)^2 du\\ &= \int_{[0,1]^{d+1}} \mathrm{E}\left[\left(\Psi(f)(u) - \eta_0(u) - \sum_{j=1}^{d}\eta_j(u, x_j)\right)^2 \Big| \mathbf{X} = \mathbf{x}\right]\\ &\times p(\mathbf{x})\,d\mathbf{x}\,du.\end{aligned}$$

Since the density responses $f_i$ are not observed, we replace them with their estimators $\hat{f}_i$ in (2.8). Ideally, estimates $\hat{f}_i$ are obtained with a modified kernel estimator (Petersen and Müller 2016), where for a symmetric probability density that serves as kernel $\kappa$ and is supported on $[-1, 1]$ and a bandwidth $b$ we write $\kappa_b(u) = b^{-1}\kappa(b^{-1}u)$ and take

$$\hat{f}_i(v) = \frac{N_i^{-1}\sum_{\ell=1}^{N_i} w_b(v)\kappa_b(v - Y_{i,\ell})}{\int_0^1 N_i^{-1}\sum_{\ell=1}^{N_i} w_b(v)\kappa_b(v - Y_{i,\ell})\,dv}, \tag{2.9}$$

where $w_b(v)$ is defined by $w_b(v)^{-1} = \int_0^1 \kappa_b(v - y)\,dy$, and $\kappa, b$ are possibly different from $K$ and $h_j$, respectively.

Note that $\int_0^1 \hat{f}_i(v)\,dv = 1$ and that the normalization factor $w_b$ serves to remove boundary bias. Petersen and Müller (2016) showed that $\hat{f}_i$ as defined at (2.9) is uniformly consistent. For a related discussion in the case of locally linear smooth backfitting estimation, see Mammen and Park (2006)

The minimization of the resulting objective functional with $\hat{f}_i$ inserted in place of $f_i$ in (2.8) is equivalent to minimizing

$$\int_{[0,1]^d} n^{-1}\sum_{i=1}^{n}\left(\Psi(\hat{f}_i)(u) - \eta_0(u) - \sum_{j=1}^{d}\eta_j(u, x_j)\right)^2 K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}_i)\,d\mathbf{x} \tag{2.10}$$

pointwise for each $u \in [0, 1]$. The resulting pointwise estimators of $g_j$ in practice might not be smooth, depending on the smoothness of density response estimators $\hat{f}_i$. This may be undesirable in case it is believed that the true component functions $g_j$ are smooth and can be easily remedied by inserting an additional local smoothing step in the direction of $u$. Namely, for each $u \in [0, 1]$, minimize

$$\int_0^1 \int_{[0,1]^d} n^{-1}\sum_{i=1}^{n}\left[\Psi(\hat{f}_i)(v) - \eta_0(u) - \sum_{j=1}^{d}\eta_j(u, x_j)\right]^2 \times K_{\mathbf{h}}(\mathbf{x}, \mathbf{X}_i)\,d\mathbf{x} \cdot \omega_{h_0}(u - v)\,dv \tag{2.11}$$

with respect to $\eta_0(u)$ and $(\eta_1(u, \cdot), \dots, \eta_d(u, \cdot))$, where $\omega$ and $h_0$ are a kernel and a bandwidth that may differ from $K$ and $h_j$, respectively, and $\omega_{h_0}(u) = h_0^{-1}\omega(h_0^{-1}u)$. In the minimization we impose the constraints

$$\int_0^1 \hat{g}_j(u, x_j)\hat{p}_j(x_j)\,dx_j = 0, \quad 1 \le j \le d, \tag{2.12}$$

where (2.12) is an empirical version of $\mathrm{E}g_j(u, X_j) = 0$ as at (2.4). Our theory to be presented in the next section also includes the special case (2.10) of (2.11).

Define $\lambda_0(u) = \int_0^1 \omega_{h_0}(u-v)\,dv$, and let

$$\begin{aligned}\hat{g}_0(u) &= \lambda_0(u)^{-1} n^{-1} \sum_{i=1}^{n} \int_0^1 \omega_{h_0}(u-v)\Psi(\hat{f}_i)(v)\,dv,\\ \tilde{g}_j(u,x_j) &= \lambda_0(u)^{-1}\hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^{n} K_{h_j}(x_j, X_{i,j})\\ &\quad \times \int_0^1 \omega_{h_0}(u-v)\Psi(\hat{f}_i)(v)\,dv.\end{aligned} \tag{2.13}$$

By considering the Frèchet differentials of the objective functional at (2.11), we find that the minimizer of (2.11) satisfies the following system of integral equations:

$$\begin{aligned}&\hat{g}_j(u,x_j)\\ &\quad = \tilde{g}_j(u,x_j) - \hat{g}_0(u)\\ &\qquad - \sum_{k\neq j}^{d} \int_0^1 \hat{g}_k(u,x_k)\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k, \quad 1 \le j \le d.\end{aligned} \tag{2.14}$$

The system of the equations at (2.14) may be understood to solve for the tuples of the univariate $\hat{g}_0$ and bivariate $g_j$ with the constraints (2.12), although it is derived from the pointwise minimization of (2.11) for each $u \in [0,1]$. This is a consequence of the fact that the kernels $\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}$ of the integral operators involved in (2.14) are independent of $u \in [0,1]$. Specifically, the integral operators $\pi_j$ are given by

$$\pi_j(\eta)(u,x_j) = \int_0^1 \eta(u,\mathbf{x})\frac{\hat{p}(\mathbf{x})}{\hat{p}_j(x_j)}\,d\mathbf{x}_{-j}. \tag{2.15}$$

With the convention that $\hat{g}_0(u,\mathbf{x}) = \hat{g}_0(u)$ and $\hat{g}_j(u,\mathbf{x}) = \hat{g}_j(u,x_j)$ for $1 \le j \le d$, we may write (2.14) as

$$\hat{g}_j = \tilde{g}_j - \hat{g}_0 - \sum_{k\neq j} \pi_j(\hat{g}_k), \quad 1 \le j \le d.$$

This is in contrast to the backfitting equations studied in Zhang, Park, and Wang (2013), where the corresponding integral operators depend on $u$.

In practice, we obtain the solution of (2.14) by an iterative algorithm. Let $(\hat{g}_j^{[0]} : 1 \le j \le d)$ be an initial tuple of estimators satisfying the constraints (2.12). The updated solution in the $r$th cycle of the backfitting iteration is given by

$$\begin{aligned}&\hat{g}_j^{[r]}(u,x_j)\\ &\quad = \tilde{g}_j(u,x_j)\\ &\qquad - \hat{g}_0(u) - \sum_{k=1}^{j-1} \int_0^1 \hat{g}_k^{[r]}(u,x_k)\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k\\ &\qquad - \sum_{k=j+1}^{d} \int_0^1 \hat{g}_k^{[r-1]}(u,x_k)\frac{\hat{p}_{jk}(x_j,x_k)}{\hat{p}_j(x_j)}\,dx_k, \quad 1 \le j \le d.\end{aligned} \tag{2.16}$$

Since $\int_0^1 \tilde{g}_j(u,x_j)\hat{p}_j(x_j)\,dx_j = \hat{g}_0(u)$ for all $1 \le j \le d$ and the initial estimators $\hat{g}_j^{[0]}$ satisfy the constraints (2.12), it holds that all subsequent updates $\hat{g}_j^{[r]}$ also satisfy (2.12). We stop the backfitting iteration if the difference between two successive updates is less than a threshold value. Our theory in the next section reveals that the iterative algorithm converges in the sup-$L_2$ metric $d(\cdot,\cdot)$ defined by

$$\begin{aligned}d(\hat{\mathbf{g}}^{[r+1]}, \hat{\mathbf{g}}^{[r]})^2 = \sup_{u\in[0,1]} \max_{1\le j\le d} \int_0^1 \left(\hat{g}_j^{[r+1]}(u,x_j) - \hat{g}_j^{[r]}(u,x_j)\right)^2\\ \times p_j(x_j)\,dx_j.\end{aligned}$$

The uniform convergence of the iteration for $u$ is a consequence of the fact that the kernels of the integral operators defined at (2.15) are independent of $u \in [0,1]$.

## 3. Theoretical Results

The following asymptotic results are derived within the framework of the additive density regression model (2.2). As mentioned before, the random densities $f_i$ that serve as responses are viewed as realizations of $L_2$-processes on $[0,1]$, taking values in $\mathcal{F}$, with $f$ as a generic element, and the $(\mathbf{X}_i, \varepsilon_i)$ are independent copies of $(\mathbf{X}, \varepsilon)$ satisfying $E(\varepsilon_i|\mathbf{X}_i) = 0$. We need several assumptions. Let $\mathcal{G}(r)$ denote the space of differentiable density functions $q$ with $\|q'\|_\infty < \infty$ such that $\max\left(\|q\|_\infty, \|1/q\|_\infty\right) \le r$.

(N) The number of observations $N_i$ from which the $i$th density is estimated is such that $(N_i : 1 \le i \le n)$ is independent of $((\mathbf{X}_i, f_i) : 1 \le i \le n)$ and $(Y_{i,\ell} : 1 \le i \le n, \ell \ge 1)$, with

$$P(m(n) \le N_i \le M(n) \text{ for all } 1 \le i \le n) \to 1 \tag{3.1}$$

for some sequences $m \equiv m(n) < M \equiv M(n)$ that converge to $\infty$ as $n \to \infty$.

(T) The random densities $f_i$ and their derivatives satisfy

$$\begin{aligned}&\max_{1\le i\le n} \|f_i\|_\infty \le r_0, \quad \max_{1\le i\le n} \|1/f_i\|_\infty \le r_0,\\ &\max_{1\le i\le n} \|f_i'\|_\infty \le r_0, \quad \max_{1\le i\le n} \|\Psi(f_i)\|_\infty \le r_0\end{aligned}$$

for some constant $0 < r_0 < \infty$. For each $0 < r < \infty$, there exists a constant $0 < D(r) < \infty$ that depends solely on $r$ such that, for all $u_1, u_2 \in \mathcal{G}(r)$,

$$\|\Psi(u_1) - \Psi(u_2)\|_\infty \le D(r)\|u_1 - u_2\|_\infty.$$

(K) The kernels $K$ and $\omega$ are symmetric, bounded, nonnegative, supported on a compact set, say $[-1,1]$, have bounded derivatives and satisfy $\int K = \int \omega = 1$.

(B) The bandwidths $h_j$, $1 \le j \le d$, $h_0$ and $b$ satisfy $h_j \asymp n^{-1/5}$, $h_0 \asymp n^{-\alpha_0}$ for some $\alpha_0 > 0$, $b \to 0$, $mb/\log M \to \infty$ with $m \asymp n^{\alpha_1}$ for some $0 < \alpha_1 < \infty$.

(J) The two-dimensional joint densities $p_{jk}(x_j,x_k)$, $1 \le j \neq k \le d$, are bounded away from zero and infinity on $[0,1]^2$ and are partially continuously differentiable.

(C) The component functions $g_j(u,x_j)$ are twice partially continuously differentiable, $1 \le j \le d$.

(M) The error process $\varepsilon$ satisfies $\sup_{u,x_j\in[0,1]} E(\varepsilon(u)^2|X_j = x_j) < \infty$ for $1 \le j \le d$, and $E\|\varepsilon\|_\infty^k < \infty$ for some $k > 5/2$.

Under Assumption (N), we may treat $N_i$ as bounded below by $m$ and bounded above by $M$ in the following, while (T) is needed for the transformation approach (Petersen and Müller 2016). Assumptions (K) and (C) are standard in kernel smoothing. The bandwidth sequences $h_j$ in Assumption (B) are postulated to have rates that are known to be optimal for univariate kernel smoothing. We do not fix the sizes of other bandwidths $h_0$ and $b$ and only require the boundedness for the two-dimensional densities $p_{jk}$, rather than that of the $d$-dimensional joint density $p$ of $\mathbf{X}$, which does not need to be fully supported on $[0,1]^d$. This leads to enhanced generality in the applicability of the proposed method. Because of the additional local smoothing step across $u$ in the estimation of $g_j$, we also do not need $\varepsilon(u)$ to be smooth as a function, and it is sufficient to require the two conditions on the (conditional) moments in (M).

Our first main result is on the uniqueness of the solution of the system of integral equations (2.14) and the uniform convergence of the iterative algorithm (2.16).

*Theorem 1.* Assume (N), (K), (B), (J), and (C). Then, with probability tending to one, there exists a unique solution $(\hat g_j(u,\cdot) : 1 \le j \le d)$ of (2.14), for each $u \in [0,1]$, subject to the constraints (2.12). Furthermore, there exist absolute constants $\gamma \in (0,1)$ and $C > 0$ such that, with probability tending to one, it holds that

$$\int_0^1 \left(\hat g_j^{[r]}(u,x_j) - \hat g_j(u,x_j)\right)^2 p_j(x_j)\,dx_j \le C\cdot\gamma^r\cdot\left(1+\sum_{k=1}^d \int_0^1 \hat g_k^{[0]}(u,x_k)^2 p_k(x_k)\,dx_k\right) \qquad (3.2)$$

for all $u \in [0,1]$.

We note that this result implies that the convergence of the algorithm (2.16) is uniform for $u \in [0,1]$ if the initial estimators $\hat g_j^{[0]}$ are bounded. This is mainly due to the fact that the kernel of the integral operators $\pi_j$ defined at (2.15) are independent of $u$. The uniqueness of the solution of (2.14) and the convergence of the algorithm are consequences of the fact that the operator $T = (I-\pi_d)\circ\cdots\circ(I-\pi_1)$ is a contraction with probability tending to one. A crucial element of the proof of $\mathbb{P}(\|T\|_{\mathrm{op}} \le \gamma) \to 1$ as $n \to \infty$ for some constant $0 < \gamma < 1$ is the bound

$$\int_{[0,1]^2} \left(\frac{\hat p_{jk}(x_j,x_k)}{\hat p_j(x_j)\hat p_k(x_k)} - \frac{p_{jk}(x_j,x_k)}{p_j(x_j)p_k(x_k)}\right)^2 p_j(x_j)p_k(x_k)\,dx_j\,dx_k = o_p(1);$$

see Mammen, Linton, and Nielsen (1999), Yu, Park, and Mammen (2008), Lee, Mammen, and Park (2010), and Lee, Mammen, and Park (2012) for technical details.

Our second result provides stochastic expansions of $\hat g_j$. One might think that one can ignore $\hat g_0$ as defined at (2.13) in the theoretical developments since it involves one-dimensional smoothing, while the other components $\hat g_j$ for $1 \le j \le d$ involve two-dimensional smoothing, so that $\hat g_0$ has a faster rate of convergence than $\hat g_j$. However, we find that this is not the case. As we will see below and in the technical details in the appendix, the main error of $\hat g_0$ as an estimator of $g_0 = \mathrm{E}\Psi(f)$ is rooted in the error of $\hat f_i$ and thus is not always smaller than the error of $\hat g_j$, where $\hat f_i$ are the kernel estimators of the random densities $f_i$. In the following Theorem 2, we include the error $\hat g_0 - g_0$ to demonstrate how the estimation errors $\Psi(\hat f_i) - \Psi(f_i)$ affect the estimation of $g_j$.

To simplify notations, we use in the following:

$$g_j^{(r,s)}(u,x_j) = \frac{\partial^{r+s}}{\partial u^r \partial x_j^s} g_j(u,x_j), \qquad r+s = 1,2$$

and denote the partial derivatives of $p_{jk}(x_j,x_k)$ by $p_{jk}^{(r,s)}(x_j,x_k)$ for $(r,s) = (0,1)$ or $(1,0)$. Define

$$\lambda_j(u) \equiv \lambda_{n,j}(u) = \int_0^1 h_0^{-j}(v-u)^j \omega_{h_0}(u-v)\,dv, \quad j \ge 0,$$

where $\omega_{h_0}(t) = h_0^{-1}\omega(h_0^{-1}t)$, and

$$\tilde g_j^{A,1}(u,x_j) = \hat p_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{i,j}) \times \int_0^1 \lambda_0(u)^{-1}\omega_{h_0}(u-v)\varepsilon_i(v)\,dv. \qquad (3.3)$$

Writing

$$\delta_{ni}(u) = \lambda_0(u)^{-1}\int_0^1 \omega_{h_0}(u-v)\big(\Psi(\hat f_i)(v) - \Psi(f_i)(v)\big)\,dv, \qquad (3.4)$$

$$a_n = b + m^{-1/2}b^{-1/2}\sqrt{\log M}, \qquad (3.5)$$

$$\mu_{\ell,j}(x_j) = \int_0^1 h_j^{-\ell}(z-x_j)^\ell K_{h_j}(x_j,z)\,dz, \quad \ell \ge 0, \qquad (3.6)$$

we note that the $\mu_{\ell,j}(x_j)$ equal the complete $\ell$th moments of $K$, $\mu_\ell = \int v^\ell K(v)\,dv$, when $x_j \in [2h_j, 1-2h_j]$, a fact that will be repeatedly used in the technical derivations.

For an arbitrary constant $0 < C < \infty$ and a positive sequence $a_n$, define

$$\tilde\Delta_j^A(u,x_j,C) = \mu_{0,j}(x_j)^{-1}p_j(x_j)^{-1} \times E\Big[K_{h_j}(x_j,X_j)\delta_n(u)I(\|\delta_n\|_\infty \le Ca_n)\Big] - E\left[\delta_n(u)I(\|\delta_n\|_\infty \le Ca_n)\right]$$

and observe that the $\tilde\Delta_j^A(u,x_j,C)$ satisfy

$$\sup_{u,x_j\in[0,1]} |\tilde\Delta_j^A(u,x_j,C)| = O(a_n),$$
$$\int_0^1 \tilde\Delta_j^A(u,x_j,C)\mu_{0,j}(x_j)p_j(x_j)\,dx_j \equiv 0 \qquad (3.7)$$

for all $C$, where we use the fact that $\int_0^1 K_{h_j}(x_j,v)\,dx_j = 1$ for all $v \in [0,1]$.

Define

$$\tilde\Delta_j^B(u,x_j) = \mu_2\sum_{k\neq j} h_k^2 E\left(g_k^{(0,1)}(u,X_k)\cdot\frac{p_{jk}^{(0,1)}(X_j,X_k)}{p_{jk}(X_j,X_k)}\,\middle|\, X_j = x_j\right) + h_j^2\cdot\left[\frac{\mu_{2,j}(x_j)}{\mu_{0,j}(x_j)} - \left(\frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)}\right)^2\right]\cdot g_j^{(0,1)}(u,x_j) \cdot\frac{p_j'(x_j)}{p_j(x_j)}. \qquad (3.8)$$

We note that $\tilde{\Delta}_j^A$ and $\tilde{\Delta}_j^B$ are nonstochastic, and writing $h_+^2 = \sum_{j=1}^d h_j^2$, that the $\tilde{\Delta}_j^B(u, x_j)$ are of magnitude $h_+^2$, uniformly for $u, x_j \in [0, 1]$. Let $(\Delta_j^A(\cdot, C) : 1 \le j \le d)$ be the solution of the system of equations

$$\Delta_j^A(u, x_j, C) = \tilde{\Delta}_j^A(u, x_j, C) - \sum_{k \ne j} \int_0^1 \Delta_k^A(u, x_k, C) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\, dx_k, \quad 1 \le j \le d, \tag{3.9}$$

and likewise define $(\Delta_j^B : 1 \le j \le d)$ with $\tilde{\Delta}_j^B$ replacing $\tilde{\Delta}_j^A(\cdot, C)$ in the system of equations (3.9). The two systems of equations determine only the sum functions $\sum_{j=1}^d \Delta_j^A(\cdot, C)$ and $\sum_{j=1}^d \Delta_j^B$, and we invoke the additional constraints

$$\int_0^1 \Delta_j^A(u, x_j, C) p_j(x_j)\, dx_j = 0, \quad 1 \le j \le d,$$
$$\int_0^1 \Delta_j^B(u, x_j) p_j(x_j)\, dx_j = \mu_2 h_j^2 \int_0^1 g^{(0,1)}(u, x_j) p_j'(x_j)\, dx_j, \quad 1 \le j \le d. \tag{3.10}$$

Finally, let

$$\gamma_j(u, x_j) = h_0 \cdot \frac{\lambda_1(u)}{\lambda_0(u)} \cdot g_j^{(1,0)}(u, x_j) + h_j \cdot \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} \cdot g_j^{(0,1)}(u, x_j) + \frac{1}{2} h_0^2 \cdot \frac{\lambda_2(u)}{\lambda_0(u)} \cdot g_j^{(2,0)}(u, x_j) + h_0 h_j \cdot \frac{\lambda_1(u)}{\lambda_0(u)} \cdot \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} \cdot g_j^{(1,1)}(u, x_j) + \frac{1}{2} h_j^2 \cdot \frac{\mu_{2,j}(x_j)}{\mu_{0,j}(x_j)} \cdot g_j^{(0,2)}(u, x_j).$$

*Theorem 2.* Assume (N), (T), (K), (B), (J), (C), and (M). Then, there exists a constant $0 < C_0 < \infty$ such that for all $C \ge C_0$ it holds that

$$\hat{g}_j(u, x_j) = g_j(u, x_j) + \tilde{g}_j^{A,1}(u, x_j) + \Delta_j^A(u, x_j, C) + \Delta_j^B(u, x_j) + \gamma_j(u, x_j) + r_{nj}(u, x_j) + o_p(h_0^2) + O_p\left(n^{-1/2}\sqrt{\log n}\right),$$

uniformly for $u, x_j \in [0, 1]$, where $r_{nj}(u, x_j)$ satisfy

$$\sup_{u, x_j \in [0,1]} |r_{nj}(u, x_j)| = O_p(a_n + h_+^2),$$
$$\sup_{u \in [0,1], x_j \in [2h_j, 1-2h_j]} |r_{nj}(u, x_j)| = o_p(a_n + h_+^2).$$

Here, $h_+$ is defined after (3.8). The lower bound $m$ on the number of observations per density affects the rates of the remainder terms in Theorem 2, as revealed by the definition of $a_n$, in Equation (3.5). In addition to $m$, $a_n$ also depends on the bandwidth $b$ that is used for constructing the kernel density estimators and reflects a trade-off between presmoothing and smooth backfitting. Theorem 2 further demonstrates that $\Delta_j^A(\cdot, C)$ are the only leading effects of the errors $\Psi(\hat{f}_i) - \Psi(f_i)$ that affect the estimation of $g_j$. These effects do not appear in any other leading terms, and neither in the constraints on $\Delta_j^B$ (3.10). Note that changing $\Delta_j^A(u, x_j, C)$ to $\Delta_j^A(u, x_j, C')$ for some $C' \ne C$ does not alter the stochastic expansion as long as $C, C' \ge C_0$ since the difference in the corresponding expansions is of order $o_p(a_n)$, as emerges from the proof of Theorem 2.

We now present the asymptotic distributions of the component function estimators $\hat{g}_j(u, x_j)$. For this, we let $n^{1/5} h_j \to c_j$ for some constants $0 < c_j < \infty$, $1 \le j \le d$, while allowing the magnitude of the bandwidth $h_0$ to be smaller than or equal to that of $h_j$, setting $n^{1/5} h_0 \to c_0$ for some constant $0 \le c_0 < \infty$ including $c_0 = 0$. We consider points $u$ and $x_j$ in $(0, 1)$. For such interior points, there is some simplification in the bias expansion since $u$ and $x_j$, respectively, belong to $[h_0, 1 - h_0]$ and $[2h_j, 1 - 2h_j]$ eventually as $n$ tends to infinity. Specifically, those terms in $\gamma_j$ that involve $\lambda_1(u)$ and $\mu_{1,j}(x_j)$ vanish and we can replace $\lambda_0(u)$ and $\lambda_2(u)$ by the corresponding complete moments, 1 and $\lambda_2 \equiv \int u^2 \omega(u)\, du$, respectively. Likewise, we may replace $\mu_{0,j}(x_j)$ and $\mu_{j,2}(x_j)$ by 1 and $\mu_2$, respectively.

To state the next main result, we choose bandwidths $b$ so that $n^{2/5} a_n \to 0$ as $n \to \infty$, whence the $\Delta_j^A$ are negligible. With

$$\tilde{\beta}_j(u, x_j) = \mu_2 \sum_{k \ne j} c_k^2 E\left(g_k^{(0,1)}(u, X_k) \cdot \frac{p_{jk}^{(0,1)}(X_j, X_k)}{p_{jk}(X_j, X_k)} \,\middle|\, X_j = x_j\right) + \mu_2 \cdot c_j^2 \cdot g_j^{(0,1)}(u, x_j) \cdot \frac{p_j'(x_j)}{p_j(x_j)},$$

we can write $\Delta_j^B = n^{-2/5} \beta_j$, where $(\beta_j : 1 \le j \le d)$ is the solution of the system of equations

$$\beta_j(u, x_j) = \tilde{\beta}_j(u, x_j) - \sum_{k \ne j} \int_0^1 \beta_k(u, x_k) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)}\, dx_k, \quad 1 \le j \le d,$$

subject to the constraints

$$\int_0^1 \beta_j(u, x_j) p_j(x_j)\, dx_j = \mu_2 c_j^2 \int_0^1 g^{(0,1)}(u, x_j) p_j'(x_j)\, dx_j, \quad 1 \le j \le d.$$

The asymptotic variances of $\hat{g}_j(u, x_j)$ are obtained from the stochastic term $\tilde{g}_j^{A,1}(u, x_j)$. The asymptotic covariances of $\tilde{g}_j^{A,1}(u, x_j)$ and $\tilde{g}_k^{A,1}(u, x_k)$ for $j \ne k$ are of smaller order than $n^{-2/5}$ from the standard theory of kernel smoothing. This means that $\hat{g}_j(u, x_j)$ and $\hat{g}_k(u, x_k)$ are asymptotically independent since they are jointly asymptotically normal. To obtain the leading variance terms requires an additional condition on the error process, where in addition to (M) we assume

(M′) For each $1 \le j \le d$, the conditional covariance $E(\varepsilon(u)\varepsilon(v) | X_j = x_j)$, as a function of $(u, v, x_j)$ is continuous on $\{(u, v) \in [0, 1]^2 : v = u\} \times [0, 1]$.

*Theorem 3.* Let $u$ and $x_j$ be fixed points in $(0, 1)$, $1 \le j \le d$. Assume that $a_n$ converges to zero faster than $n^{-2/5}$. Then, under the assumptions of Theorem 2 and (M′), it holds that

$n^{2/5}(\hat{g}_j(u,x_j) - g_j(u,x_j))$ for different $j$ are asymptotically independent, and that

$$\begin{aligned} &n^{2/5}(\hat{g}_j(u,x_j) - g_j(u,x_j)) \\ &\quad \xrightarrow{d} N\Big(c_0^2\lambda_2 g_j^{(2,0)}(u,x_j)/2 + c_j^2\mu_2 g_j^{(0,2)}(u,x_j)/2 + \beta_j(u,x_j), \\ &\quad p_j(x_j)^{-1}E\big(\varepsilon(u)^2|X_j = x_j\big)\int K^2\Big). \end{aligned}$$

## 4. Numerical Illustrations

### 4.1. Simulation study

We demonstrate numerical applications of the proposed additive density regression model (2.2) for density responses when choosing the transformation $\Psi$ as the log-quantile transformation $\Psi_1$ (1.1) and the additive surface in (2.3) as $g(u,\mathbf{x}) = g_1(u,x_1) + g_2(u,x_2)$, with

$$\begin{aligned} g_1(u,x_1) &= \sin(2\pi u)(2x_1 - 1) \quad \text{and} \\ g_2(u,x_2) &= \sin(2\pi u)\sin(2\pi x_2) \end{aligned} \tag{4.1}$$

for $u, x_1, x_2 \in [0,1]$ and $g_0 = 0$. For a given covariate vector $\mathbf{X} = \mathbf{x} \in \mathbb{R}^2$, the fitted model is defined by the conditional Fréchet mean $\mu(v|\mathbf{x}) = \Psi_1^{-1}(g(\cdot,\mathbf{x}))(v)$ of the random response densities $f$, as described in Section 2.1 and model (2.1). More specifically, $\Psi_1^{-1}(g(\cdot,\mathbf{x})) = \theta_0(\mathbf{x})\exp\{-g(F(\cdot|\mathbf{x}),\mathbf{x})\}$, where $\theta_0(\mathbf{x}) = \int_0^1 \exp\{g(v,\mathbf{x})\}\,dv$ and the conditional distribution function $F(\cdot|\mathbf{x})$ and conditional quantile function $Q(\cdot|\mathbf{x})$ satisfy

$$Q(u|\mathbf{x}) = F^{-1}(u|\mathbf{x}) = \theta_0(\mathbf{x})^{-1}\int_0^u \exp\{g(v,\mathbf{x})\}\,dv. \tag{4.2}$$

Here, $\theta_0(\mathbf{x})$ is a normalization factor such that $\mu(\cdot|\mathbf{x})$ is supported on the interval $[0,1]$ for all $\mathbf{x} \in [0,1]^2$ (Petersen and Müller 2016).

Using (4.2), we implemented random sampling based on error-contaminated random quantile functions $\mathcal{Q}(u|\mathbf{X}) = \theta_\varepsilon(\mathbf{X})^{-1}\int_0^u \exp\{g(v,\mathbf{X}) + \varepsilon(v)\}\,dv$, where $\theta_\varepsilon(\mathbf{x}) = \int_0^1 \exp\{g(v,\mathbf{x}) + \varepsilon(v)\}\,dv$ and $\varepsilon(u) = \varepsilon_1\sin(\pi u) + \varepsilon_2\sin(2\pi u)$ represents the error process of the model (2.2) as a random $\mathrm{L}_2$-element, where $\varepsilon_1$ and $\varepsilon_2$ were chosen as mean zero independent normal random variables with $\mathrm{var}(\varepsilon_1) = 0.1^2$ and $\mathrm{var}(\varepsilon_2) = 0.05^2$. For the random covariate vector $\mathbf{X}_i = (X_{i,1}, X_{i,2})^\top$, we generated $\mathbf{X}_i = (\Phi(V_{i,1}), \Phi(V_{i,2}))^\top$, where $\Phi$ is the standard normal CDF and $\mathbf{V}_i = (V_{i,1}, V_{i,2})^\top \sim N_2(\mathbf{0}, \Sigma)$ are bivariate normal random vectors with mean zero and a covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

We note that both $X_{i,1}$ and $X_{i,2}$ have marginal distributions on $[0,1]$ and are correlated with each other. Then, for $U_{i,1}, \ldots, U_{i,N_i} \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$ independent of $\mathbf{X}_i$, we obtained random samples $\mathbf{Y}_i = \{Y_{i,j} = \mathcal{Q}(U_{i,j}|\mathbf{X}_i) : 1 \le j \le N_i\}$ for each $1 \le i \le n$, so that $Y_{i,1}, \ldots, Y_{i,N_i} \overset{\text{iid}}{\sim} f_i \equiv \Psi_1^{-1}(g(\cdot,\mathbf{X}_i) + \varepsilon_i)$, where the $f_i$ are random response densities and $\varepsilon_i$ are random copies of $\varepsilon$. Ideally, we would have available random copies $(\mathbf{X}_1,f_1), \ldots, (\mathbf{X}_n,f_n)$ of $(\mathbf{X},f)$ satisfying $\mathrm{E}(\Psi_1(f)(u)|\mathbf{X}) = \Psi_1(\mu(\cdot|\mathbf{X}))(u)$ for all $u \in [0,1]$, from which we would aim to infer the additive component functions $g_1, g_2$ and the fitted model (2.1). However, in almost all applications, the response densities $f_i$ need to be inferred from data that they generated. Accordingly, we assume $N_i = N$ iid observations are available for each response distribution and consider scenarios with $N = 200$, 400, and 800 in order to assess how this number affects the estimation performance of the proposed method. For the sample size $n$ of available data points (predictors and response density pairs) to fit the additive density regression, we choose $n = 100$, 400, and 1000.

In Figure 1, we depict random density responses generated in the simulation setting described above. The conditional Fréchet mean $\mu(v|\mathbf{X}) = \Psi_1^{-1}(g(\cdot,\mathbf{X}))(v)$ is illustrated by some fixed covariate vectors $\mathbf{X} = (x_1,x_2)^\top$ together with random realizations of the density response $f = \Psi_1^{-1}(g(\cdot,\mathbf{X}) + \varepsilon)$ at each given $\mathbf{X} = (x_1,x_2)$, respectively. This demonstrates how covariates affect the shape of density realizations through the additive model (4.1) and the nonlinear quantile transformation $\Psi_1$ and how the random densities vary with the additional error in the transformed space.

We write $\mathcal{X}_n = \{(\mathbf{X}_i, \mathbf{Y}_i) : 1 \le i \le n\}$ for the generated sample, $\hat{g}_j(\cdot,\cdot) \equiv \hat{g}_j(\cdot,\cdot;\mathcal{X}_n,\mathbf{h})$ for the smooth backfitting estimators (2.14) of the additive component surfaces $g_j(\cdot,\cdot)$, $j = 1,2$, based on the sample $\mathcal{X}_n$ and $\mathbf{h} = (h_0,h_1,h_2)$ for the bandwidth vector used in (2.13), where $h_0$, $h_1$ and $h_2$ are chosen to adapt to the smoothness of the component function estimators in $u$, $x_1$ and $x_2$, respectively. Throughout we suppress the dependency on $N$ for simplicity of notation. For data-adaptive bandwidth selection, we implemented a shrinkage bandwidth selector (Han, Müller, and Park 2018), based on $K$-fold cross validation (CV). A more detailed description of implementation and bandwidth selection can be found in the online supplement.

We examined the performance of the component estimation for the additive surface estimates $\hat{g}_j(u,x_j)$ in terms of mean integrated squared error (MISE), approximated by

$$\mathrm{MISE}(\hat{g}_j) \approx B^{-1}\sum_{b=1}^{B}\int_0^1\int_0^1 \Big(\hat{g}_j^{(b)}(u,x_j) - g_j(u,x_j)\Big)^2\,du\,dx_j, \tag{4.3}$$

where $\hat{g}_j^{(b)}(u,x_j)$ are the component estimators of $g_j(u,x_j)$ from the $b$th Monte Carlo (MC) sample $\mathcal{X}_n^{(b)} = \{(\mathbf{X}_i^{(b)}, \mathbf{Y}_i^{(b)}) : 1 \le i \le n\}$ and the shrinkage bandwidth selection procedure was applied separately at each Monte Carlo run. The integrated squared bias (ISB) and integrated variance (IV) were also reported, where

$$\mathrm{ISB}(\hat{g}_j) \approx \int_0^1\int_0^1 \Big(\bar{\hat{g}}_j(u,x_j;\mathbf{h}) - g_j(u,x_j)\Big)^2\,du\,dx_j,$$

$$\mathrm{IV}(\hat{g}_j) \approx B^{-1}\sum_{b=1}^{B}\int_0^1\int_0^1 \Big(\hat{g}_j^{(b)}(u,x_j) - \bar{\hat{g}}_j(u,x_j)\Big)^2\,du\,dx_j,$$

with $\bar{\hat{g}}_j(u,x_j) = B^{-1}\sum_{b=1}^{B}\hat{g}_j^{(b)}(u,x_j)$. From the results in Table 1, it can be seen that MISE becomes smaller as $n$ and $N$ increase. It turns out that the sample size $n$ mainly affects IV, while ISB depends primarily on $N$, which is expected, since $N$ determines the precision with which the response densities

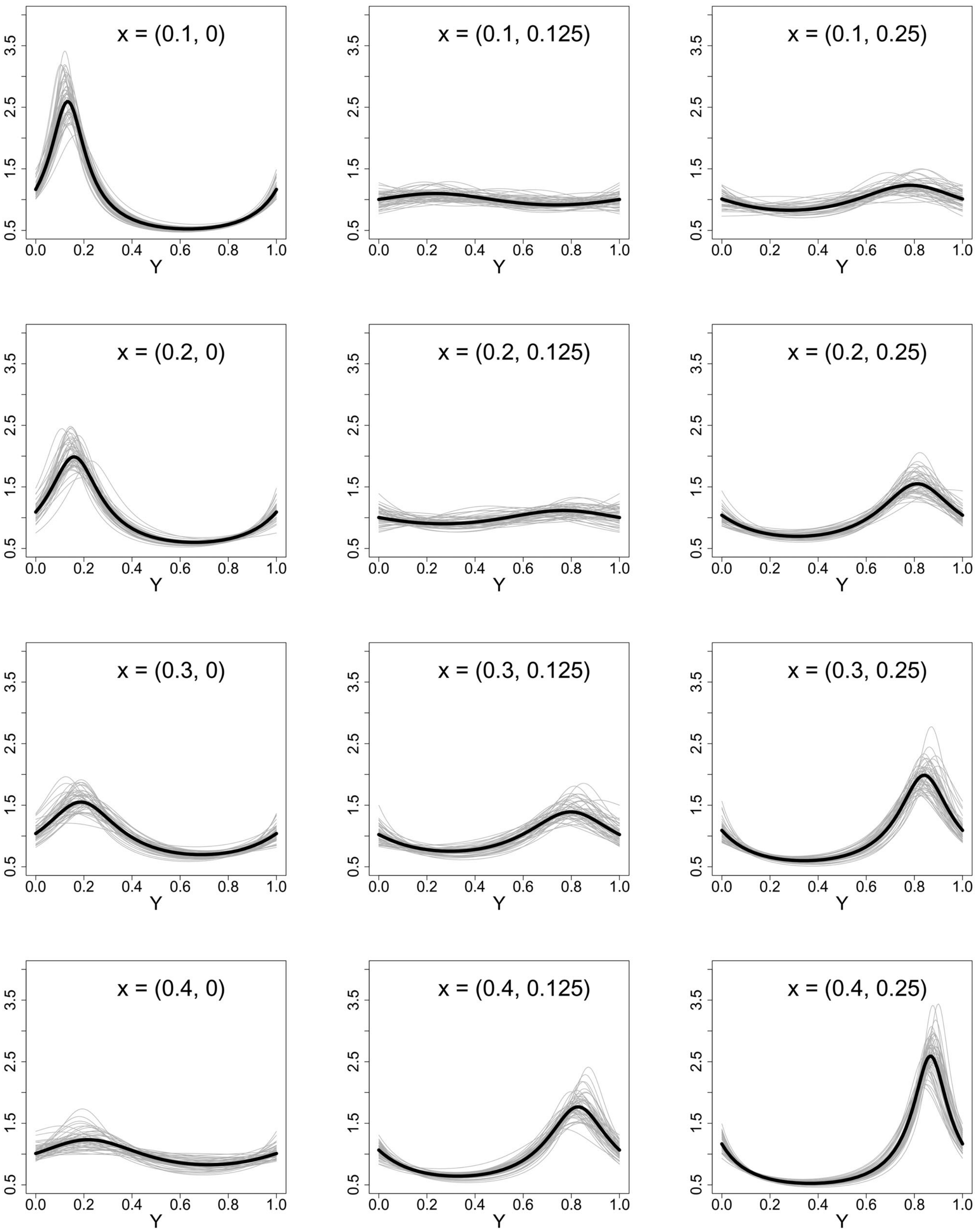

**Figure 1.** Examples of random densities within the simulation setting. For a selection of 12 fixed covariate vectors $\mathbf{x} = (x_1, x_2)^\top$, the solid black line is the density function that corresponds to the conditional Fréchet mean, which depends on the covariates through $\mu(\cdot|\mathbf{x}) = \Psi_1^{-1}(g(\cdot, \mathbf{x}))$. Also depicted are 30 random copies of $f = \Psi_1^{-1}(g(\cdot, \mathbf{x}) + \varepsilon)$, demonstrating the extent to which the additional noise process $\varepsilon$ affects the densities through the back transformation $\Psi_1^{-1}$.

**Table 1.** Mean integrated squared error (MISE), integrated squared bias (ISB) and integrated variance (IV) of the smooth backfitting estimation for additive surface components for $j = 1, 2$, for $B = 500$ Monte Carlo samples with sample sizes $n = 100$, 400, 1000, and $N = 200$, 400, and 800. Here, the sample size $n$ denotes the number of observed density responses and $N$ stands for the number of observations generated from each density.

| | | $N = 200$ | | $N = 400$ | | $N = 800$ | |
|---|---|---|---|---|---|---|---|
| Sample size | Criterion | $j = 1$ | $j = 2$ | $j = 1$ | $j = 2$ | $j = 1$ | $j = 2$ |
| $n = 100$ | MISE | 0.0195 | 0.0304 | 0.0159 | 0.0255 | 0.0131 | 0.0214 |
| | ISB | 0.0150 | 0.0251 | 0.0119 | 0.0208 | 0.0094 | 0.0170 |
| | IV | 0.0045 | 0.0053 | 0.0040 | 0.0047 | 0.0037 | 0.0044 |
| $n = 400$ | MISE | 0.0169 | 0.0230 | 0.0131 | 0.0178 | 0.0103 | 0.0139 |
| | ISB | 0.0154 | 0.0211 | 0.0118 | 0.0162 | 0.0091 | 0.0125 |
| | IV | 0.0015 | 0.0019 | 0.0013 | 0.0016 | 0.0012 | 0.0014 |
| $n = 1000$ | MISE | 0.0151 | 0.0200 | 0.0121 | 0.0164 | 0.0101 | 0.0134 |
| | ISB | 0.0144 | 0.0190 | 0.0114 | 0.0165 | 0.0094 | 0.0125 |
| | IV | 0.0007 | 0.0010 | 0.0007 | 0.0009 | 0.0007 | 0.0009 |

can be assessed. Figure 2 provides graphical illustrations for the average performance of the proposed estimator in terms of bias $\bar{\hat{g}}_j(u, x_j) - g_j(u, x_j)$ and variance $B^{-1}\sum_{b=1}^{B}\left(\hat{g}_j^{(b)}(u, x_j) - \bar{\hat{g}}_j(u, x_j)\right)^2$, obtained from $B$-Monte Carlo runs for $j = 1, 2$.

The proposed additive density regression model can also be used to predict response densities for a given predictor level, where the estimated mean density regression is used to predict a new response. To evaluate this prediction method and to compare it with an alternative approach, we assume for the purposes of our simulation that for each of the sample elements $\{(\mathbf{X}_i, f_i) : 1 \le i \le n\}$ we have a second observation with a new error, leading to a second sample $\{(\mathbf{X}_i, f_i^*) : 1 \le i \le n\}$. We note that $f_i$ and $f_i^*$ share the same predictor levels $\mathbf{X}_i$ and are iid copies. Given the actually observed data $\{(\mathbf{X}_i, \mathbf{Y}_i) : 1 \le i \le n\}$, where $Y_{ij} \sim_{\text{iid}} f_i$, possible predictors for $f_i^*$ in this scenario include: (1) The fitted additive density regression model $\hat{\mu}(\cdot|\mathbf{X}_i)$ as in (2.5), evaluated at predictor level $\mathbf{X}_i$ and denoted as predictor $\hat{f}_{i1}^*$; and (2) Using the $N$ observed data points in $\mathbf{Y}_i$ to nonparametrically estimate the density $f_i$, and thus a second predictor for $f_i^*$ is the kernel density estimate $\hat{f}_i$ (2.9), denoted as predictor $\hat{f}_{i2}^*$.

Since in practice, the correct transformation $\Psi$ is unknown and as a misspecified transformation leads to an incorrect model and increased errors, we developed a data-based transformation selection method for the proposed additive density regression model (2.2), selecting the best transformation by minimizing cross-validation (CV) prediction error, where squared prediction error is the squared Wasserstein distance between predicted and observed densities. To implement this method, we first consider a set of transformations that satisfy the basic requirements, fit the model for each one and then select the transformation that minimizes the CV prediction error. In our numerical illustrations, we included the log-quantile density transformation $\Psi_1$ (1.1) and the log-hazard transformation $\Psi_2$ (1.2), then selected the transformation associated with the smallest CV prediction error; further details are in the online supplement.

For the implementation of additive density regression with transformation selection and the above described kernel method as comparison method, we evaluated the performance of response density prediction for estimates $\hat{f}_{i1}$ (additive density regression with transformation selection) and $\hat{f}_{i2}$ (kernel method) by the average-squared prediction error (ASPE)

$$\text{ASPE}_\ell = B^{-1}n^{-1}\sum_{b=1}^{B}\sum_{i=1}^{n} d_W(f_i^{*,b}, \hat{f}_{i\ell}^{*,b})^2, \quad \ell = 1, 2, \qquad (4.4)$$

based on $B$ Monte Carlo runs. Here $d_W(f, g)^2 = \int_0^1 \left(F^{-1}(u) - G^{-1}(u)\right)^2 du$ is the Wasserstein L$_2$-distance between two distribution functions (Villani 2003), $\hat{f}_{i1}^{*,b} = \hat{\mu}^{(b)}(\cdot|\mathbf{X}_i^{(b)}; \hat{\mathbf{h}}^{(b)})$ is the conditional Fréchet mean estimator of $\mu(\cdot|\mathbf{X}_i)$ based on the $b$th Monte Carlo training sample $\mathcal{X}_n^{(b)}$ and using data-driven shrinkage bandwidths $\hat{\mathbf{h}}^{(b)}$, $\{(\mathbf{X}_i^{(b)}, f_i^{*,(b)}) : 1 \le i \le n\}$ is the $b$th Monte Carlo test sample, and $\hat{f}_{i2}^{*,b}$ is the kernel estimator for the $b$th Monte Carlo sample.

We note that the alternative kernel estimator can only be reasonably used when one already has observed one response at a given predictor level and aims to predict a second response at the same level. Therefore, we restrict the comparison to this situation, which does not commonly occur in practice, so that the kernel density estimator in general cannot be used for prediction purposes. The prediction error for this alternative kernel estimator stems from two sources, the discrepancy of a kernel estimator based on a sample of $N$ observations from the true density $f_i$; and the deviation of a second response at the same predictor level from a first response due to the error in the densities $f_i$ themselves, as the response densities are assumed to include a random error in addition to being generated by the additive model, in analogy to regular regression models.

Comparing predictors $\hat{f}_{i1}^*$ using the proposed additive density regression with transformation selection and $\hat{f}_{i2}^*$ using kernel estimators, we find that in almost all cases the additive density regression estimator outperforms the kernel estimator, with simulation results reported in Table 2. The kernel method tends to achieve somewhat smaller prediction errors than the additive model when $N$ is large and $n$ is small, as then the kernel density estimator is closer to the actual density $f_i$, so that this part of the prediction error becomes smaller, while for smaller $n$ the additive model may suffer from higher variance. However, since the kernel estimator cannot use the covariate information, it does not lead to improved prediction for increasing sample sizes $n$, in contrast to the additive density regression method, which rapidly improves with increasing sample size, through improved estimation of the additive component functions $g_j$.

### 4.2. Popularity of Baby Names in the United States

We apply the proposed additive density regression to the United States (US) baby names data, which are available

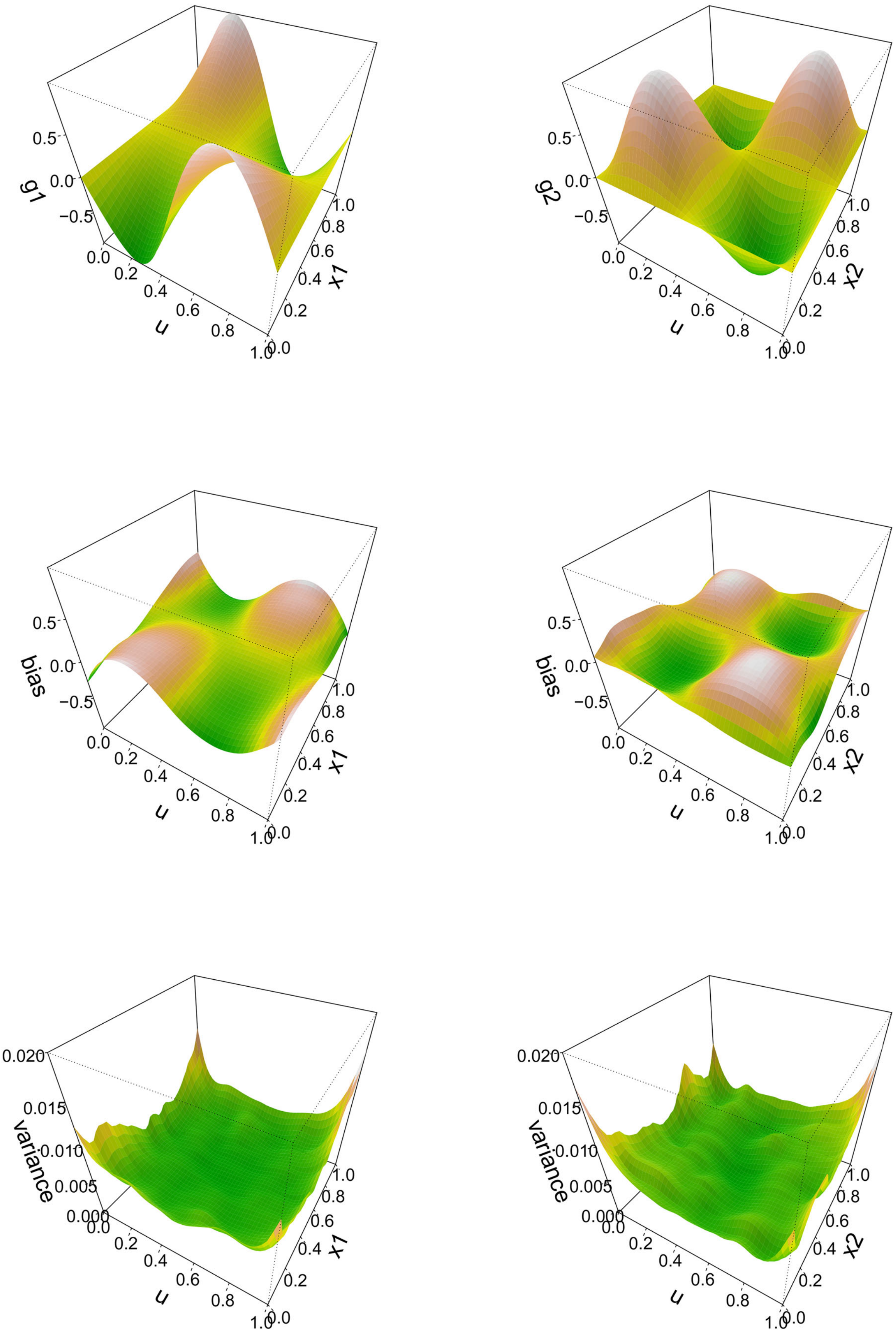

**Figure 2.** Top panels: True additive component surfaces $g_1$ and $g_2$. The average performance of the proposed method is illustrated in terms of bias (middle panels) and variance (bottom panels) of the additive component estimates, obtained from 500 Monte Carlo runs with sample sizes $n = 100$ and $N = 200$, respectively. Here, $n$ refers to the number of data points $(\mathbf{X}_i, f_i)$ and $N$ to the number of observations generated by each response density $f_i$.

**Table 2.** Average-squared prediction errors (ASPE) of fitted response densities for the proposed additive density regression with transformation selection and the kernel density estimator (KDE) based on data from a first response when the second response to be predicted is sampled at the same predictor levels. The average prediction errors are based on $B = 500$ Monte Carlo repetitions with sample sizes $n = 100, 400, 1000$ and $N = 200, 400$, and 800, where the sample size $n$ denotes the number of observed density responses and $N$ stands for the number of observations generated from each density. The ASPEs are given up to a factor of $10^{-2}$ and the corresponding standard deviations are in parenthesis.

| Sample size | | Average squared prediction error ($\times 10^{-2}$) | |
|---|---|---|---|
| $N$ | $n$ | Additive modeling | KDE |
| 200 | 100 | 0.0701 (0.0263) | 0.1186 (0.0135) |
| | 400 | 0.0421 (0.0060) | 0.1181 (0.0064) |
| | 1000 | 0.0383 (0.0028) | 0.1179 (0.0043) |
| 400 | 100 | 0.0601 (0.0234) | 0.0744 (0.0083) |
| | 400 | 0.0346 (0.0039) | 0.0744 (0.0041) |
| | 1000 | 0.0318 (0.0020) | 0.0746 (0.0026) |
| 800 | 100 | 0.0521 (0.0195) | 0.0511 (0.0056) |
| | 400 | 0.0292 (0.0030) | 0.0510 (0.0028) |
| | 1000 | 0.0269 (0.0014) | 0.0511 (0.0017) |

from the US Social Security Administration for births after 1879. Frequencies with which a name is given are recorded by calendar year, for details see *https://www.ssa.gov/oact/babynames/background.html*. The raw data are available at *https://catalog.data.gov/organization/ssa-gov*. We focus here on male baby names that have newly appeared and thus been given for the first time between 1935 and 1985, and view the distribution of a given name as a function of the time after it appeared for the first time, which will be designated as time 0 across all names.

The distribution of a name over time after it appeared for the first time quantifies its popularity trend. The relative time scale where the first appearance of the name occurs at time 0 allows us to directly compare different names. To obtain the distribution of a name over calendar time after its first appearance, we truncate the data at 30 years after appearance of the name and then construct relative frequencies for the first 30 years during which the name is being given. Since the most recent name, we include in this study appeared in 1985, and the data ended in 2016, the distribution of the frequency of the name for the first 30 years after it appears is equally available for all names. We included only names with more than 30 instances over their first 30 years after appearance and names that newly appeared between 1935 and 1985. To satisfy the basic requirements of the proposed approach as in Section 3, we selected names for which the marginal distribution of the covariates of interest was compactly supported and bounded away from zero on a domain, where the lower and upper limits of this domain were empirically defined by the minimum and maximum of the observed covariates, respectively. These preprocessing steps led to a sample of $n = 2118$ baby names.

We considered two scenarios of additive density modeling for the baby name densities as responses, employing the log-quantile transformation $\Psi = \Psi_1$ which resulted from the CV selection of the best transformation when compared against the log-hazard transformation $\Psi_2$. In a first scenario (Model 1), we utilized two predictors, the calendar year when the name first appeared and the total frequency of the name over its first 30 years, motivated by the idea that trends in name popularity may have changed over the years, and that the shape of the popularity trend curve is related to the total popularity of a name. In a second scenario (Model 2), we considered the same predictors but additionally added the frequency of the name in the first year it appeared, and also the cumulative frequencies observed for the name over the first 5, 10, and 20 years. Model 1 played the role of a most parsimonious approach and Model 2 that of a full model. As Model 2 captures more features than Model 1, it is of interest how many features are actually needed for reasonably good predictions of the name popularity trends.

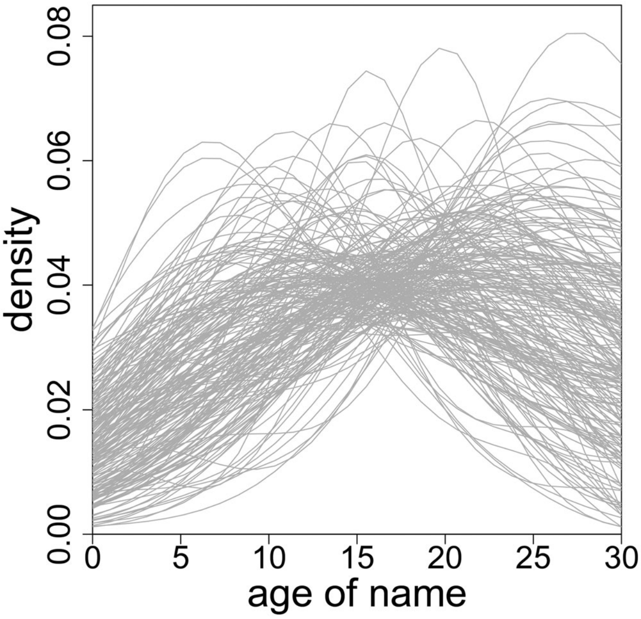

**Figure 3.** A subsample of 200 densities for male baby name distributions over an interval of 30 years, where each name started to appear at time 0. The total sample size of included baby names is $n = 2118$.

We mention here that alternatively, one could also take the entire initial phase of the popularity curve of a name on a specified interval during the initial phase and view it as a functional predictor. In scenarios with functional predictors, our procedures can be implemented with vectorized versions of the predictor function, such as functional principal components, extending the additive model for the case where both predictors and responses are functional (see, e.g., Müller and Yao 2008) to the case where responses are density functions.

Initial inspection showed that most of the baby name densities or popularity trend curves have well-defined modes. Examples are in Figure 3. Note that the raw data are relative frequencies over yearly bins, which then can be smoothed to construct densities. For this we used local linear kernel smoothers (Müller, Wang, and Capra 1997), for which we employed Gaussian kernels with bandwidths chosen by 5-fold cross-validation (CV). The resulting smooth curves are constrained to be nonnegative and standardized to integrate to 1, so that they are bona fide density functions (Gajek 1986). We view these reconstructed densities as our responses $f_i$, $i = 1, \ldots, n$.

For additive surface estimation, we apply the implementation strategy of smooth backfitting, as introduced in the previous subsection 4.1, including the data-adaptive selection of bandwidths for the mean part and additive components followed by the shrinkage bandwidth selector with 5-fold CV. To achieve

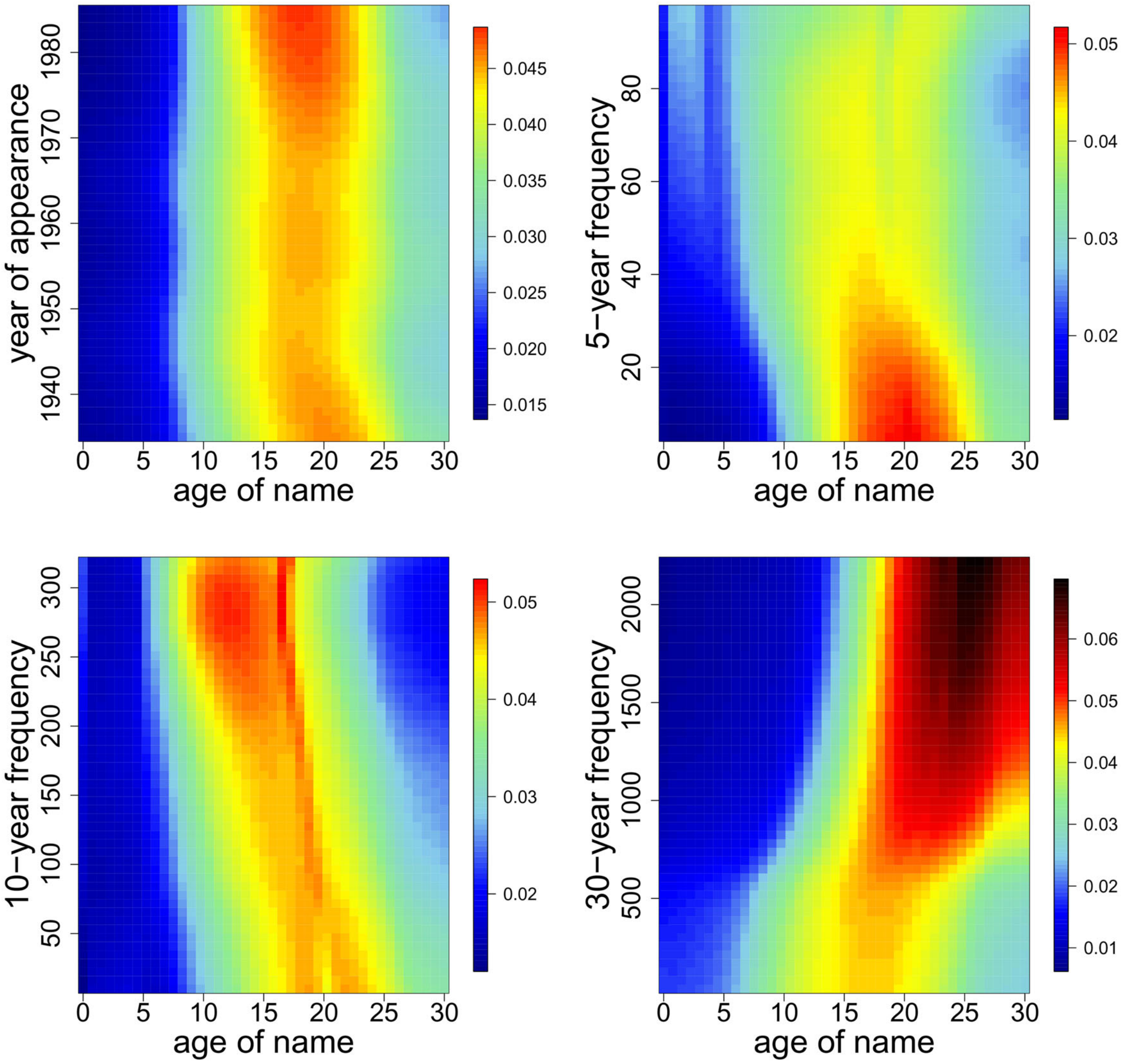

**Figure 4.** Fitted additive density regression model (2.1) for male baby names for a model with four predictors, applying estimates (2.5). The predictors are calendar year of the first appearance of the name (top left, FVE = 0.71%), 5-year frequency of the name (top right, FVE = 7.69%), 10-year frequency (bottom left, FVE = 6.49%) and 30-year frequency (bottom right, FVE = 15.78%). In our analysis, we eliminated two predictors, frequency of the name in the first year after its appearance and 20-years frequencies through backward elimination.

parsimonious modeling we consider a backward elimination procedure for Model 2, where the contribution of individual additive components is quantified by an empirical version of the fraction of variance explained (FVE) criterion. Here, the empirical FVE of the *j*th component $X_j$ is defined by $V_j/V_\infty$, where $V_j = \sum_{i=1}^n d_W(f_i, \hat{\mu}_{i,0})^2 - \sum_{i=1}^n d_W(f_i, \hat{\mu}_{i,j})^2$ and $V_\infty = \sum_{i=1}^n d_W(f_i, \hat{\mu}_{i,0})^2$ with $\hat{\mu}_{i,0} = \Psi^{-1}(\hat{g}_0)$ and $\hat{\mu}_{i,j} = \Psi^{-1}(\hat{g}_0 + \hat{g}_j(\cdot, X_{i,j}))$. We note that $V_j$ corresponds to the notion of extra sum of squared errors in multiple linear regression, which measures the reduction in the sum of squared errors due to the addition of a predictor; see also Petersen and Müller (2016).

We then find the best model by backward elimination, where predictors are successively removed, considering at each step to remove the predictor that has the smallest FVE among the included predictors. Here, FVE is sequentially computed after every elimination step. We stop the backward elimination procedure if the mean-squared error (MSE) increases over the previous step, where the MSE is defined by $n^{-1} \sum_{i=1}^n d_W(f_i, \hat{\mu}_i^{(r)})^2$ and $\hat{\mu}_i^{(r)}$ is the fitted density of $f_i$ in the *r*th elimination step, where $\hat{\mu}_i^{(0)} = \hat{\mu}_i \equiv \Psi^{-1}(\hat{g}_0 + \sum_{j=1}^d \hat{g}_j(\cdot, X_{i,j}))$. In our analysis, $d = 6$ for Model 2, which provides the starting model for the elimination procedure. We then ran the backward elimination procedure, whereby two of the 6 predictors where removed, the frequency of the name in the year it first appeared, and its 20-year frequency. This left four predictors for the final additive density regression model, namely the calendar year when the name appeared first, and its cumulative 5-, 10- and 30-year frequencies.

Figure 4 demonstrates the effects of the additive components on the fitted densities by heat maps, which are better interpretable than the plots of the additive component functions $g_j$, of which an example is shown in Figure 2 for the simulation setting, while here we illustrate the effects of individual predictors on the estimates (2.5) in model (2.1). We find that names that first appeared between 1935 and 1945 have slightly delayed modes in their densities compared to names that appeared after 1945. Moreover, overall more popular names have a slower increase in their densities and their mode occurs after more than 20 years

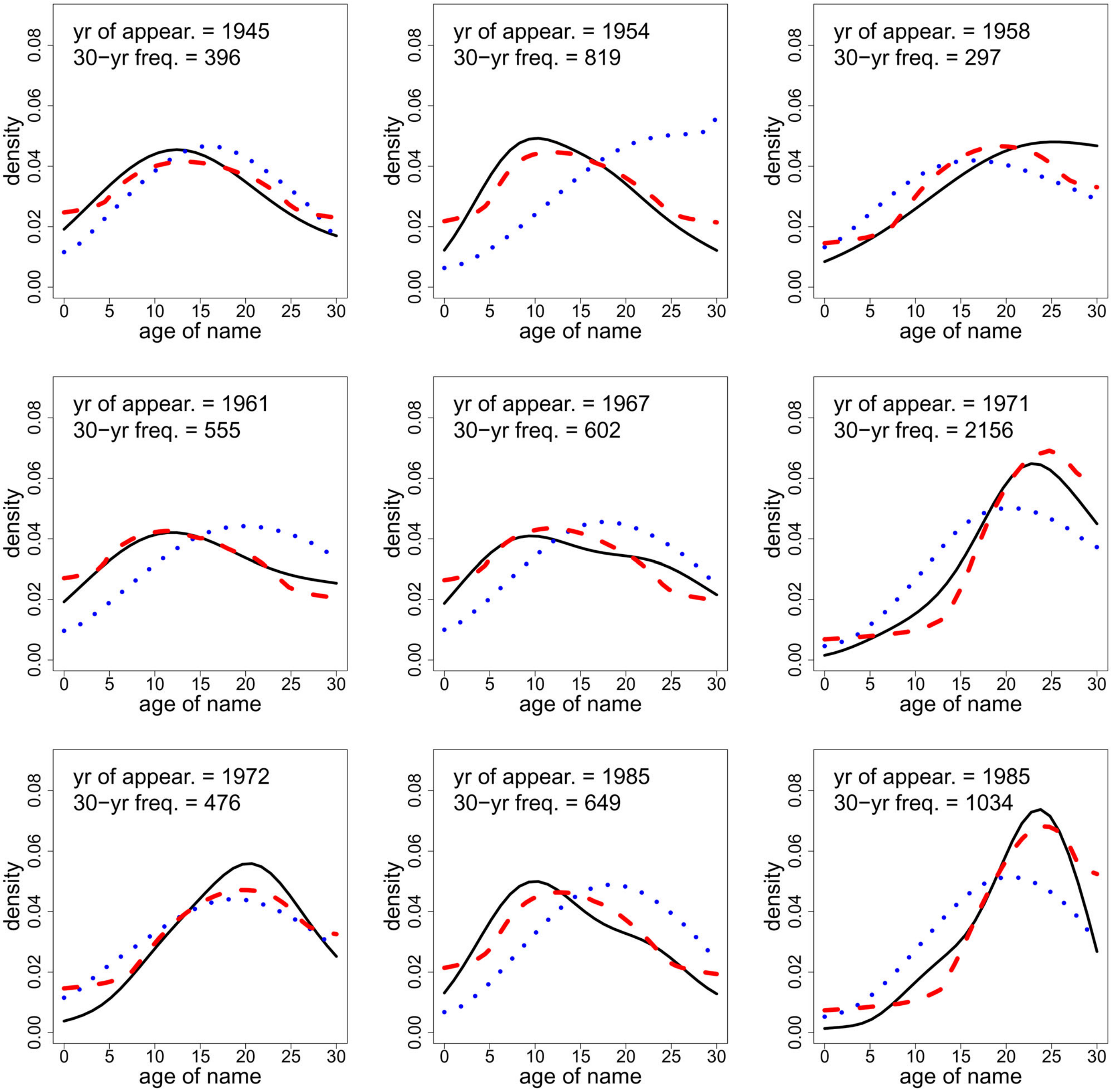

**Figure 5.** Examples of density prediction, comparing Model 1 (with blue dotted lines as predicted densities) and Model 2 (red dashed lines), where the black solid lines correspond to observed (estimated) baby name density responses.

from the time the name first appears, while overall less popular names have faster increasing densities with earlier modes.

Five- and ten-year frequencies that remained included as predictors after the backwards elimination procedure also have notable effects as shown in the upper right and lower left panel of Figure 4. The heat map corresponding to the 5-year frequency illustrates that less popular names during the first 5 years tend to have later and more expressed modes than those which are more popular during the first five years. The 10-year frequency of the name as predictor is associated with a monotone decreasing trend in the peak locations, such that lower 10-year frequencies are associated with later mode locations and higher frequencies with earlier locations. It is also worth mentioning that the calendar year when the name appeared still constitutes an important predictor for the baby name dynamics even though the corresponding FVE was small with 0.71%, since the reduced model excluding this calendar year predictor had significantly larger MSE than the model with this predictor included. The effect of the overall 30-year frequency on the densities was similar as in Model 1, if not even more pronounced.

We also studied the prediction performance of the proposed additive density regression model, with examples of density prediction illustrated in Figure 5. We compared the prediction performance between two additive models by the square root of mean prediction error ($\times 10^{-2}$), obtained from 5-fold cross-validation, which for Model 1 and Model 2 after backward elimination were found to be 2.468 and 1.881, respectively. Therefore, we find that Model 2 with backward elimination outperforms Model 1 for prediction.

## Supplementary Material

The online supplementary materials contain the detailed bandwidth selection procedure and the proof of Theorem 2.

## Funding

Research supported by NSF Grants DMS-1407852 and DMS-1712864. Research supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. 2019R1A2C3007355).

## References

Bhattacharya, P. K., and Gangopadhyay, A. K. (1990), "Kernel and Nearest-neighbor Estimation of a Conditional Quantile," *The Annals of Statistics*, 18, 1400–1415. [1]

Bigot, J., Gouet, R., Klein, T., and López, A. (2017), "Geodesic PCA in the Wasserstein space by convex PCA," *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, 53, 1–26. [1]

Delicado, P. (2007), "Functional k-sample Problem When Data Are Density Functions," *Computational Statistics*, 22, 391–410. [1]

——— (2011), "Dimensionality Reduction When Data Are Density Functions," *Computational Statistics*, 55, 401–420. [1]

Dunson, D., Pillai, N., and Park, J.-H. (2007), "Bayesian Density Regression," *Journal of the Royal Statistical Society*, Series B, 69, 163–183. [1]

Egozcue, J. J., Diaz-Barrero, J. L., and Pawlowsky-Glahn, V. (2006), "Hilbert Space of Probability Density Functions Based on Aitchison Geometry," *Acta Mathematica Sinica*, 22, 1175–1182. [1]

Gajek, L. (1986), "On Improving Density Estimators Which Are Not Bona Fide Functions," *Annals of Statistics*, 14, 1612–1618. [11]

Hall, P., and Müller, H.-G. (2003), "Order-preserving Nonparametric Regression, With Applications to Conditional Distribution and Quantile Function Estimation," *Journal of the American Statistical Association*, 98, 598–608. [1]

Hall, P., Wolff, R. C. L., and Yao, Q. (1999), "Methods for Estimating a Conditional Distribution Function," *Journal of the American Statistical Association*, 94, 154–163. [1]

Han, K., Müller, H.-G., and Park, B. U. (2018), "Smooth Backfitting for Additive Modeling With Small Errors-in-variables, With an Application to Additive Functional Regression for Multiple Predictor Functions," *Bernoulli*, 24, 1233–1265. [7]

Kneip, A. and Utikal, K. J. (2001), "Inference for Density Families Using Functional Principal Component Analysis," *Journal of the American Statistical Association*, 96, 519–542. [1]

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile Smoothing Splines," *Biometrika*, 81, 673–680. [1]

Lee, Y. K., Mammen, E., and Park, B. U. (2010), "Backfitting and Smooth Backfitting for Additive Quantile Models," *Annals of Statistics*, 38, 2857–2883. [5]

——— (2012), "Flexible Generalized Varying Coefficient Regression Models," *Annals of Statistics*, 40, 1906–1933. [5]

Li, Q., Lin, J., and Racine, J. S. (2013), "Optimal Bandwidth Selection for Nonparametric Conditional Distribution and Quantile Functions," *Journal of Business & Economic Statistics*, 31, 57–65. [1]

Mammen, E., Linton, O., and Nielsen, J. (1999), "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions," *Annals of Statistics*, 27, 1443–1490. [5]

Mammen, E., and Park, B. U. (2006), "A Simple Smooth Backfitting Method for Additive Models," *Annals of Statistics*, 34, 2252–2271. [3]

Menafoglio, A., Guadagnini, A., and Secchi, P. (2014), "A Kriging Approach Based on Aitchison Geometry for the Characterization of Particle-size Curves in Heterogeneous Aquifers," *Stochastic Environmental Rearch and Risk Assessment*, 28, 1835–1851. [1]

Menafoglio, A., Secchi, P., and Guadagnini, A. (2016), "A Class-kriging Predictor for Functional Compositions With Application to Particle-size Curves in Heterogeneous Aquifers," *Mathematical Geosciences*, 48, 463–485. [1]

Müller, H.-G., Wang, J.-L., and Capra, W. B. (1997), "From Lifetables to Hazard Rates: The Transformation Approach," *Biometrika*, 84, 881–892. [11]

Müller, H.-G., and Yao, F. (2008), "Functional Additive Models," *Journal of the American Statistical Association*, 103, 1534–1544. [11]

Panaretos, V. M., and Zemel, Y. (2016), "Amplitude and Phase Variation of Point Processes," *The Annals of Statistics*, 44, 771–812. [1]

Petersen, A., Chen, C.-J., and Müller, H.-G. (2018), "Quantifying and Visualizing Intraregional Connectivity in Resting-State Functional Magnetic Resonance Imaging with Correlation Densities," *Brain Connectivity*, 9, 37–47. [1]

Petersen, A., and Müller, H.-G. (2016), "Functional Data analysis for Density Functions by Transformation to a Hilbert Space," *Annals of Statistics*, 44, 183–218. [1,2,3,5,7,12]

——— (2018), "Fréchet Regression for Random Objects With Euclidean Predictors," *Annals of Statistics*, 47, 691–719. [1,2]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, Springer Series in Statistics (2nd ed.), New York: Springer. [1]

Roussas, G. G. (1969), "Nonparametric Estimation of the Transition Distribution Function of a Markov Process," *The Annals of Mathematical Statistics*, 40, 1386–1400. [1]

Sen, R., and Ma, C. (2015), "Forecasting Density Function: Application in Finance," *Journal of Mathematical Finance*, 5, 433. [1]

Talská, R., Menafoglio, A., Machalová, J., Hron, K., and Fišerová, E. (2018), "Compositional Regression With Functional Response," *Computational Statistics & Data Analysis*, 123, 66–85. [1]

Villani, C. (2003), *Topics in Optimal Transportation*, Graduate Studies in Mathematics, Vol. 58, American Mathematical Society, Providence, Rhode Island. [9]

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and its Application*, 3, 257–295. [1]

Yu, K., Park, B. U., and Mammen, E. (2008), "Smooth Backfitting in Generalized Additive Models," *Annals of Statistics*, 36, 228–260. [5]

Zhang, X., Park, B. U., and Wang, J.-L. (2013), "Time-varying Additive Models for Longitudinal Data," *Journal of the American Statistical Association*, 108, 983–998. [4]

Zhang, Z., and Müller, H.-G. (2011), "Functional Density Synchronization," *Computational Statistics & Data Analysis*, 55, 2234–2249. [1]