Nonparametric Estimation of Repeated Densities with Heterogeneous Sample Sizes

Jiaming Qiu, Xiongtao Dai, and Zhengyuan Zhu

Department of Statistics, Iowa State University, Ames, IA

ABSTRACT

We consider the estimation of densities in multiple subpopulations, where the available sample size in each subpopulation greatly varies. This problem occurs in epidemiology, for example, where different diseases may share similar pathogenic mechanism but differ in their prevalence. Without specifying a parametric form, our proposed method pools information from the population and estimate the density in each subpopulation in a data-driven fashion. Drawing from functional data analysis, low-dimensional approximating density families in the form of exponential families are constructed from the principal modes of variation in the log-densities. Subpopulation densities are subsequently fitted in the approximating families based on likelihood principles and shrinkage. The approximating families increase in their flexibility as the number of components increases and can approximate arbitrary infinite-dimensional densities. We also derive convergence results of the density estimates formed with discrete observations. The proposed methods are shown to be interpretable and efficient in simulation studies as well as applications to electronic medical record and rainfall data. Supplementary materials for this article are available online.

1. Introduction

Density estimation is one of the most fundamental tasks in statistics and machine learning. We consider the problem of repeated density estimation, namely, to estimate the distributions of a variable in multiple subpopulations with similar nature. For example, the distributions of interest can be the age distributions of patients with different diseases, or the precipitation distributions in multiple regions. Repeated density functions can be modeled as the realizations of a random density element that is subject to the positivity and unit integral constraints. Moreover, density functions are in practice unobserved and are accessed through discrete samples, rendering additional difficulties for data analysis. The primary objective of this work is to address these difficulties and derive reliable yet flexible density estimates for the subpopulations given discrete observations, in a fully data-driven and efficient manner.

Most of the existing work consider random densities that are either completely observed or are reconstructed from densely sampled observations. This type of density objects was first analyzed by Kneip and Utikal (2001) by directly applying functional principal component analysis (FPCA), a technique that has later been applied to analyze time-varying densities (Huynh et al. 2011; Tsay 2016; Chu et al. 2018). Recognizing the constraints in density functions, Delicado (2011) and Petersen and Müller (2016) proposed to analyze densities by using one-to-one transformations to map them to an unconstrained space, where the transformed densities are then represented using FPCA and subsequently back transformed into densities. Different transformations have been considered, such as the log-hazard and log-quantile density transformations (Petersen and Müller 2016; Kim, Lee, and Park 2020), the quantile transformation (Panaretos and Zemel 2016), and the square root transformation with the Riemannian logarithm map (Srivastava, Jermyn, and Joshi 2007; Dai 2021) which generates a Hilbert sphere geometry. In particular, Hron et al. (2016) applied the centered log-ratio transformation to obtain FPCA for densities as compositional data in the Bayes space (Egozcue, Díaz-Barrero, and Pawlowsky-Glahn 2006; Delicado 2011; Egozcue et al. 2013; van den Boogaart, Egozcue, and Pawlowsky-Glahn 2014). We will adopt the centered log-ratio transformation (Hron et al. 2016) as a building block to obtain positive density estimates, noting that none of the methods cited above are directly applicable in our setting detailed next.

We instead consider fitting sparsely observed random densities with as few as a handful of data points available. Subpopulations with small sample sizes arise in a number of possible situations when the sampling mechanism dictates heterogeneous sample sizes. A first situation is when the availability and cost of observation highly varies among subpopulations (Wakefield and Elliott 1999), with some groups easy but the rest hard to reach (Bonevski et al. 2014). This scenario is illustrated by the Medical Information Mart for Intensive Care (MIMIC) data (Johnson et al. 2016) in our first data application, where patient records are ample for common diseases such as diabetes and coronary heart disease but scarce for uncommon conditions. A second generating mechanism is that the data collection process had been established for different duration in different
subpopulations (Fithian and Wager 2015), which is illustrated by our second data application considering rainfall recorded by weather stations around the Haihe River basin in northern China that differ in the number of years when data are available.

Our proposal is a nonparametric approach that directly targets the repeated densities, which are modeled as realizations of a density-valued random element, and the discrete observations from each subpopulation are modeled as iid realizations from the corresponding density. Information is pooled over all subpopulations in a two-step procedure. First, we identify low-dimensional structures to parsimoniously approximate the densities using data from subpopulations densely sampled by applying FPCA to the centered log-ratio of densities that lie in an unconstrained $L^2$ space. When transformed back to densities, the low-dimensional structures constitute exponential families, where the sufficient statistics are the eigenfunctions of the transformed densities, and each density element is compactly represented by its natural parameter or, equivalently, moment coordinate (Amari 2016). Second, a new and potentially sparsely observed subpopulation is fitted within the approximating family, obtaining maximum likelihood estimate (MLE), and the coordinate estimate is then shrunk toward the population using distributional information, further enforcing information borrowing. Two shrinkage estimates, namely the maximum a posteriori (MAP) estimate and best linear unbiased prediction (BLUP), are proposed based on shrinking the natural parameter and moment coordinate, respectively.

To illustrate the idea, suppose that the random densities are generated from the normal distribution family $\mathcal{P} = \{p_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma > 0\}$ with densities $p_{\mu,\sigma}(x) \propto \exp\{x(\mu/\sigma^2) - x^2/(2\sigma^2)\}$ truncated to $x \in [-1, 1)$. Observe that a random element $p$ taking values in $\mathcal{P}$ satisfies $\log p \subset \text{span}\{1, x, x^2\}$ and thus lies in a low-dimensional space. The low-dimensional structure can be recovered by applying dimension reduction techniques such as FPCA on $\log p$. When inferring the distribution of a subpopulation, information borrowing is performed within the low-dimensional space and is analogous to that for the classical linear mixed effects model. Remarkably, the proposed methods estimate the low-dimensional structure and pool information fully nonparametrically without imposing parametric shape assumptions. A worked simulation example is included in Section S2.1 of the supplementary materials.

The proposed methods combine the flexibility from nonparametric density estimation and the statistical efficiency and interpretability offered by parametric inference and information borrowing. Our fitting of a density with sparse observations follows likelihood principles in an estimated exponential family, so the required amount of observations by the proposed methods could be substantially fewer than what would otherwise be required for non-parametric estimates. Moreover, the exponential family approximation has advantages in computation and interpretability. The approximating family increases in flexibility as its dimension increases and can approximate an arbitrarily complex density. Rates of convergence of the proposed MLE under the Kullback–Leibler (KL) divergence and $L^1$ norm are established under the asymptotics that the number of training subpopulations, the sample size of each subpopulation, and the sample size of the new subpopulation diverge to infinity. This result holds for general infinite-dimensional random densities under smoothness and boundedness conditions. If the random density actually lies in a finite-dimensional exponential family, then the proposed method converges under the KL divergence to the truth with a nearly-parametric rate. If the sample size in the new subpopulation is fixed, the proposed MLE converges to the MLE in the best approximating exponential family. The literature on modeling sparsely observed random densities is scant, and a closely related topic is replicated point processes. Given sparsely sampled replicated point processes, Wu, Müller, and Zhang (2013), taking a similar approach as Kneip and Utikal (2001), proposed to pool information over multiple point process realizations using local smoothing and directly model the intensity functions using FPCA. Gervini (2016) and Gervini and Khanal (2019) considered semi-parametric models to decompose the intensity and the log-intensity functions of the point processes, respectively. These methods are applicable to sparsely sampled processes where the intensity estimates can be normalized to produce density estimates. However, the method of Wu, Müller, and Zhang (2013) can produce negative estimates and thus needs to use projection to nonnegative functions, and Gervini (2016) formulated a nonnegative decomposition for the intensity function which is costly in computation. In contrast, our methods approximate the structure of the underlying density family using a sequence of exponential families, obtained from the training subpopulations, which is subsequently applied to produce intrinsically positive density estimates for any sparsely observed subpopulations. Gervini and Khanal (2019) used a similar multiplicative structure for the intensity functions as in our model and is thus considered as a comparison in our simulation studies. However, this method resorts to constrained optimization for smoothing spline fitting and information borrowing and thus can be computationally costly. Panaretos and Zemel (2016) defined the amplitude and phase variation for point processes and considered the time-warping problem under a Wasserstein geometry; the methods focused on densely observed Poisson point processes so is inapplicable to the situation we consider.

The proposed models and estimation are introduced in Section 2. Numerical properties of the proposed methods are investigated and compared to alternative parametric and nonparametric density estimates in Section 3. Data illustrations with age-at-admission of ICU patients and rainfall data in the Haihe River basin are included in Section 4. Theoretical results are stated in Section 5. Additional simulation results, discussion, and detailed proofs are included in the supplementary materials.

2. Models and Methods

2.1. Data Setup and Goals

Let $\{X_{ij} : j = 1, \ldots, N_i, i = 1, \ldots, n\}$ be the observations of a continuous variable in subpopulation $i = 1, \ldots, n$; for example, the age-at-admission to the Intensive Care Unit (ICU) as described in Section 4.1 grouped by primary diagnosis and sex. Here $X_{i1}, \ldots, X_{iN_i}$ are iid observations from the $i$th subpopulation with density $p_i$. Densities $p_1, \ldots, p_n$ are modeled as iid realizations of a random density $p$ taking values in family $\mathcal{P}$, where $\mathcal{P}$ consists of densities that are positive on a common compact
interval of interests $T \subset \mathbb{R}$. For example, the susceptibility age profiles of different diseases may be determined by latent random characteristics, so the age distribution of a disease can be regarded the realization of a density-valued random element.

To derive a low rank approximation to the densities, we adopt $p_0$ as the density of the $i$th subpopulation, while the distribution of the random density $p$ as the population.

Our goal is to estimate the density $p_0$ of a new subpopulation with a few iid observations $X_{0j}$, $j = 1, \ldots, N$ from $p_0$, where the number of observations $N$ can potentially be very small. For example, this problem occurs when the goal is to estimate the age distribution for an uncommon disease. We will first obtain low-dimensional structures of $P$ from subpopulations where ample observations are available, for example, from records for more common conditions, and apply the structure to estimate the density in the subpopulation with a small sample size.

### 2.2. Proposed Repeated Density Estimation

#### 2.2.1. Constructing Approximating Family

We propose to approximate distribution family $P$ with low rank exponential families by using a functional data analytic approach. The scenario where the distribution of the random density $p$ is given is discussed here, and the scenario where the distribution must be estimated is described in Section 2.2.4.

The collection of densities forms an infinite-dimensional constraint manifold instead of a Hilbert space. To borrow from the rich literature studying functional data as objects in a Hilbert space, we follow a transformation approach (Petersen and Müller 2016; Hron et al. 2016) and analyze the transformed densities as Hilbertian elements. The centered log-ratio transformation (Egozcue, Díaz-Barrero, and Pawlowsky-Glahn 2006), denoted as $\psi : P \rightarrow L^2(T)$, maps a positive density $q \in P$ into the $L^2$ space, obtaining

\[
(\psi q)(t) = \log q(t) - \frac{1}{|T|} \int_T \log q(x)dx, \quad t \in T. \tag{2.1}
\]

To derive a low rank approximation to the densities, we adopt the FPCA method based on the transformed trajectory $f = \psi p$ as proposed by Hron et al. (2016) and apply the Karhunen–Loève expansion. Assuming that $f$ is square integrable, obtain

\[
f(t) = (\psi p)(t) = \mu(t) + \sum_{k=1}^{\infty} \eta_k \varphi_k(t), \quad t \in T, \tag{2.2}
\]

where $\mu(t) = Ef(t)$ is the mean function, $\eta_k = \int_T (f(t) - \mu(t)) \varphi_k(t)dt$ is the $k$th component score, and $(\lambda_k, \varphi_k)$ is the $k$th eigenvalue–eigenfunction pair of the covariance function $G(s,t) = \text{cov}(f(s) - \mu(s), f(t) - \mu(t))$ satisfying $\lambda_k \varphi_k(t) = \int_T G(s,t) \varphi_k(s)ds$ on $t \in T$, for $k = 1, 2, \ldots$. The eigenfunctions are orthonormal, and the eigenvalues are nonnegative and non-increasing, satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

The transformed density $f$ resides in the Hilbert subspace orthogonal to the constant functions, so the mean $\mu$ and eigenfunctions $\varphi_k$ lie in the same space. Applying the back-transformation $\psi^{-1}$ to the affine spans of the mean and the leading eigenfunctions lead to a low rank approximating density family to $P$ that take the form of an exponential family. More precisely, for $K = 1, 2, \ldots$, the $K$-dimensional approximation $P_K$ to $P$ is the exponential family

\[
P_K = \left\{ p_0 : \theta \in \mathbb{R}^K, B_K(\theta) < \infty \right\}, \quad \text{with} \tag{2.3}
\]

\[
p_0(t) = \exp \left( \mu(t) + \sum_{k=1}^{K} \theta_k \varphi_k(t) - B_K(\theta) \right), \quad t \in T, \tag{2.4}
\]

where $B_K(\theta) = \log \left( \int_T \exp \left( \mu(t) + \sum_{k=1}^{K} \theta_k \varphi_k(t) \right) dt \right)$ is the normalizing constant. Family $P_K$ is identifiable as will be shown in Proposition S4.1 in the supplementary materials. The random density $p$ can then be approached by the truncated version $p_K \in P_K$, defined by

\[
p_K(t) \propto \exp(\mu(t) + \sum_{k=K+1}^{\infty} \eta_k \varphi_k(t)), \quad t \in T. \tag{2.5}
\]

The model components in (2.4) enjoy classical interpretation within an exponential family: $\theta_k$ is the natural parameter which compactly describes the density, the eigenfunctions $\varphi_k$ is the sufficient statistic, and mean $\mu$ is the baseline measure. Analogous to the case for FPCA, the leading eigenfunctions $\varphi_k$, $k = 1, \ldots, K$ encode the principal modes of variation in the transformed densities, and the $\eta_k$ are the scores explaining the most variation; $P_K$ provides the most parsimonious $K$-dimensional description of the random (log-)densities. Hence, performing density estimation within $P_K$ will effectively borrow information from the typical shapes of the densities determined by the leading eigenfunctions or the form of sufficient statistic.

#### 2.2.2. Fitting Densities within the Approximating Family

Given a small sample $X_{0j}$, $j = 1, \ldots, N$ on $T$ from a new subpopulation, an estimate for the underlying density $p_0$ is then obtained as a best fit within the approximating family $P_K$. To assess the goodness of fit, various approaches are available from information theory (e.g., Amari and Nagaoka 2000; Amari 2016), and we adopt the information loss as quantified by the Kullback–Leibler (KL) divergence, which is defined for two positive densities $q_1$ and $q_2$ supported on $T$ as

\[
D(q_1 || q_2) = \int_T q_1(t) \log \frac{q_1(t)}{q_2(t)} \, dt. \tag{2.6}
\]

The KL divergence $D(q_1 || q_2)$ from $q_1$ to $q_2$ quantifies the information loss when one approximates $q_1$ by $q_2$. To estimate $p_0$, we propose to use the best approximating element in $P_K$ minimizing the KL divergence from the empirical distribution. This leads to the maximum likelihood estimate (MLE) defined by $p_\hat{\theta} \in P_K$, where

\[
\hat{\theta} = \text{argmax}_{\theta : B_K(\theta) < \infty} \left( \theta^T \bar{\varphi}_0 - B_K(\theta) \right), \tag{2.7}
\]

and $\bar{\varphi}_0 \in \mathbb{R}^K$ is the sufficient statistic constructed from the sample $X_{0j}$, in which the $k$th element is $N^{-1} \sum_{j=1}^{N} \varphi_k(X_{0j}), k = 1, \ldots, K$. 

2.2.3. Incorporating Population-Level Information

In addition to the typical shapes of random densities, we borrow information from the distribution of these random densities in the population to reduce variation when fitting a new density. This will result in a significant improvement especially if the sample size in the new subpopulation is small.

The distribution of $p$ is captured by those of the component scores $\eta_k$, $k = 1, \ldots, K$. Incorporating such distributional information into the likelihood function, we propose maximum a posteriori (MAP) estimate $\hat{p}_{\text{MAP}} = p_{\theta_{\text{MAP}}} \in \mathcal{P}_K$, where the parameter estimate $\hat{\theta}_{\text{MAP}}$ is defined as

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta : B_K(\theta) < \infty} \left( \theta^T \hat{\varphi}_0 - B_K(\theta) + N^{-1} \log \pi(\theta) \right), \quad (2.8)$$

and $\pi$ is the unconditional density of the $\eta_k$. Analogous to the PACE procedure (Yao, Müller, and Wang 2005), $\pi$ is constructed using independent marginal distributions being normal distributions with mean $E \eta_k = 0$ and variance $\text{var}(\eta_k) = \lambda_k$, $k = 1, \ldots, K$. Here $\hat{\theta}_{\text{MAP}}$ is a shrinkage of the MLE $\hat{\theta}$ toward the population distribution $\pi$ of the natural parameter.

To accommodate nonnormal unconditional distributions for the $\eta_k$, we propose a second approach to incorporate the population-level knowledge by working with the moment parameterization. The moment parameter, also referred to as the one-to-one mappings derived from (2.9).

$$\xi = \partial_0 B_K(\theta). \quad (2.9)$$

For brevity, we use $\xi(\cdot) : \theta \mapsto \xi(\theta)$ and $\theta(\cdot) : \xi \mapsto \theta(\xi)$ to denote the one-to-one correspondence with natural parameters (see, e.g., Amari 2016), satisfying $\hat{\xi} = \partial_0 B_K(\theta)$.

Motivated by the fact that the MLE of the natural parameter is the method of moments estimate under moment parameterization, we consider shrinkage estimate based on the moment parameterization. Let $\eta_0$ be the natural parameter of the truncated random densities $p_{0,K} \in \mathcal{P}_K$ of $p_0$ according to (2.3) and $\tau_0 = \xi(\eta_0) \in \mathbb{R}^K$ the corresponding moment coordinate. We propose to estimate the moment parameter $\tau_0$ by the best linear unbiased predictor (BLUP) of $\tau_0$ given the sample sufficient statistic $\hat{\varphi}_0$, defined as

$$\hat{\eta}_{\text{BLUP}} = \text{BLUP}(\tau_0 | \hat{\varphi}_0) \triangleq \text{cov}(\tau_0, \hat{\varphi}_0)\text{var}(\hat{\varphi}_0)^{-1}(\hat{\varphi}_0 - E \hat{\varphi}_0) + E \tau_0 = \text{var}(\tau_0)\text{var}(\hat{\varphi}_0)^{-1}(\hat{\varphi}_0 - E \tau_0) + E \tau_0, \quad (2.10)$$

where the last equality is due to $E(\hat{\varphi}_0 | \tau_0) = \tau_0$ and double expectation. Note that here the covariance and expectation are computed under the joint distribution of $(\tau_0, \hat{\varphi}_0)$. We thus obtain a plug-in estimate $\hat{\theta}_{\text{BLUP}} = \theta(\hat{\eta}_{\text{BLUP}})$ of the natural parameter.

The MAP (2.8) and BLUP (2.10) estimates not only borrow typical shapes of densities via the sufficient statistic $\varphi$ analogous to the MLE (2.7), but also take advantage of the distribution of the random density through the component scores. We refer to the information embedded in the component scores of log-densities as population-level information. The MAP and BLUP estimates are, respectively, a shrinkage of the MLE toward the mean under the natural and the moment coordinate. As will be demonstrated in our numerical investigation later, the shrinkage is crucial for reducing information loss especially given small samples.

2.2.4. Pre-smoothing and Unknown Distribution of $p$

Oftentimes, the random densities $p_1, \ldots, p_n$ and their distribution are unavailable but only discrete samples $\{X_{ij} : j = 1, \ldots, N_j; i = 1, \ldots, n\}$ are accessible. In this case, we use pre-smoothing to construct pilot density estimates for the samples, analogous to the approach taken by Petersen and Müller (2016) and Han, Müller, and Park (2020). Many density estimators can be applied, for example kernel density estimate (KDE) and logspline (Kooperberg and Stone 1991), and for theory we consider the KDE

$$\hat{p}_i(t) = \left( \int_{\mathcal{X}} \sum_{j=1}^{N_i} \frac{t - X_{ij}}{h} \, dx \right)^{-1} \sum_{j=1}^{N_i} \frac{t - X_{ij}}{h}, \quad t \in \mathcal{T}, \quad (2.11)$$

where $\kappa$ is a density function. In the pre-smoothing step, the pilot density estimator is applied on each sample to obtain pre-smoothed densities $\hat{p}_1, \ldots, \hat{p}_n$.

FPCA is then performed on the centered log-ratio transformed densities $\tilde{f}_i = \psi \hat{p}_i$, $i = 1, \ldots, n$ to obtain sample mean $\hat{\mu}(t) = n^{-1} \sum_{i=1}^n \hat{f}_i(t)$, sample covariance function $\hat{G}(s, t) = n^{-1} \sum_{i=1}^n ([\hat{f}_i(t) - \hat{\mu}(t)] [\hat{f}_i(s) - \hat{\mu}(s)])$, the associated eigenvalues $\lambda_k$ and eigenfunctions $\hat{\varphi}_k$, and the component scores $\hat{\eta}_k \in \mathbb{R}^K$ in which the $k$th element is $\hat{\eta}_k = \{\tilde{f}_i - \hat{\mu}, \hat{\varphi}_k\}$, for $k = 1, \ldots, n-1$. Note that $\int_{\mathcal{T}} \hat{\mu} = \mu$ and $\int_{\mathcal{T}} \hat{\varphi}_k = 0$, so applying the inverse transformation $\psi^{-1}$ leads to a sample version of the approximating families $\hat{\mathcal{P}}_K$ as

$$\hat{\mathcal{P}}_K = \left\{ \hat{p}_0 : \theta \in \mathbb{R}^K, \hat{B}_K(\theta) < \infty \right\}, \quad \text{with} \quad (2.12)$$

$$\hat{p}_0(t) = \exp \left( \hat{\mu}(t) + \sum_{k=1}^K \theta_k \hat{\varphi}_k(t) - \hat{B}_K(\theta) \right), \quad t \in \mathcal{T}, \quad (2.13)$$

where $\hat{B}_K(\theta) = \log \left( \int_{\mathcal{T}} \exp \left( \hat{\mu}(t) + \sum_{k \leq K} \theta_k \hat{\varphi}_k(t) dt \right) \right)$ is the normalizing constant. Inference for new samples is conducted based on $\hat{\mathcal{P}}_K$ with sufficient statistic $\hat{\varphi}_0 \text{ave} \in \mathbb{R}^K$, in which the $k$th element is $N^{-1} \sum_{i=1}^n \hat{\varphi}_k(X_{ij})$; to avoid double over-scripts here we use subscript "ave" to denote the average. The distribution of random densities can be accessed via the empirical distribution of the sample component scores. The proposed estimators in full sample-based versions are then defined using the estimated quantities. The MLE and MAP within $\hat{\mathcal{P}}_K$ are $\hat{\theta}_{\text{MLE}}$ and $\hat{\theta}_{\text{MAP}}$, respectively, where

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta : \hat{B}_K(\theta) < \infty} \left( \theta^T \hat{\varphi}_0 \text{ave} - \hat{B}_K(\theta) \right), \quad (2.14)$$

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta : \hat{B}_K(\theta) < \infty} \left( \theta^T \hat{\varphi}_0 \text{ave} - \hat{B}_K(\theta) + N^{-1} \log \hat{\pi}(\theta) \right), \quad (2.15)$$

in parallel to (2.7) and (2.8), and $\hat{\pi}$ is the product of normal densities with mean 0 and variance $\hat{\sigma}^2_{\eta_k}$, the sample variance of $(\hat{\eta}_k, i = 1, \ldots, n)$, for $k = 1, \ldots, K$. 


For BLUP, $E_{T_0}$ and $\text{var}(\tau_0)$ in (2.10) are estimated by the sample mean $\bar{\tau}_{\text{ave}} = n^{-1} \sum_{i=1}^{n} \tau_i$ and sample covariance matrix $\Sigma_{\bar{T}}$ of moment coordinates $\{\bar{\tau}_i = \xi(\bar{\eta}_i) : i = 1, \ldots, n\}$ in $\tilde{\mathcal{P}}_K$. Further decompose $\text{var}(\tilde{\psi}_0) = E \text{var}(\tilde{\psi}_0 | \tau_0) + \text{var}(\tau_0)$, and estimate the first term by

$$\Sigma_{\bar{\psi}_{\text{ave}}} = \frac{1}{nN} \sum_{i=1}^{n} \int_{T} (\tilde{\psi}(t) - \bar{\tau}_i) (\tilde{\psi}(t) - \bar{\tau}_i)^T \tilde{p}_i(t) dt,$$

where $\tilde{\psi} : t \mapsto (\tilde{\psi}_1(t), \ldots, \tilde{\psi}_K(t)) \in \mathbb{R}^K$ is the sufficient statistic. In combine, the plug-in BLUP estimate for the moment parameter within $\tilde{\mathcal{P}}_K$ is obtained as

$$\hat{\xi}_{\text{BLUP}} = \Sigma_{\bar{T}} (\Sigma_{\bar{\psi}_{\text{ave}}} + \Sigma_{\bar{T}})^{-1} (\bar{\psi}_{0,\text{ave}} - \bar{\tau}_{\text{ave}}) + \bar{\tau}_{\text{ave}}, \quad (2.16)$$

and the associated natural parameter $\hat{\theta}_{\text{BLUP}} = \partial_\theta \tilde{B}_K(\theta)$ is the solution to $\hat{\xi}_{\text{BLUP}} = \partial_\theta \tilde{B}_K(\theta)$.

When the random densities $p_1, \ldots, p_n$ are observed in their entirety, a scenario considered for the interest of theory and referred to as the subpopulations being completely observed, it suffices to skip the pre-smoothing step and proceed with the observed densities. In this case, FPCA is performed on the transformed trajectories $f_i = \psi p_i$, $i = 1, \ldots, n$ to obtain sample mean and covariance as $\hat{\mu}_i(t) = n^{-1} \sum_{i=1}^{n} f_i(t)$ and $\hat{G}_i(s, t) = n^{-1} \sum_{i=1}^{n} (f_i(s) - \hat{\mu}_i(s))(f_i(t) - \hat{\mu}_i(t))$, eigenvalues and eigenfunctions $\hat{\lambda}_k$ and $\hat{\varphi}_k$, and sample component scores $\hat{\eta}_{ik} = [f_i - \hat{\mu}_i, \hat{\varphi}_k]$, $k = 1, \ldots, n - 1$. The approximating family $\tilde{\mathcal{P}}_K$ with completely observed densities is defined as

$$\tilde{\mathcal{P}}_K = \{\hat{\rho}_0 : \theta \in \mathbb{R}^K, \tilde{B}_K(\theta) < \infty\}, \quad (2.17)$$

$$\hat{\rho}_0(t) = \exp \left( \hat{\mu}_i(t) + \sum_{k=1}^{K} \theta_k \hat{\varphi}_k(t) - \tilde{B}_K(\theta) \right), \quad t \in T, \quad (2.18)$$

for $K = 1, 2, \ldots$, where $\tilde{B}_K(\theta) = \log \int_{T} \exp (\hat{\mu}(t) + \sum_{k=1}^{K} \theta_k \hat{\varphi}_k(t)) dt$ is the normalizing constant. The proposed MLE, MAP, and BLUP are then defined analogously to those within $\tilde{\mathcal{P}}_K$. We use tilde overscripts to denote estimated quantities constructed with fully observed densities.

We refer to the process of constructing approximating exponential families from the discrete observations from $p_1, \ldots, p_n$ as the training step, which emphasizes that the $K$-dimensional families $\tilde{\mathcal{P}}_K$ are flexibly derived from the data despite themselves being parametric families compactly described by a few natural parameters. The random densities $p_1, \ldots, p_n$ are referred to as the training densities associated with the training subpopulations, and the corresponding samples as training samples. To assess performance, we call the process constructing density estimates from new observations within the approximating families the fitting or testing step; correspondingly, $p_0$ is the density of a testing subpopulation, from which the discrete sample is referred to as a testing sample.

The number of components $K$ controls the flexibility of $\tilde{\mathcal{P}}_K$ and is the key tuning parameter trading off bias and variance when fitting a density. We propose to select $K$ using a penalized approach by considering the Akaike information criterion (AIC), and set $K = K^*$ that minimizes $\text{AIC}(K) \triangleq 2K - 2 \log \hat{\rho}_K$, where $\hat{\rho}_K$ is the likelihood under the estimates constructed using either (2.14), (2.15), or (2.16). In practice, we observe that the optimal $K^*$ often increases as the fitting sample size increases. Intuitively, the proposed component selection automatically uses a more flexible model given a larger sample, in which case the additional variance that comes with the flexibility becomes more affordable.

The proposed procedure is computationally efficient. The proposed MLE, MAP, and BLUP estimates are performed within exponential families, so the optimization tasks in obtaining (2.14)–(2.16) are convex and thus easy to solve. In addition, the discrete samples are succinctly summarized using the sufficient statistics.

### 3. Simulation Studies

We study the numerical properties for the proposed methods and alternative parametric and non-parametric density estimators under various scenarios. Independent training samples $\{X_{ij} : j = 1, \ldots, N_i\}$ were drawn from densities $p_i$, $i = 1, \ldots, n$, where $p_i$ are independent copies of a random density $p$. Independent testing samples $\{X_{ij}^* : j = 1, \ldots, N_i^*\}$ were then drawn from additional independent copies $p_i^*$ of $p$ for $l = 1, \ldots, n^*$. Typically, training sample sizes were larger, for example, $N_i = 200$, while testing samples were smaller, for example, $N_i^* = 10$. The small sample sizes in the testing samples pose a challenge for the density estimators compared. The underlying family $\mathcal{P}$ and the distribution of $p \in \mathcal{P}$ varied between scenarios. The experiment under each scenario was repeated 500 times.

The proposed methods (2.14), (2.15), and (2.16) are referred to in the comparison as FPCA_MLE, FPCA_MAP, and FPCA_BLUP. Our proposals were compared against the maximum likelihood estimate within the ground truth family $\mathcal{P}$ as an idealized parametric approach, referred to as MLE; non-parametric methods including KDE (Sheather and Jones 1991), logspline (Kooperberg and Stone 1991), and locfit (Loader 1996), where the latter two approximates the log-densities with splines and local polynomials, respectively; and a repeated point processes (RPP) approach proposed by Gervini and Khanal (2019). These methods do not require training.

The mean KL divergence for the density estimates was used to assess the overall estimation performance. For density estimate $\hat{\rho}_i^*$ for the $l$th testing sample, information loss was evaluated by KL divergence $D(\hat{\rho}_i^* || \tilde{\rho}_i^*)$ defined in (2.6), and we examined the average

$$\text{MKL} = \frac{1}{n^*} \sum_{l=1}^{n^*} D(\hat{\rho}_l^* || \tilde{\rho}_l^*)$$

over $n^*$ testing samples to quantify performance. Smaller MKL indicates better estimation performance. Results for an alternative error metric, the mean integrated squared errors (MISE), and additional simulation cases are included in the supplementary materials.

#### 3.1. Flexible Exponential Families

We first set the density families to be either the truncated normal or bimodal distributions. For the truncated normal scenario, each sample was generated from $N(\mu, \sigma^2)$ truncated on $[-3, 3]$
with $\mu \sim \text{Unif}(-2, 2)$ and $\sigma \sim \text{Unif}(2, 4)$. For the bimodal scenario, each sample was generated according to density $p(x) \propto \exp((4 + \theta)x - (26.5 + \theta)x^2 + 47x^3 - 25x^4), x \in [0, 1]$ with $\theta \sim \text{Unif}(0, 10)$. In each Monte Carlo experiment, $n = 50$ training samples of size $N_i = 200$ each and $n^* = 100$ testing samples with sizes $N = 10$ or 50 were created.

The mean KL divergence in Figure 3.1 shows that the proposed FPCA_MLE using either the KDE as pre-smoother or with the completely observed training densities both produced estimates comparable to the MLE, but without requiring explicit specification of a parametric family. Among all the non-parametric density estimators, the proposed methods are shown to be stable and best-performing, especially when the testing samples are small, in which case the proposed methods outperformed MLE by borrowing strength among different samples, whereas MLE does not pool information. In particular, for truncated normal family with $N^* = 10$, on average the information loss of the proposed MAP and BLUP were around 20% less compared to that of MLE within ground truth family, and more than 40% less compared to that of KDE. Moreover, the AIC-driven number of components $K^*$ is overall appropriate, as shown in Figure S2.2 in the supplementary materials.

### 3.2. Gaussian Mixture Family

Contrary to the previous scenario where the underlying families were exponential families, we next estimate densities from a family consisting of mixtures of Gaussian distributions, which forms a multi-modal nonexponential family. This is to further demonstrate the flexibility of the proposed methods to a “mis-specified” case and also demonstrate the data-adaptive selection of the number of components.

The samples were generated following a mixture of three Gaussian distributions with density $p(x) \propto \sum_{l=1}^{3} \theta_l \phi((x - \mu_l)/\sigma_l)$ truncated to $x \in [-3, 3]$, where $\phi$ is the density of $N(0, 1)$, $(\theta_1, \theta_2, \theta_3) \sim \text{Dirichlet}(1/3, 1/3, 1/3)$, $\mu_l \sim \text{Unif}(-5, 5)$, and $\sigma_l \sim \text{Unif}(0.5, 5)$, for $l = 1, 2, 3$. Here, instead of performing a direct MLE which is known to be unstable for mixture distributions, the expectation-maximization (EM)
algorithm in the variant of Lee and Scott (2012) is used as the baseline parametric density estimator.

The left panel of Figure 3.2 compares the relative error of various density estimates to that of the EM algorithm as a baseline reference. The proposed methods enjoyed the least information loss for all but the larger sample size when EM method became the best. When the testing sample size $N^* = 25$, on average, the information loss of proposed MAP and BLUP were 20% less than that of RPP, and all at least five times better compared to the EM algorithm. Remarkably, this is achieved by using only the first few components, on average, as demonstrated in the right panel of Figure 3.2. This shows that the proposed methods are effective alternatives to the EM algorithm, and overcome the local maximum issue by performing convex optimization within approximating exponential families. In addition, by using the AIC-driven component selection, the proposed estimates used more components as the size of fitting sample increased, suggesting that the AIC selection adapts to the increased complexity of the underlying family as unveiled by a larger sample. Indeed, the larger the sample size, the more representative the sample will be for the underlying distribution, which has eight free parameters in our setup. By using more components, the proposed methods can accommodate the increased complexity revealed by the data such as multi-modality, which may otherwise be hidden when fitting with fewer components.

4. Real Data Applications
4.1. MIMIC Data

We consider the age-at-admission distributions of patients with critical conditions in the Medical Information Mart for Intensive Care (MIMIC) database (Johnson et al. 2016), collected from patients admitted to critical care units at a large hospital in Boston between 2001 and 2012. The number of patients under different primary diagnosis are highly different. Ample data points are available to establish a precise non-parametric density estimate for common diseases, as shown in the left panels of Figure 4.1, whereas observations for uncommon conditions are scarce and cannot afford a flexible nonparametric approach such as KDE, as displayed in right panels of Figure 4.1. Estimating the densities in the hard-to-observe subpopulations of patients with uncommon conditions is the emphasis in our analysis. Intuitively, different diseases, even the rare ones, may share common underlying age-related risk factors, so that we may borrow information from common diseases to model rare ones and the proposed methods may succeed.

Sex and diagnosis code (ICD9) combinations were used to define the subpopulations, where we excluded birth-related, unclassified, or unspecified diagnosis. A total of 940 subpopulations with more than five observations each were analyzed, among which 75 subpopulations with at least 75 observations were used for training while the rest for testing. For example, the upper left panel in Figure 4.1 shows a training sample of 109 female patients with juvenile type diabetes with ketoacidosis, while the bottom right panel shows a testing sample consisting of 27 male patients with abdominal aortic aneurysm rupture. We focused on estimating the densities for age range $T = [14, 85]$. KDE was used as the pre-smother and the bandwidth was set to the median of the bandwidths in the training samples selected by the method of Sheather and Jones (1991). Patients in the training subpopulations outside of the age range of interest were included in pre-smoothing for better boundary performance. As illustrated in right panels of Figure 4.1, the proposed density estimates produced sensible results given small test samples, which were smoother than the estimates given by KDE and avoided multiple spurious modes.
Figure 4.1. The age-at-admission distribution of patients with either sex and a different primary diagnosis, each combination defining a subpopulation. Four training subpopulations with ample observations and four testing ones with scarce observations are shown in the left and the right panels, respectively. Along with the kernel density estimate (KDE), the proposed density estimates (MLE, MAP, and BLUP) are displayed for the testing subpopulations.

The modes of variation in the density functions obtained by varying each of the first three natural parameters of \( \hat{P}_K \) are shown in the left panels of Figure 4.2. The first three modes of variations correspond to a shift in the age distribution toward the old, the elderly, and the middle age, respectively. This suggests the existence of latent disease characteristics that affect the age profiles of different conditions.

The effects of shrinkage in the proposed MAP and BLUP estimators are demonstrated in the right panels of Figure 4.2 for the four rare conditions displayed in the right panels of Figure 4.1.
Table 1. Cross-entropy (4.1) of the proposed estimators (MLE, MAP, and BLUP) and the KDE, applied to the MIMIC data.

| Estimator | 5 ≤ N ≤ 10 | 10 < N ≤ 35 | 35 < N ≤ 75 |
|-----------|-------------|-------------|-------------|
|           | n* Mean | SD | n* Mean | SD | n* Mean | SD |
| MLE       | 423   | 4.60 | 1.62 | 343   | 4.03 | 0.32 | 78   | 3.93 | 0.22 |
| MAP       | 4.13  | 0.42 | 3.97 | 0.25 | 3.92 | 0.21 |
| BLUP      | 4.11  | 0.36 | 3.97 | 0.24 | 3.92 | 0.21 |
| KDE       | 402   | 5.56 | 2.42 | 341   | 4.16 | 0.37 | 78   | 3.99 | 0.22 |

NOTE: The results are stratified by size N of testing samples, where the number of testing samples involved is listed as n*.

Subpopulations with smaller sample sizes were typically shrunk more toward the population mean using the MAP and BLUP methods that pool population information from the training samples, in which case the estimate differed more notably from the MLE.

Visually, the proposed methods outperformed KDE as in Figure 4.1 and on artificially sparsified samples, detailed in Section S1.2 of the supplementary materials. To numerically assess the density estimates, we evaluated the cross-entropy instead of the KL divergence to assess information loss, since the underlying densities are unknown. Given any density estimator \( \hat{p} \), we approached the cross-entropy \( H(p, \hat{p}) = -E_p \log \hat{p} \), which is the quantity in \( D(p|\hat{p}) \) that depends on the estimate \( \hat{p} \), by an unbiased leave-one-out estimate

\[
H_{LOO}(p, \hat{p}) \triangleq \frac{1}{N} \sum_{j=1}^{N} \log \hat{p}_{-j}(X_j),
\]

(4.1)

where \( \hat{p}_{-j} \) is a density estimate constructed with all but the jth observations. A summary of cross-entropy applied on 844 testing samples using the proposed methods and KDE are reported in Table 1, where a smaller cross-entropy indicates a better estimate. Occasionally, KDE resulted in non-finite cross entropy, leading to a smaller number of samples reported. The results show that the proposed methods produced density estimates that better reflect sample information compared to KDE by having smaller mean cross-entropy, and also more stable performance with smaller standard deviations.

4.2. Precipitation Return Level Estimation

We then analyze the precipitation data in the Haihe region to illustrate the versatility of the proposed methods in providing good estimates for both the body of the density function and the tail probability. We will demonstrate that our proposed methods can borrow information from the stations with longer records to construct approximating distribution families to model annual maxima of hourly rainfall, which lead to better estimates of densities and return levels for those stations having shorter records.

Predicting the likelihood of extreme precipitation is of great importance because of the severe damage such events may cause to economy and social life. We focus on heavy rainfall during short time period. The likelihood of extreme precipitation event in a certain region can be quantified using the \( T \)-year return level, which is defined as the level expected to be exceeded once every \( T \) years, or equivalently the \( 1 - 1/T \) quantile of the marginal distribution under stationarity and independence assumptions. Sample quantile is a straight forward estimate for a small \( T \), which is referred to as the empirical return level, but this approach requires more than \( T \) data points being available; otherwise, distributional models are necessary to estimate the return level for a longer return period.

We focus on estimating the 5- to 30-year return levels of the annual maximum hourly rainfall at various locations in the Haihe River basin, a vital region in northern China that includes mega-cities such as Beijing and Tianjin. Precipitation data from 232 weather stations around the Haihe River basin from 1961 to 2012 used in this analysis are available from the National Meteorological Information Center (NMIC) and the China Meteorological Administration (CMA). All the 232 stations investigated had at least 30 years of records, and each station defines a subpopulation. As the precipitation distributions are heavily right-skewed, the pre-smoothing density estimates could be zero within the working domain, making it impossible to compute the centered log-ratios. We therefore work with the log-precipitation; see Section S1.4 of the supplementary materials for detail.

We compared return levels estimation and the information loss of the proposed methods with standard approaches. To imitate the situation of newer weather stations having shorter records, we randomly selected \( n = 50 \) stations as training sites and retained the full records, and another \( n^* = 100 \) testing sites where records from randomly chosen 10 consecutive years were used for fitting. Estimates of the density and 5- to 30-year return levels were obtained from nonparametric approaches including our proposed methods and standard KDE, as well as parametric approaches of maximum likelihood estimation within the families of Gamma distribution and generalized extreme value (GEV) distribution, of which the density is \( p(t|\theta) = \theta_2^{-1} \omega(t)^{\theta_1+1} e^{-\omega(t)} \) where \( \omega(t) = \exp(-(t - \theta_1)/\theta_2) \) if \( \theta_2 \neq 0 \), and \( \omega(t) = (1 + \theta_3(t - \theta_1)/\theta_2)^{-1/\theta_3} \) otherwise. Cross-entropy (4.1) was used to evaluate the overall fit of the density estimates. For return level estimates, we used the relative difference of the estimates to the empirical return levels that are obtained from full records in the testing stations, which measures the fit in the right tail. The process was repeated 10 times, and the overall average cross-entropies and the relative differences of the return levels averaged over the stations are reported in Figure 4.3. KDE and logKDE were excluded from return the level prediction since they cannot reliably estimate the tail behavior.

The proposed MLE, MAP, and BLUP methods were all significantly better than the KDE and GEV approaches under the information loss. The estimates using Gamma family were comparable to our results but underperformed the MAP and BLUP methods which borrow information from the population. For return level estimation, the proposed MAP and BLUP methods had the least variation compared to MLE, Gamma, and GEV methods, where the last three methods do not pool information. For shorter (5- and 10-year) return levels, the bias of all methods were comparable, but for longer (20- and 30-year) return levels, the bias of our methods were significantly smaller than the Gamma and GEV methods. In summary, our proposed methods had the best performance in estimating the return levels as well. Hence, the proposed methods provide better fits compared to classical approaches in both the bulk and the tail parts of
the distributions and lead to better quantification of interesting characteristics of the distribution such as return levels, using just a small number of discrete observations.

These numerical results come as a positive surprise, as we only used 10 years of records at the testing sites and yet were able to obtain good estimates for 20- and even 30-years return levels. The proposed approaches are able to estimate the tail structure from the data rather than relying on assumptions, and borrow information from the population enforced through parameter shrinkage.

5. Theoretical Results

We establish theoretical results for the proposed methods including the identifiability of the approximating families, consistency of the MLE within those families, and consistency of the component scores. While the main motivation and justification for the proposed methods are the small-sample performance as demonstrated in the simulations and real data applications, in this section we also establish large-sample properties to justify the asymptotic validity. Sections 5.1 and 5.2 investigate the large testing sample scenario where the testing sample size \( n \), the number of training samples \( N_t \), and the training sample sizes \( N_i \) all diverge, while Section 5.3 considers the small testing sample case when \( N \) is small and kept fixed but \( n \) and \( N_i \) diverge. We focus on the MLE for theoretical discussions, since for a large testing sample the shrinkage effects of the BLUP and MAP estimates would be negligible, making them similar to the MLE. Proofs and additional discussions are included in the supplementary materials.

For theoretical development, densities in the underlying family are required to be smooth and uniformly bounded away from zero and infinity. To handle the boundary issue in the pre-smoothing step, we focus on the domain \( T \) but assume that any density of interest \( q \in \mathcal{P} \) is the truncated version of a density \( q_e \in \mathcal{P}_e \) with a slightly enlarged support than \( T \), satisfying \( q(t) = q_e(t)/\int_T q_e(s)ds \) for \( t \in T \). Denoting \( p_{ci} \) as the extended version of \( p_i \), we have \( \psi p_{ci} = \psi p_i \) on \( T \), so the approximating families will be identical no matter whether they are constructed with \( p_i \) or \( p_{ci} \); a similar identity holds for the pre-smoothed versions. For theory, we assume that the training data \( X_{ij} \) are generated from the extended random densities \( p_{ci} \). The same pre-smoother \( \hat{p}_i \) as in (2.11) will be applied, but the observations are now assumed to be drawn from \( p_{ci} \); all other modeling is identical as described in Section 2.

(A1) All densities \( q_e \in \mathcal{P}_e \) are twice differentiable and strictly positive on the support \( T_e := \{ s : \inf_{t \in T} |s - t| \leq \varepsilon_T \} \).

Moreover, there is a constant \( M > 1 \) such that, for all \( q_e \in \mathcal{P}_e \), \( \sup_{t \in T_e} \max(|q_e(t)|, |1/q_e(t)|, |q_e''(t)|) \leq M \).

Under (A1), the approximating families \( \mathcal{P}_K, \mathcal{P}_K^*, \) and \( \mathcal{P}_K \) as defined in (2.3), (2.17), and (2.12) are identifiable, thanks to the fact that the entries of the sufficient statistic are orthogonal to each other and to the constant function, as shown in Proposition S4.1 in the supplementary materials.

5.1. Consistency of MLE as \( N \) Diverges

The proposed MLE within low-dimensional approximating family \( \mathcal{P}_K^* \) are shown to be consistent. The MLE is computed with an additional iid sample \( X_1, \ldots, X_N \) of size \( N \), whose density \( p_0 \) supported on \( T \) is an independent and identical copy.
of the training densities \( p_i \). The approximating family \( \hat{P}_K \) is obtained from \( n \) discretely observed training populations each with \( N_i \geq m \) observations. Theoretical results for this practical scenario are presented here, and results for MLE obtained within \( P_K \) and \( \hat{P}_K \) are included in the supplementary materials. For brevity, we use \( \hat{p} = \hat{p}_{N,K} \in \hat{P}_K \) to denote the MLE density estimate with parameter \( \hat{\theta} = \hat{\theta}_{N,K} \) defined in (2.14), and use \( \hat{p} \) instead of \( p_0 \). The consistency result relies on the following conditions imposed on the covariance of the transformed random density.

(C1) For process \( f = \psi p \in L^2(T) \), the eigenvalues and eigenfunctions of the integral operator \( G \) associated with the covariance function \( G \) satisfy, as \( J \to \infty \),

\[
\left( \sum_{k=1}^{J} \| \psi_k \|_\infty \right) \left( \sum_{k=j}^{J} \lambda_k \right) = O(1),
\]

\[
\left( \sum_{k=1}^{J} \| \psi_k \|_\infty \right)^{2/3} \left( \sum_{k=j}^{J} \lambda_k \right) = o(1).
\]

(C2) For all \( k = 1, \ldots, \), rank \( G \), the nonzero eigenvalues are distinct; the estimated eigenfunctions are properly aligned so that \( \langle \psi_k, \tilde{\psi}_k \rangle \geq 0 \) and \( \langle \psi_k, \psi_k \rangle \geq 0 \).

Condition (C1) ensures that the trajectories are well-represented by smooth finite approximations by assuming suitably fast eigenvalue decays and smooth eigenfunctions. The integrated Brownian motion satisfies (C1) whereas the Brownian motion does not, as shown in Example S3.1 in the supplementary materials. Condition (C2) is standard for simplifying presentation. We assume rank \( G = \infty \) in discussion hereafter unless otherwise stated.

Conditions (A1), (C1), and (C2) in combine are sufficient for the consistency of the MLE within \( P_K \) and \( \hat{P}_K \), as shown in Theorem S6.2 of the Supplementary Materials. When discrete training samples, and sequences \( \{a_n\}, \{b_n\} \) are defined in (5.1) and (5.2).

\begin{align*}
A_K \sqrt{K/N} \to 0, & \quad KA_K (1/\sqrt{n} + b_n) \to 0, \\
\sum_{k \leq K} \delta_k^{-1} (1/\sqrt{n} + a_n) = O(1), & \quad 0 < K \leq K,
\end{align*}

where

\[
A_K = 2 \max \left( |T|^{-1/2}, \left( \sum_{k=1}^{K} ||\psi_k||_\infty \right)^{1/3} \right),
\]

then

\[
D(p||\hat{P}_{N,K}) = O_p \left( \left( \frac{1}{n} + b_n \right) \left( K + \sum_{k \leq K} \delta_k^{-1} \right)^2 + \sum_{k > K} \lambda_k + \frac{K}{N} \right).
\]

\begin{proof}
See Theorem S6.2 and Proposition S5.3 in the supplementary materials.
\end{proof}

Note that the result is highly flexible and suits any general infinite-dimensional families. In particular, we do not assume the underlying \( P \) to be a finite-dimensional exponential family. The three summands in the rate correspond to the training error, a bias term due to the approximating family, and a variance term from fitting with \( N \) observations, respectively. Varying the number of components \( K \) trades off errors from these three terms, as a more complex model with a larger \( K \) will have larger training and fitting errors but smaller approximation bias. Quantity \( A_K \) is closely related to (C1) and the equivalence of \( L^\infty \) and \( L^2 \) norm in the eigenspaces of covariance operator; see Proposition S4.2 of the supplementary materials.

An immediate corollary is the convergence of the densities in the \( L^1 \) norm by Pinsker’s inequality (Wainwright 2019), which states that the KL divergence dominates squares of \( L^1 \) norm of differences.
Corollary 5.1. Under the conditions of Theorem 5.1,
\[
    ||\hat{p}_{N,K} - p||^2_1 = O_p\left(\left(\frac{1}{n} + b_n^2\right)\left(K + \sum_{k \leq K} \delta_k^{-1}\right)^2 + \sum_{k > K} \lambda_k + \frac{K}{N}\right).
\]

Our general result directly implies the next corollary if the underlying family is actually a $K_0$-dimensional exponential family for some finite $K_0$, with no approximation bias resulting from truncating the log-densities. Note that (C1) is not required, and by convention we set $\lambda_k, \varphi_k$ to be zero for $k > K_0$.

Corollary 5.2. Under (A1), (K1), (C2), (C3), if $\mathcal{P}$ is a $K_0$-dimensional exponential family for some finite $K_0$, then for any fixed $K \geq K_0$,
\[
    D(p||\hat{p}_{N,K}) = O_p\left(h^4 + \frac{1}{n} + \frac{1}{m^2 h^2} + \frac{1}{N}\right).
\]

Proof. See Corollary S6.1 in the supplementary materials.

5.1.2. Selecting $K$

The following proposition shows an example where the requirements in Theorem 5.1 are satisfied. For sequences $x_n$ and $y_n$, write $x_n \succ y_n$ if $x_n = O(y_n)$ and $y_n = O(x_n)$.

Proposition 5.1. Suppose that (A1), (K1), (C2), and (C3) hold, and that

1. eigenfunctions coincide with the polynomial or trigonometric basis, with $||\varphi_k||_\infty \asymp k$;
2. eigenvalues have polynomial decay: $\lambda_k \asymp k^{-r}$ with $r \geq 3$ as $k \to \infty$; and
3. the number of components $K$ used for fitting satisfies $K = \tilde{K}^* \asymp \min \left(\frac{mh}{\log n}, \left(h^4 + \frac{1}{n} + \frac{1}{m^2 h^2}\right)^{-1/(4r+4)}, N^{1/r}\right)$.

Then the conditions for Theorem 5.1 are satisfied, implying
\[
    D(p||\hat{p}_{N,K^*}) = O_p\left(h^4 + \frac{1}{n} + \frac{1}{m^2 h^2}\right)^{(r-1)/(4r+4)} + \left(h^4 + \frac{1}{n} + \frac{1}{m^2 h^2}\right)^{(r-1)/(4r+4)} + N^{-1+1/r}.
\]

Proof. See Proposition S6.2 in the supplementary materials for details.

The optimal $\tilde{K}^*$ is derived by balancing the dominating terms in (5.3) within the feasible domain. Note that the sum of first two terms in the RHS of (5.4) corresponds to the errors due to approximating family $\mathcal{P}_K$, whereas the last term is the variance term for fitting $N$ observations with $K$ parameters. In fact, if the fitting sample is small, then the optimal number of components would scale as $\tilde{K}^* \asymp N^{1/r}$, and $D(p||\hat{p}_{N,K^*}) = O_p\left(N^{-1+1/r}\right)$; the rate is the same as if we are working with the family $\mathcal{P}_K$ instead of $\hat{p}_K$.

5.2. Consistency of Component Scores

We next establish the convergence of the component scores of pre-smoothed training trajectories $\hat{f}_i = \hat{\psi}_i p_i, i = 1, \ldots, n$, which are used to capture the distribution of random densities for the MAP and BLUP methods. Let $\hat{f}_{i,k} = \hat{\mu} + \sum_{k \leq K} \hat{\eta}_{ik} \varphi_k$ be the truncated pre-smoothed log-density with $\hat{\eta}_{ik} = (\hat{f}_i - \hat{\mu}, \varphi_k)$ for $k = 1, \ldots, K, i = 1, \ldots, n$. The next proposition shows that the estimated component scores $\hat{\eta}_{ik}$ of the pre-smoothed log-density converge uniformly to the truth $\eta_{ik} = (\mu, \varphi_k)$.

Proposition 5.2. Under (A1), (K1), (C2), and (C3), for any $K < \infty$, as $n \to \infty$, we have
\[
    \sup_{1 \leq n} \sup_{k \leq K} |\hat{\eta}_{ik} - \eta_{ik}| = O_p\left(\sqrt{\frac{\log n}{m^4}} + (n^{-1/2} + b_n) \sup_{k \leq K} \delta_k^{-1}\right).
\]

Proof. See Proposition S5.4 and Proposition S5.3 in the supplementary materials.

To extend the previous result to $K \to \infty$, we need $n, K \to \infty$ in such a way that $(n^{-1/2} + b_n) \sup_{k \leq K} \delta_k^{-1} \to 0$, which is satisfied under the conditions of Theorem 5.1 and $K = \tilde{K}^*$ from Proposition 5.1, implying that the sample component scores are uniformly consistently estimated.

5.3. Consistency of the MLE in a Small and Fixed Sample

The following result further characterizes the asymptotic behavior of the proposed MLE for a given new fitting sample with a small and fixed sample size $N$, where $K$ is finite and fixed and there is a large amount of observations from training subpopulations.

Proposition 5.3. Under (A1), (K1), (C3), and a fixed $K$, for a given sample $X_1, \ldots, X_N$ drawn from density $p \in \mathcal{P}$, suppose that the MLE $\hat{p}_{N,K}^o \in \mathcal{P}_K$ exists and is unique with corresponding parameters $\hat{\theta} \in \mathbb{R}^K$. Then the MLE $\hat{p}_{N,K} \in \mathcal{P}_K$ exists and is unique with probability tending to 1 as $n \to \infty$, where $n$ is the number of training subpopulations used to construct $\hat{p}_K$. The corresponding parameter $\hat{\theta}_n \in \mathbb{R}^K$ converges to $\hat{\theta}$ in probability with
\[
    ||\hat{\theta}_n - \hat{\theta}||_2 = O_p\left(\frac{1}{\sqrt{n}} + a_n + b_n\right),
\]
\[
    D(p||\hat{p}_{N,K}^o) = O_p\left(\frac{1}{\sqrt{n}} + a_n + b_n\right).
\]

Proof. See Proposition S6.3 and Corollary S6.2 of in the supplementary materials.

6. Discussion

We develop new methods to borrow information from densely observed subpopulations for estimating new densities with sparse observations, which demonstrates preferable theoretical and numerical properties. The applicability of the proposed methods hinges on two conditions, namely a common support
for the densities and the availability of densely observed training subpopulations. Due to the pre-smoothing process, the underlying density \( \hat{p}_i \) must be positive on \( \mathcal{T} \) for constructing centered log-ratios, and there must be sufficiently dense observations to ensure that the pre-smoothed density \( \tilde{p}_i \) is bounded away from zero. Scaling the raw observations so that they are more evenly distributed on the support, as we did for the precipitation data and further explained in Section S1.4 of the supplementary materials, can alleviate the support issue. Alternative transformations that can handle zero density values can also be considered, such as the quantile transformation (Panaretos and Zemel 2016). Our method requires densely observed training subpopulations to estimate the exponential families. Given subpopulations with heterogeneous sample sizes, the threshold for dense subpopulation can be set according to model stability and estimation performance, as detailed in Section S1.1. If all subpopulations have only sparse observations, our modeling framework would require oversmoothing; alternatives are considered in Wu, Müller, and Zhang (2013) and Gervini (2016).

**Supplementary Materials**

The supplementary materials include additional practical considerations, simulation studies, and proofs of the theoretical results.

**Acknowledgments**

We thank the associate editor and referees for their comments that have greatly improved this article. The majority of the work is completed while Xiongtao Dai is at Iowa State University.

**Disclosure Statement**

The authors report there are no competing interests to declare.

**Funding**

Qiu is supported in part by NSF grant TRIPODS CCF-1934884. Dai is supported in part by NSF grant DMS-2113713.

**ORCID**

Jiaming Qiu [https://orcid.org/0000-0002-2174-0337](https://orcid.org/0000-0002-2174-0337)

Xiongtao Dai [https://orcid.org/0000-0002-6996-5930](https://orcid.org/0000-0002-6996-5930)

Zhengyuan Zhu [https://orcid.org/0000-0002-2266-0646](https://orcid.org/0000-0002-2266-0646)

**References**

Amari, S.-i. (2016), *Information Geometry and its Applications*, Volume 194 of Applied Mathematical Sciences, Tokyo: Springer. [177,178,179]

Amari, S.-i., and Nagaoka, H. (2000), *Methods of Information Geometry*, Volume 191 of Translations of Mathematical Monographs, Providence, RI: American Mathematical Society; Oxford: Oxford University Press. [178]

Bonevski, B., Randell, M., Paul, C., Chapman, K., Tsywman, L., Bryant, J., Brozek, L., and Hughes, C. (2014), “Reaching the Hard-to-Reach: A Systematic Review of Strategies for Improving Health and Medical Research with Socially Disadvantaged Groups,” *BMC Medical Research Methodology*, 14, 1–29. [176]

Chu, B., Huynh, K., Jacho-Chávez, D., and Kryjvitso, O. (2018), “On the Evolution of the United Kingdom Price Distributions,” *The Annals of Applied Statistics*, 12, 2618–2646. [176]

Dai, X. (2021), “Statistical Inference on the Hilbert Sphere with Application to Random Densities,” *Electronic Journal of Statistics* (to appear). [176]

Delgado, P. (2011), “Dimensionality Reduction When Data are Density Functions,” *Computational Statistics & Data Analysis*, 55, 401–420. [176]

Egozcue, J. J., Díaz-Barreiro, J. L., and Pawlowsky-Glahn, V. (2006), “Hilbert Space of Probability Density Functions based on Aitchison Geometry,” *Acta Mathematica Sinica (English Series)*, 22, 1175–1182. [176,178]

Egozcue, J. J., Pawlowsky-Glahn, V., Tolosana-Delgado, R., Ortego, M. L., and van den Boogaart, K. G. (2013), “Bayes Spaces: Use of Improper Distributions and Exponential Families,” *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas*, 107, 475–486. [176]

Fithian, W., and Wager, S. (2015), “Semiparametric Exponential Families for Heavy-Tailed Data,” *Biometrika*, 102, 486–493. [177]

Gervini, D. (2016), “Independent Component Models for Replicated Point Processes,” *Statistical Spatial*, 18, 474–488. [177,188]

Gervini, D., and Khanal, M. (2019), “Exploring Patterns of Demand in Bike Sharing Systems via Replicated Point Process Models,” *Journal of the Royal Statistical Society, Series C*, 68, 585–602. [177,180,181,182]

Han, K., Müller, H.-G., and Park, B. U. (2020), “Additive Functional Regression for Densities as Responses,” *Journal of the American Statistical Association*, 115, 997–1010. [179]

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., and Filzmoser, P. (2016), “Simple Principal Component Analysis for Density Functions in Bayes Spaces,” *Computational Statistics & Data Analysis*, 94, 330–350. [176,178]

Huynh, K. P., Jacho-Chávez, D. T., Petrunia, R. J., and Voia, M. (2011), “Functional Principal Component Analysis of Density Families with Categorical and Continuous Data on Canadian Entrant Manufacturing Firms,” *Journal of the American Statistical Association*, 106, 858–878. [176]

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghasemi, M., Moody, B., Szelovits, P., Celi, L. A., and Mark, R. G. (2016), “MIMIC-III, a Freely Accessible Critical Care Database,” *Scientific Data*, 3, 1–9. [176,182]

Kim, D., Lee, Y. K., and Park, B. U. (2020), “Principal Component Analysis for Hilbertian Functional Data,” *Communications for Statistical Applications and Methods*, 27, 149–161. [176]

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. [176,177]

Koopkoerberg, C., and Stone, C. J. (1991), “A Study of Logslope Density Estimation,” *Computational Statistics & Data Analysis*, 12, 327–347. [179,180,181,182]

Lee, G., and Scott, C. (2012), “EM Algorithms for Multivariate Gaussian Mixture Models with Truncated and Censored Data,” *Computational Statistics & Data Analysis*, 56, 2816–2829. [182]

Loader, C. R. (1996), “Local Likelihood Density Estimation,” *The Annals of Statistics*, 24, 1602–1618. [180,181,182]

Panaretos, V. M., and Zemel, Y. (2016), “Amplitude and Phase Variation of Point Processes,” *The Annals of Statistics*, 44, 771–812. [176,177,188]

Petersen, A., and Müller, H.-G. (2016), “Functional Data Analysis for Density Functions by Transformation to a Hilbert Space,” *The Annals of Statistics*, 44, 183–218. [176,178,179]

Sheather, S. J., and Jones, M. C. (1991), “A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation,” *Journal of the Royal Statistical Society, Series B*, 53, 683–690. [180,182]

Srivastava, A., Jernyn, I., and Joshi, S. (2007), “Riemannian Analysis of Probability Density Functions with Applications in Vision,” in 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. [176]

Tsai, R. S. (2016), “Some Methods for Analyzing Big Dependent Data,” *Journal of Business and Economic Statistics*, 34, 673–688. [176]

van den Boogaart, K. G., Egozcue, J. J., and Pawlowsky-Glahn, V. (2014), “Bayes Hilbert Spaces,” *Australian & New Zealand Journal of Statistics*, 56, 171–194. [176]

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (1st ed.), Cambridge, UK: Cambridge University Press. [186]

Wakefield, J., and Elliott, P. (1999), “Issues in the Statistical Analysis of Small Area Health Data,” *Statistics in Medicine*, 18, 2377–2399. [176]

Wu, S., Müller, H.-G., and Zhang, Z. (2013), “Functional Data Analysis for Point Processes with Rare Events,” *Statistica Sinica*, 23, 1–23. [177,188]

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), “Functional Data Analysis for Sparselongitudinal Data,” *Journal of the American Statistical Association*, 100, 577–590. [179]