Supplement to “Nonparametric Estimation of Repeated Densities with Heterogeneous Sample Sizes”

Jiaming Qiu\textsuperscript{1}, Xiongtao Dai\textsuperscript{1}, and Zhengyuan Zhu\textsuperscript{1}

\textsuperscript{1}Department of Statistics, Iowa State University

Contents

\textbf{S1 Practical Considerations} \hspace{2cm} 3
- S1.1 Determine Training Subpopulations \hspace{2cm} 3
- S1.2 Additional Evaluation for MIMIC \hspace{2cm} 8
- S1.3 Number of Components $K$ Selected by AIC and FVE \hspace{2cm} 10
- S1.4 Scaling the Responses for Heavy-tailed Densities \hspace{2cm} 11

\textbf{S2 Additional Simulation Results} \hspace{2cm} 11
- S2.1 Random Intercept Family \hspace{2cm} 11
- S2.2 AIC-driven $K^*$ in Finite Dimensional Family and MISE of Simulation Results \hspace{2cm} 14

\textbf{S3 Notation and Assumptions for Theories} \hspace{2cm} 19
- S3.1 List of Assumptions \hspace{2cm} 21

\textbf{S4 Some Lemmas and Proofs} \hspace{2cm} 25
- S4.1 Identifiability \hspace{2cm} 25
- S4.2 Equivalence of $L^2$ and $L^\infty$ norm in $S_K$ \hspace{2cm} 27

\textbf{S5 Asymptotic Behaviors of Sample Mean, Covariance, Eigenvalues, Eigenfunctions, and Component Scores} \hspace{2cm} 27
- S5.1 Pre-smoothing \hspace{2cm} 27
- S5.2 Limiting Behaviors \hspace{2cm} 36
- S5.3 Perturbation Theory \hspace{2cm} 38
- S5.4 Projection Error \hspace{2cm} 41

\textbf{S6 Results related to KL divergence} \hspace{2cm} 46
- S6.1 A Uniform Convergence Result in KL Divergence \hspace{2cm} 48
- S6.2 Main Result for MLE Consistency \hspace{2cm} 49
- S6.3 Optimal $K$ \hspace{2cm} 53
S6.4 Effect of Training
S1 Practical Considerations

For implementation, we adopted \texttt{nloptr}, a \texttt{R} version of NLopt (Johnson 2020) for optimization, and \texttt{fdapace} (Carroll et al. 2020) for functional data analysis. An \texttt{R} package implementing the proposed methods will be made available on GitHub.

For pre-smoothing, we considered KDE and logspline in numerical examples. For the bandwidth selection for KDE, with the Gaussian kernel, we first applied the method of Sheather and Jones (1991) to obtain optimal bandwidths $h^*_i$, $i = 1, \ldots, n$ for the training samples, and then utilized the median bandwidth $h^* = \text{median}(h^*_1, \ldots, h^*_n)$ to each training sample for more stable estimation.

As both the theoretical derivation and the practical experience indicates, comparing to the entire data range, setting a slightly smaller working domain $T$ can greatly improve stability of the presmoothing steps in the proposed methods, since the transformation $\psi$ is more than a point-wise map and takes the entire density as input. Though, working domain should also cover most observations of fitting samples, since the approximating families are not defined outside $T$.

Alternative to truncating working domains is to assume a common compact support for all densities in $\mathcal{P}$. One issue there is that the KDE would be vulnerable to boundary issue. To address, one could use any boundary correction KDE, such as a weighted version (c.f. e.g., Petersen and Müller (2016), or section 11.3 of Bickel and Doksum (2015)).

The replicated point process (RPP) approach in our simulation was adopted from Gervini and Khanal (2019). There was a periodicity restriction for the densities in the original paper, though we used the Matlab implementation that would be otherwise identical but without the periodicity restriction, as post by the original first author.

The “GEV” in the precipitation example (Subsection 4.2, Figure 4.3) stands for estimation by fitting a generalized extreme value (GEV) distribution, whose density takes the form of

\[ p(t|\theta) = \theta_3^{-1} \omega(t) \theta_3 + 1 e^{-\omega(t)}, \quad (S1.1) \]

where $\omega(t) = \exp(-(t - \theta_1)/\theta_2)$ if $\theta_3 = 0$, and $\omega(t) = (1 + \theta_3(t - \theta_1)/\theta_2)^{-1/\theta_3}$ otherwise. We use \texttt{fevd} function of the \texttt{R} package \texttt{extRemes} (Gilleland and Katz 2016) for this method.

S1.1 Determine Training Subpopulations

As pointed out during the review process, the training subpopulations are subject to user’s selection, a proper choice of which remains unanswered. We recommend setting a cutoff size, like the $m = 75$ as in the medical data (MIMIC) (Subsection 4.1 of main text), so that all subpopulations of larger size would be used as training. There is a trade-off between number of training samples $n$ and $m$, while according to our theory we would like both quantities to be sufficiently large.
Overall, the model stability and density estimation performance should be considered for a sensible choice. The number of training subpopulations $n$, the bandwidth $h$ of pre-smoothing KDE, the estimated component scores $\hat{\eta}$, mean $\hat{\mu}$ and eigenfunctions $\hat{\phi}_k$ can be used to guide the choice of such cutoff size.

Ideally, $n$ should be sufficient so that the distributional information for the random density $p$ can be well represented by the component scores $\hat{\eta}_{ik}$. The bandwidth $h$ should be suitable that it neither undersmooth nor oversmooth. In particular, the $h$ here is selected as the median of all SJ bandwidth (Sheather and Jones 1991) of individual training subpopulations, thus if the cutoff size $m$ were to be small, $h$ would be larger as shown in the top left panel of Figure S1.3, resulting in an oversmoothed $\hat{\mu}$ and $\hat{\phi}_k$ as in the first column of Figure S1.1.

Moreover, the estimated mean $\hat{\mu}$ and eigenfunctions $\hat{\phi}_k$ shall be stable when slightly shifting $m$ near the chosen value, as shown in the Figure S1.1. In contrast, greater variation will suggest differences in the pattern of random density $p$ among subpopulations smaller or larger than $m$, which makes information borrowing across subpopulations questionable.

In addition, the distribution of estimated component scores should resemble that of the Gaussian distribution, which is assumed when we construct distributional information for MAP and BLUP. As shown in the first column of Figure S1.2, with smaller $m$, such distribution tends to be ill-posed, largely resulted from the smaller subpopulations. On the other hand, picking larger $m$ could lead to nicer-looking distributions, but it risks the representativity of overall population, since there would be only a handful of training subpopulations, the $\hat{\phi}_k$ also digress from the other ones comparing the rightmost column to other panels in Figure S1.1.

The performance on testing samples can also be consulted, but one shall be particularly careful about what is being compared. In our case, the performance of most testing subpopulations were unaffected, as shown in the bottom panels of Figure S1.3. However, since the cross-entropy is used, one must only compare the performance on an identical set of testing subpopulations. By definition (4.1), cross-entropy for different subpopulations cannot be compared directly, since $H_{LOO}(p_1, \hat{p}_1) < H_{LOO}(p_2, \hat{p}_2)$ does not necessarily indicate better performance for $p_1$ than $p_2$.

For example, in the lower middle panel of Figure S1.3, let $S$ be the collection of all testing subpopulations with size in $(10, 35]$, the solid lines and ribbons correspond to mean and inter-quartile range of $\{i \in S : H_{LOO}(p_i, \hat{p}_{i,m})\}$, where $\hat{p}_{i,m}$ is the estimated density with the proposed methods trained with all subpopulations larger than $m$. Noting the training and testing subpopulations must be distinct, $S$ will change for changing $m \in (10, 35]$, thus reasonable comparison in performance for different $m$ can only be made for $m > 35$. Indeed, observe that the cross entropy tends to be smaller for larger samples as shown in lower panels of Figure S1.3. Therefore, as $m$ increases on $(10, 35]$ and consequently $S$ starts to include larger subpopulations, the average of cross entropy over entire $S$ would decrease, creating a misleading artifact.

In combine, we settled at $m = 75$ for the MIMIC data, and in coincidence the corresponding $n = 75$ in our MIMIC data.
Figure S1.1: For the medical data (MIMIC) in Subsection 4.1 of main text and in addition to Figure S1.3, the variation of estimated mean function and eigenfunctions for the MIMIC data, varying the minimal training sample sizes $m = 10, 15, \ldots, 200$, where the ribbons show the range of estimated values and the solid lines indicate the results for $m = 25, 75, 125, 175$, respectively.
Figure S1.2: For the medical data (MIMIC) in Subsection 4.1 of main text and in addition to Figure S1.3, the changes in the pairwise joint distribution of the first three estimated component scores (on different rows) as the threshold $m$ for the training sample size varies (in different columns). Each dot represents one training subpopulation colored according to the sample size $N_i$. 
Figure S1.3: For the medical data (MIMIC) in Subsection 4.1 of main text, as the minimal training sizes $m$ varies among $10, 15, 20, \ldots, 200$, the changes in bandwidth $h$ used in pre-smoothing (top left) and the number $n$ of training samples (top right), and the variation of cross-entropy errors in the testing samples of size $5 \leq N \leq 10$, $10 < N \leq 35$, and $35 < N \leq 75$ respectively, where the solid lines represent the mean errors, while ribbons show the inter-quartile range. The cross entropy plots are separated by sizes to ensure that testing samples are identical among all plotted $m$ in each panel.
S1.2 Additional Evaluation for MIMIC

We now compare the proposed methods to KDE with artificially constructed sparse samples. In setup identical to that in Subsection 4.1 of main text, additional sparse samples were created by leaving out majority of observations in the subpopulation. Density estimation using the proposed methods and KDE based on the sparsified samples were compared to KDE based on the original densely observed samples. For example, with a full sample of $N = 400$ for female aortic valve disorder, we randomly selected 20 observations to form a sparsified sample. Then density estimations were computed using the proposed methods (trained without this subpopulation) and KDE with the sparsified sample. In addition, another KDE is computed with the full sample ($N = 400$) as dense case. As shown in Figure S1.4 of the eight largest original subpopulations, the proposed methods (colored curves) were much closer to the KDE obtained from the dense samples in (dashed black), regarded as the underlying truth, compared to the KDE constructed with the sparsified samples (solid black). Note that the MAP and BLUE methods that borrow distributional information clearly outperformed KDE, and even the MLE method that borrows only the typical shape of the densities outperformed KDE in most cases. Note that unlike Figure 4.1 in the main text where $N$ denote the total number of observations available per subpopulation, $N$ here denote the number of observations within range [14, 85].
Figure S1.4: For the largest few subpopulations in the medical data (MIMIC) in Subsection 4.1 of main text, the estimated densities from the proposed MLE, MAP, BLUP methods (colored curves), and KDE (solid black curve) based on artificially sparsified samples of size 20, and the KDE based on original dense samples (dotted lines). The histograms of original dense samples are shown in each panel, with the size noted as $N$. The bottom rug plot in each panel represents the sparsified sample.
S1.3 Number of Components $K$ Selected by AIC and FVE

In the end of Section 2, we proposed to select number of components via AIC while fitting sparse subpopulations. The following Figure S1.5 and Figure S1.6 show the AIC driven $K^*$ and the fraction of variation explained (FVE) for the MIMIC and precipitation data. The FVE of the first $K$ components is defined as $\sum_{k=1}^{K} \frac{\lambda_k}{\sum_{j=1}^{\infty} \lambda_j}$, and in practice the denominator is computed by summing all positive eigenvalues. From the FVE plots, it is clear that for both data set, most variation in the random densities could be explained by the first few components. Therefore the AIC-driven $K^*$ is overall proper, as the majority of which reside below 3.

Figure S1.5: In addition to Figure 4.2 of main text for the medical data (MIMIC), here we show the boxplot of fraction of variation explained (FVE) of the FPCA result (right) and the overall distribution of number of components selected by AIC while estimating fitting samples (left).

Figure S1.6: In addition to Figure 4.3 of main text for the precipitation data, here we show the boxplot of fraction of variation explained (FVE) of FPCA results among 10 repeated experiments (right) and the overall distribution of number of components selected by AIC while estimating fitting samples (left).

In addition, Figure S2.2 also illustrate the performance of $K^*$ in simulation
involving finite dimensional families.

**S1.4 Scaling the Responses for Heavy-tailed Densities**

Suppose that the original responses \( \{ Y_{ij} \in \mathbb{R}^+ : i = 1, \ldots, n; j = 1, \ldots, N_i \} \) have a heavy right-tail and the pre-smoothing density estimates for the training sub-populations are occasionally zero, which will create an issue when considering log-densities. To address this issue, we log-scale the responses and obtain \( X_{ij} \triangleq \log Y_{ij} \) lying in a compact working domain \( T \triangleq [\min(X_{ij}) - \delta, \max(X_{ij}) + \delta] \) for some adjustment \( \delta > 0 \). Using generic notations, denote \( Y \) as the random variable associated with responses in original scale with density \( p_Y(y) \) for \( y > 0 \), and the density of \( X \) as \( p_X(x) \) for \( x \in T \).

The proposed methods now proceed by applying the FPCA to the transformed trajectories \( f_{X,i} = \psi p_{X,i}, i = 1, \ldots, n \), leading to mean function \( \mu_X \) and eigenfunctions \( \varphi_{X,k} \), \( k \in \mathbb{N}^+ \), with which we construct approximating families for the distribution of \( X \), whose density functions would take form of

\[
p_{X,\theta}(x) \propto \exp \left( \mu_X(x) + \sum_{k=1}^{K} \theta_k \varphi_{X,k}(x) \right), x \in T,
\]

where \( \theta_1, \ldots, \theta_K \) are natural parameters. In the end, reverse the log-scaling to arrive at approximating families for distribution of \( Y \), densities in which take form of

\[
p_{Y,\theta}(y) \propto \exp \left( \mu(\log y) - \log y + \sum_{k=1}^{K} \theta_k \varphi_k(\log y) \right), y \in \mathbb{R}^+.
\]

The distribution of component scores is preserved before and after scaling, since all parameters \( \theta_1, \ldots, \theta_K \) are unchanged from \( p_{X,\theta} \) to \( p_{Y,\theta} \), making the proposed MAP and BLUP methods still valid.

**S2 Additional Simulation Results**

**S2.1 Random Intercept Family**

It is of particular interest to compare the proposed methods with classical parametric approaches when the data are generated according to a random intercept model

\[
X_{ij} = A_i + \sigma_e \varepsilon_{ij}
\]  

(S2.1)

for stratum \( i = 1, \ldots, n \) and observation \( j = 1, \ldots, N_i \) within each stratum, where the \( A_i \) follow i.i.d. zero-mean normal distribution with variance \( \sigma_a^2 \), the \( \varepsilon_{ij} \) are i.i.d. noise with mean zero and variance one, and \( \sigma_a \) and \( \sigma_e \) are standard deviation parameters regarded as unknown.

Two cases for the distribution of noise \( \varepsilon_{ij} \) were considered, a normal distribution and a heavy-tailed \( t \)-distribution with 3 degrees of freedom. In each
experiment, \( n = 100 \) training samples were created each with sample size uniformly distributed on the integers between 75 and 100. Next, \( n^* = 100 \) testing samples each with size ranging from 10 to 20 with equal probability were generated. Standard deviations were set to be \( \sigma_a = 1 \) and \( \sigma_e = 1 \). The mean KL divergence were evaluated for the fitted estimates on domain \( \mathcal{T} = [-10, 10] \).

We compare against an additional density estimator constructed by the classical linear mixed-effects model approach. Under normality assumption, a parametric density estimator for a new subpopulation given \( N \) new observations \( \{X^*_j : j = 1, \ldots, N\} \) is constructed as the density of \( N(\hat{A}, \hat{\sigma}_e^2) \), where \( \hat{A} = X^*(1 + \hat{\sigma}_e^2 N^{-1}\hat{\sigma}_a^{-2})^{-1} \) is the BLUP for random effect, and \( \hat{\sigma}_e \) and \( \hat{\sigma}_a \) are residual maximum likelihood estimates obtained from the training samples. This parametric approach is referred to as linear mixed-effects (LME) method in our comparisons.

Results reported in Figure S2.1 shows that KDE performed the worst, the LME method excelled only when the model is correctly specified under normal noises but suffered considerably under \( t \)-distributed noises. In contrast, the proposed methods were stable under both setup. In the normal noise case shown in the left panel of Figure S2.1, even the proposed FPCA_MLE performed better than MLE, suggesting the approximating families more closely resemble the underlying truth. Indeed, the underlying density family is one-dimensional, and the proposed FPCA_MLE utilized one component as selected by AIC for more than 80% of the experiments, while the MLE has to estimate two unknown parameters since the variance is unknown. Moreover, as shown in the \( t \) noise scenario, the proposed methods outperformed KDE and LME, even if the underlying family is not an exponential family. This shows that the approximation of an arbitrary family through exponential families is fruitful in density estimation.

One shall not be surprised that the LME method outperformed others when the model is well-specified. The strength of LME would be, especially when homoscedasticity would be granted, that the estimates of variance components could pool information from all subpopulations and hence improve performance. In comparison, the MLE method within the ground truth family does not borrow strength across subpopulations. The proposed methods, especially FPCA_MAP and FPCA_BLUP, which utilize population-level knowledge of parameters also outperformed MLE, suggesting that, similar to LME, they successfully pool information from the training samples.
Figure S2.1: Boxplot of Mean KL divergence, one dot for one replication. For FPCA related methods, classic KDE used for pre-smoothing. Normal noise means data generated following model \((S2.1)\) and \(\varepsilon_{ij} \sim N(0,1)\), t noise means same model but \(\varepsilon_{ij} \sim t_3\).
S2.2 AIC-driven $K^*$ in Finite Dimensional Family and MISE of Simulation Results

Figure S2.2 illustrates the performance of AIC selected $K^*$ in finite dimensional families. The choice of $K^*$ by AIC is overall appropriate, selecting more components in larger samples to unveil more information. The majority of the $K^*$ were no greater than the actual (finite) dimension of the underlying family, which are two and one for the normal and the bimodal family, respectively.

![Figure S2.2: With setup identical to that of Figure 3.1 in the main text, this figure shows the number of components used as selected by AIC.](image)

In addition to the mean KL divergence results shown in Figure 3.1, Figure 3.2 of main text, and Figure S2.1, mean integrated squared errors (MISE) were also considered. The behavior of MISE is rather similar to that of the MKL, as demonstrated in the following Figure S2.3, Figure S2.4, and Figure S2.5. More precisely, following the set up in the beginning of Section 3 in the main text, we define the MISE as

$$MISE = \frac{1}{n^*} \sum_{l=1}^{n^*} \int_{\tau} (\hat{p}_l^* (t) - p_l^* (t))^2 dt,$$

where $n^*$ is the number of testing samples, $p_l^*$ is the true density of the $l$th sample and $\hat{p}_l^*$ is its estimate for $l = 1, \ldots, n^*$. 

14
Figure S2.3: Boxplot of MISE in analogy to Figure S2.1, one dot for one replication. For FPCA related methods, classic KDE used for pre-smoothing. Normal noise means data generated following model (S2.1) of main text and $\varepsilon_{ij} \sim N(0,1)$, t noise means same model but $\varepsilon_{ij} \sim t_3$. 
Figure S2.4: Boxplots of MISE in analogy to Figure 3.1 of main text. RPP, repeated point processes approach (Gervini and Khanal 2019); locfit, local polynomial density estimate (Loader 1996); logspline, adaptive logspline (Kooperberg and Stone 1991). Both discrete training samples of sizes $N_i = 200$ with KDE for pre-smoothing and completely observed densities are considered for our proposed approaches, namely FPCA_MLE, FPCA_MAP, and FPCA_BLUP.
Figure S2.5: In analogy to Figure 3.2 of main text, the left panel shows the ratio of the average MISE of compared methods over that of the EM algorithm, where a ratio smaller than 1 indicates better performance than EM. The right panel displays the average number of components $K$ selected by the proposed AIC method, with ribbons indicating one standard deviation around the means. Axes in both panels are drawn in log 10 scale. Size of training samples were 50, and logspline was used as the pre-smoother. RPP, repeated point processes approach (Gervini and Khanal 2019) with 4 components; locfit, local polynomial density estimate (Loader 1996); logspline, adaptive logspline (Kooperberg and Stone 1991); FPCA_MLE, FPCA_MAP, and FPCA_BLUP, the proposed methods in the MLE, MAP, and BLUP variants.
In addition to Figure S2.5, the absolute values of errors are shown in Figure S2.6 below, which assures that the performance of the proposed methods improve for increasing testing sample size $N$. It is also clear that as $N$ increases, the gap in performance between the proposed MLE and the two methods adopting shrinkage (BLUP and MAP) decreases, suggesting such shrinkage effect would be weaker with growing sample sizes. In addition, the results in the first two columns appear similar, suggesting the proposed methods are robust to the choice of pilot estimator in pre-smoothing step.

Interestingly, the method of Gervini and Khanal (2019), as implemented in the Matlab code provided by the original first author, performed worse as the sample size increased above 100. This may be due to numeric instability and the difficulty of solving a constrained optimization problem in their method.

![Graph showing the absolute values of average MISE (top) and KL divergence (bottom) along increasing testing sample size $N$. The proposed methods utilizing KDE (leftmost column) or logspline (middle column) for presmoothing were compared to other methods (rightmost column).](image-url)
Denote the random density as $p \in \mathcal{P}$, where $\mathcal{P}$ is a set of positive densities on compact interval $\mathcal{T} = [a, b]$. Further, let $\mathcal{T}_\epsilon = [a - \epsilon, b + \epsilon]$ be the extended domain and we assume for any $q \in \mathcal{P}$, there exists an extended version $q_e$ being density on $\mathbb{R}$ and positive on $\mathcal{T}_\epsilon$ satisfying $q(t) = q_e(t)/\int_{\mathcal{T}_\epsilon} q_e(s)ds$ for all $t \in \mathcal{T}$. We write the family of the extended densities as $\mathcal{P}_e = \{q_e : q \in \mathcal{P}\}$. We write random density $p_{e,i} \in \mathcal{P}_e$ as the extended version of $p_i$ satisfying $p_i(t) = p_{e,i}(t)/\int_{\mathcal{T}} p_{e,i}(s)ds$.

Denote the centralized log-transformed density as $f \triangleq \psi p \in L^2(\mathcal{T})$, where $\psi$ is defined in (2.1) of main text. Further let

$$
\mu(t) \triangleq Ef(t), \quad G(s, t) \triangleq \text{cov}(f(s) - \mu(s), f(t) - \mu(t))
$$

be the mean and covariance function of $f$; let $\lambda_k, \varphi_k, k = 1, 2, \cdots$ be the eigenvalues and eigenfunctions of the integral operator $G : g \mapsto \int_{\mathcal{T}} G(s, \cdot)g(s)ds$ associated with $G$; let $\delta_k \triangleq \inf_{j \neq k}|\lambda_j - \lambda_k|/2$ be the eigengaps. Define its rank to be the dimension of its image $\{Gg : g \in L^2(\mathcal{T})\}$, denoted as rank $G$, which may be infinity.

For independent copies $p_1, \ldots, p_n$ of the random density, denote $f_i = \psi p_i$, $i = 1, \ldots, n$, let

$$
\hat{\mu}(t) \triangleq \frac{1}{n} \sum_{i=1}^{n} f_i(t), \quad \hat{G}(s, t) \triangleq \frac{1}{n} \sum_{i=1}^{n} (f_i(s) - \hat{\mu}(s)) (f_i(t) - \hat{\mu}(t))
$$

be the sample mean and sample covariance function, let $\hat{\lambda}_k, \hat{\varphi}_k, k = 1, \ldots, n - 1$ be the eigenvalues and eigenfunctions associated with $\hat{G}$.

When $p_i$ are not observed but only $\{X_{ij} : 1 \leq j \leq N_i, 1 \leq i \leq n\}$ are provided, for $i = 1, \ldots, n$, let $\hat{p}_i$ be the pilot density estimates with the sample $\{X_{ij} : 1 \leq j \leq N_i\}$ in the pre-smoothing step, and $f_i = \psi \hat{p}_i$ be its corresponding centered log-ratio transformation. Denote $m = m(n) \triangleq \min_{1 \leq i \leq n} N_i$ be the minimum sample size out of the $n$ samples, and let

$$
\hat{\mu}(t) \triangleq \frac{1}{n} \sum_{i=1}^{n} \hat{f}_i(t), \quad \hat{G}(s, t) \triangleq \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i(s) - \hat{\mu}(s)) (\hat{f}_i(t) - \hat{\mu}(t))
$$

be the sample mean and sample covariance function when pre-smoothing is involved, let $\hat{\lambda}_k, \hat{\varphi}_k$ for $k = 1, \cdots, n - 1$ be the eigenvalues and eigenfunctions associated with $\hat{G}$.

For theoretical discussion involving presmoothing, in addition to

$$
\hat{p}_i(t) = \left(\int_{\mathcal{T}} \sum_{j=1}^{N_i} \kappa \left(\frac{x - X_{ij}}{h}\right)dx\right)^{-1} \sum_{j=1}^{N_i} \kappa \left(\frac{t - X_{ij}}{h}\right), \quad t \in \mathcal{T}
$$
as listed in (2.11) of the main text, we also consider

\[ \tilde{p}_{e,i}(t) = \frac{1}{N_i h} \sum_{j=1}^{N_i} \kappa \left( \frac{t - X_{ij}}{h} \right), \quad t \in \mathcal{T}, \]

where \( X_{i1}, \ldots, X_{iN_i} \) form i.i.d. sample from the extended density \( p_{e,i} \). The purpose is to alleviate boundary issue, as \( \tilde{p}_{e,i} \) can achieve a smaller bias uniformly on \( \mathcal{T} \).

Further, in light of the Karhunen–Loève expansion, given \( f \in L^2(\mathcal{T}) \), we write

\[
\begin{align*}
\hat{f}_K &\triangleq \hat{\mu} + \sum_{k=1}^{K} \hat{\eta}_k \hat{\varphi}_k, \\
\tilde{f}_K &\triangleq \tilde{\mu} + \sum_{k=1}^{K} \tilde{\eta}_k \tilde{\varphi}_k, \\
\hat{\hat{f}}_K &\triangleq \hat{\mu} + \sum_{k=1}^{K} \hat{\eta}_k \hat{\varphi}_k,
\end{align*}
\]

where \( \eta_k = \langle f - \mu, \varphi_k \rangle, \tilde{\eta}_k = \langle f - \tilde{\mu}, \tilde{\varphi}_k \rangle, \) and \( \hat{\eta}_k = \langle f - \hat{\mu}, \hat{\varphi}_k \rangle \). We denote the space spanned by eigenfunctions as \( \mathcal{S} \), i.e., \( \mathcal{S} \triangleq \text{span}\{ \varphi_k : k = 1, 2, \cdots \} \), while \( \mathcal{S}_K \triangleq \text{span}\{ \varphi_k : k = 1, \ldots, K \} \) be that with the first \( K \). Further let \( \mathcal{F} \triangleq \mu + \mathcal{S} = \{ \mu + g : g \in \mathcal{S} \} \), \( \mathcal{F}_K \triangleq \mu + \mathcal{S}_K = \{ \mu + g : g \in \mathcal{S}_K \} \); and \( \mathcal{S}, \mathcal{S}_K, \mathcal{F}, \mathcal{F}_K, \mathcal{S}_K, \mathcal{F}_K \) denote the respective quantities replacing \( \mu, \varphi \) by \( \hat{\mu}, \hat{\varphi} \) or \( \tilde{\mu}, \tilde{\varphi} \). Therefore \( f_K, \tilde{f}_K \) and \( \hat{f}_K \) are projections of \( f \) to \( \mathcal{F}_K, \tilde{\mathcal{F}}_K \) and \( \hat{\mathcal{F}}_K \) respectively. Note that it is not \( \tilde{f} \) but \( f \) that we are projecting. Hence, \( P_K = \psi^{-1} \mathcal{F}_K, \tilde{P}_K = \psi^{-1} \tilde{\mathcal{F}}_K \) and \( \hat{P}_K = \psi^{-1} \hat{\mathcal{F}}_K \), where \( \psi^{-1} \) denotes the pre-image under \( \psi \) in the space of densities. Let \( B_K, \tilde{B}_K \) and \( \hat{B}_K \) be the normalizing constant functions in corresponding exponential families as in (2.3) of main text.

In addition, for the rates of pre-smoothing errors, denote

\[
\begin{align*}
a_n &= a_n(m, h) = h^2 + \sqrt{\frac{1}{n} + \frac{1}{nmh}} \log n + \frac{\log n}{mh}, \\
b_n &= b_n(m, h) = h^2 + 1/\sqrt{nmh} + 1/mh,
\end{align*}
\]

in parallel to (5.1) and (5.2) of the main text.

For any symmetric, non-negative definite and compact operator \( G \) on \( L^2(\mathcal{T}) \) with eigenvalue \( \lambda_k \) and differentiable eigenfunctions \( \varphi_k \) for \( k \geq 1 \), let

\[
A_K \triangleq 2 \max \left( |\mathcal{T}|^{-1/2}, \left( \sum_{k=1}^{K} \| \varphi'_k \|_\infty \right)^{1/3} \right), \tag{S3.1}
\]

which would be closely related to the Lipschitz constant for functions in the space spanned by the eigenfunctions, which will be elaborated in Proposition S4.2. Also, denote the half eigen gap as \( \delta_k \triangleq \inf_{l \neq k} |\lambda_k - \lambda_l|/2 \).

For simplifying notation, for sequences of real numbers \( \{x_n\}, \{y_n\} \) and sequences of random variables \( \{X_n\}, \{Y_n\} \), as \( n \to \infty \), we sometimes write \( x_n \lesssim y_n \) or \( y_n \gtrsim x_n \) if \( x_n = o(y_n) \), \( X_n \lesssim_p Y_n \) or \( Y_n \gtrsim_p X_n \) if \( X_n = O_p(Y_n) \);
write \( x_n \sim y_n \) if \( x_n = O(y_n) \) and \( y_n = O(x_n) \); \( x_n \overset{p}{\rightharpoonup} Y_n \) if \( X_n = O_p(Y_n) \) and \( Y_n = O_p(X_n) \). For multiple indexed sequence of non-zero random variables \( \{U_{n,k}\} \) and \( \{V_{n,k}\} \), as \( n \to \infty \), we say \( U_{n,k} = O_p(V_{n,k}) \) uniformly over \( k \) if \( \sup_n \sup_k P(|U_{n,k}/V_{n,k}| > R) \to 0 \) as \( R \to \infty \).

We write the direct sum of vector spaces \( U \) and \( V \) as \( U \oplus V \triangleq \{u + v : u \in U, v \in V\} \). Also, we sometimes write \( a \wedge b \overset{\triangle}{=} \min(a, b) \) and \( a \vee b = \max(a, b) \) for \( a, b \in \mathbb{R} \).

We denote the space of squared integrable functions w.r.t. measure \( \nu \) on \( \mathcal{F} \) as \( L^2(\mathcal{F}, \nu) \) and the space of continuous functions as \( C(\mathcal{F}) \). For \( g : \mathcal{F} \to \mathbb{R} \), define \( L^\infty \) norm as \( \|g\|_\infty = \sup_{t \in \mathcal{F}} |g(t)| \), \( L^2 \) norm as \( \|g\|_2 = \sqrt{\int_\mathcal{F} g(t)^2 d\nu(t)} \). Hence, for \( G : \mathcal{F} \times \mathcal{F} \to \mathbb{R} \), we write \( \|G\|_\infty = \sup_{t,s \in \mathcal{F}} |G(s,t)| \) and \( \|G\|_2 = \sqrt{\int_{\mathcal{F} \times \mathcal{F}} G(s,t)^2 d\nu(s)d\nu(t)} \). Unless otherwise specified, we take \( \nu \) to be the Lebesgue measure and write \( L^2(\mathcal{F}) \). For operator \( G : L^2(\mathcal{T}) \to L^2(\mathcal{T}) \), define the induced operator norm \( \|G\| = \sup_{g \in L^2(\mathcal{T})} \|Gg\|_2 / \|g\|_2 \). For vector \( a \in \mathbb{R}^J \), define \( \|a\|_2 = \sqrt{\sum_{j=1}^J a_j^2} \), \( \|a\|_\infty = \max_{1 \leq j \leq J} |a_j| \).

### S3.1 List of Assumptions

Here we list all our assumptions, some of which ((A1) and (K1)) are modified from Petersen and Müller (2016).

(A1) All densities \( q_e \in \mathcal{P}_e \) are twice differentiable and strictly positive on the support \( \mathcal{T}_e := \{s : \inf_{t \in \mathcal{T}} |s-t| \leq \varepsilon_T \} \). Moreover, there is a constant \( M > 1 \) such that, for all \( q_e \in \mathcal{P}_e \), \( \sup_{t \in \mathcal{T}_e} \max(|q_e(t)|, |1/q_e(t)|, |q_e'(t)|, |q_e''(t)|) \leq M \).

Recall \( N_i, i = 1, \ldots, n \) are the sample sizes in the training subpopulations, denote \( m = \min_{1 \leq i \leq n} N_i \).

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

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

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

(C2) For all \( k = 1, \ldots, \text{rank } \mathcal{G} \), the non-zero eigenvalues are distinct; the estimated eigenfunctions are properly aligned so that \( \langle \varphi_k, \hat{\varphi}_k \rangle \geq 0 \) and \( \langle \varphi_k, \hat{\varphi}_k \rangle \geq 0 \).

(C3) \( n \to \infty, m \to \infty, \) and \( h \to 0 \) in a way such that
(i) $|\log h| = O(\log n)$ and that $(\log n)/m = o(h)$;
(ii) $h(h + 1/m) \geq \log(n)/n$.

(K1) The kernel $\kappa$ is a density function of bounded variation supported on $[-1, 1]$; $\int u^2 \kappa(u) du$ and $\int \kappa(u)^2 du$ are finite, with $\int u \kappa(u) du = 0$.

The following are some discussion on the assumptions in addition to the those regarding (C1), (C2), and (C3) in the main text.

Under (A1), we can take the following condition. The following are some discussion on the assumptions in addition to the those regarding (C1), (C2), and (C3) in the main text.

(A1) All densities $q_c \in \mathcal{P}_c$ are twice differentiable and strictly positive on

$$
T_c = \left\{ s : \inf_{t \in T} |s - t| \leq \varepsilon_T \right\}
$$

for some $\varepsilon_T > 0$. There is a constant $M > 1$ such that, for all $q_c \in \mathcal{P}_c$, $\|q_c\|_\infty$, $\|1/q_c\|_\infty$, $\|q_c'\|_\infty$, $\|q_c''\|_\infty$, $\|\psi q_c\|_\infty$, $\|\psi q_c'\|_\infty$, and $\|\psi q_c''\|_\infty$ are all bounded above by $M$.

Condition (A1') is equivalent to (A1), therefore in later discussion we write $M$ as the uniform bound for the densities, transformed densities and their derivatives. Further, as the derivatives are uniformly bounded, $\mathcal{F} = \psi \mathcal{P}$ forms a family of equicontinuous functions, and hence the random process $f = \psi p$ is mean-square continuous, implying its mean $\mu$ and covariance functions $G$ are also continuous (c.f. theorem 7.3.2 of Hsing and Eubank (2015)).

Under (A1), the covariance function $G$ is differentiable, so are its eigenfunctions. Since by definition, $\frac{\partial}{\partial s} G(s, t) = \frac{\partial}{\partial s} \text{cov}(f(s), f(t)) = \text{cov}(f(s), f'(t)) \leq 4M^2$, where the exchange of derivative and expectation granted by dominated convergence. Further $\varphi_k'(t) = \lambda_k^{-1} \int_T G(s, t) \varphi_k(s) ds < \infty$, hence $A_K$ is well defined. So that with Proposition S4.2 one shall see $\|\cdot\|_\infty$ and $\|\cdot\|_2$ are equivalent in $\mathcal{S}_K$, $\tilde{\mathcal{S}}_K$ and $\mathcal{S}_K$, i.e., there exists constants $c_1, c_2$ such that $c_1 \|g\|_\infty \leq \|g\|_2 \leq c_2 \|g\|_\infty$ for all $g \in \mathcal{S}_K \cup \tilde{\mathcal{S}}_K \cup \mathcal{S}_K$.

Being a smoothness assumption, condition (C1) is not very strong. A similar condition has been adopted by Proposition 2(i) in Petersen and Müller (2016). As we will discuss in Proposition S6.2, it can be satisfied by $\|\varphi_k\|_\infty \asymp k$ and $\lambda_k \lesssim k^{-3}$. For example, take $T = [0, 1]$, then the even part of Fourier basis $\varphi_k(t) = \cos(k\pi t)$ can be used to construct the centered log-ratio transformed trajectories as $\int_T \cos(k\pi t) dt = 0$. Then one only need to pick proper eigenvalues so that $\lambda_k \lesssim k^{-3}$ to satisfy (C1) for consistency.

More illustrative examples can be constructed using Brownian motion, as Brownian motion is too rough, yet integrated Brownian motion is smooth enough to satisfy (C1).

**Example** S3.1. Let $W(t)$ be the Brownian motion on $t \in T = [0, 1]$, to honor the integrate-to-zero constraint resulted from $\psi$, we consider the centered Brownian
motion $Z$ and centered integrated Brownian motion $Y$. On $t \in T = [0,1]$, define

$$Z(t) = W(t) - \int_T W(s)ds,$$  \hspace{1cm} (S3.2)

$$Y(t) = \int_0^t W(s)ds - \int_{0 \leq u < v \leq 1} W(u)dudv,$$ \hspace{1cm} (S3.3)

where the integral part is used to center the process such that $\int_T Z = \int_T Y = 0$, similar to the centered log-ratio transformed trajectory $\psi_p$. Then using results from Brownian motion, we have

$$\text{cov}(Z(s), Z(t)) = s \wedge t - s(2 - s)/2 - t(2 - t)/2 + 1/3,$$

$$\text{cov}(Y(s), Y(t)) = \frac{1}{2} (s \wedge t)^2 \left( s \vee t - \frac{s \wedge t}{3} \right) + \frac{1}{20}$$

$$\left(-t^2/4 - t^3/6 + t^4/24) - (s^2/4 - s^3/6 + s^4/24), \right.$$ leading to two eigen problem for positive $\lambda$ and orthonormal $\varphi$ for all $t \in [0,1]$ as

$$\int_0^1 \text{cov}(Z(s), Z(t))\varphi(s)ds = \lambda \varphi(t),$$  \hspace{1cm} (S3.4)

$$\int_0^1 \text{cov}(Y(s), Y(t))\varphi(s)ds = \lambda \varphi(t),$$ \hspace{1cm} (S3.5)

subject to

$$\int_0^1 \varphi(s)ds = 0.$$ \hspace{1cm} (S3.6)

For (S3.4), take derivative w.r.t. $t$ iteratively leads to

$$\int_t^1 \varphi(s)ds = \lambda \varphi'(t),$$ \hspace{1cm} (S3.7)

$$-\varphi(t) = \lambda \varphi''(t).$$ \hspace{1cm} (S3.8)

Solution to (S3.8) takes the form of

$$\varphi(t) = A \sin(t/\sqrt{\lambda}) + B \cos(t/\sqrt{\lambda}), t \in [0,1]$$

for some constant $A$ and $B$. With (S3.6), taking $t = 0$ in (S3.7) leads to $A = 0$, further $t = 1$ in (S3.7) leads to $\lambda_k = (k\pi)^{-2}$ for $k \in \mathbb{N}^+$. In the end, orthonormality gives $B = \sqrt{2}$. In combine, the eigenvalues and eigenfunctions of centered Brownian motion (S3.2) are $\lambda_k = (k\pi)^{-2}$ and $\varphi_k(t) = \sqrt{2} \cos(k\pi t)$ on $t \in [0,1]$. Noting $\|\varphi_k\|_\infty \approx k$, condition (C1) is not satisfied.

For (S3.5), take derivative w.r.t. $t$ iteratively leads to

$$\int_0^t s^2 \varphi(s)ds/2 + \int_t^1 (st - t^2/2)\varphi(s)ds = \lambda \varphi'(t),$$ \hspace{1cm} (S3.9)
\[ \int_{t}^{1} (s-t) \varphi(s) ds = \lambda \varphi''(t), \quad (S3.10) \]
\[ - \int_{t}^{1} \varphi(s) ds = \lambda \varphi^{(3)}(t), \quad (S3.11) \]
\[ \varphi(t) = \lambda \varphi^{(4)}(t). \quad (S3.12) \]

Solution to (S3.12) takes the form of
\[ \varphi(t) = A \sin(\lambda^{-1/4} t) + B \cos(\lambda^{-1/4} t) + C \exp(\lambda^{-1/4} t) + D \exp(\lambda^{-1/4} t) \]
for some constant \( A, B, C, \) and \( D. \)

Under (S3.6), taking \( t = 0 \) in (S3.9) and (S3.11) leads to \(-A + C - D = 0\) and \(-A - C + D = 0\), indicating \( A = 0 \) and \( C = D \), together with (S3.10) and (S3.11) at \( t = 1 \), we have
\[ B \cos \lambda^{-1/4} = C \left( \exp(-\lambda^{-1/4}) + \exp(\lambda^{-1/4}) \right), \]
\[ B \sin \lambda^{-1/4} = C \left( \exp(-\lambda^{-1/4}) - \exp(\lambda^{-1/4}) \right). \]

For nonzero \( B \) and \( C, \lambda^{-1/4} \) must be complex number. Further since eigenvalues of covariance function must be real numbers, we can set \( \lambda^{-1/4} = \sigma \sqrt{-1}, \) which leads to
\[ B \cosh \sigma = 2C \cos \sigma, \]
\[ B \sinh \sigma = -2C \sin \sigma, \]
corresponding to \( \sigma_k > 0 \) satisfying \( \sinh \sigma_k / \cosh \sigma_k = -\sin \sigma_k / \cos \sigma_k, \) i.e., \( \sigma_1 \approx 2.36502 \) and \( \sigma_k \approx \sigma_1 + k\pi. \) Plugging the eigenvalues back, \( \| \varphi_k \|_2 = 1 \) leads to
\[ \varphi_k(t) = c_k \left( \cos(\sigma_k t) \cosh(\sigma_k t) / \cosh(\sigma_k) + \cos(\sigma_k t) \right) \]
for a sequence of constants \( c_k \to 1 \) as \( k \to \infty \) such that \( \| \varphi_k \|_2 = 1 \) for all \( k. \)

Therefore,
\[ \varphi_k'(t) = -c_k \sigma_k \sin(\sigma_k t) + c_k \sigma_k \sinh(\sigma_k t) \cos(\sigma_k t) / \cosh(\sigma_k). \]

Since \( \sinh(\sigma_k t) \cos(\sigma_k) / \cosh(\sigma_k) \to 0 \) as \( \sigma_k \to \infty \) on \( t \in [0, 1], \) we have \( \| \varphi_k' \|_\infty \approx \sigma_k \approx k \) as \( k \to \infty. \) Noting \( \lambda_k \approx \sigma_k^{-4} \approx k^{-4}, \) we have (C1) satisfied. In general, the author believe the centered \( m \)-times integrated Brownian motion should be smooth enough to satisfy (C1) for \( m \geq 1. \) \( \square \)

Under (C3), we can obtain asymptotic results for the average pre-smoothing error \( \sum_i (\hat{p}_i - p_i) / n, \) which leads to desired rates for corresponding sample mean and covariance. Condition (C3). (i) is required so that Talagrand inequality (Lemma S5.1) could provide a uniform bound for the squared-bias of pre-smoothed trajectories, while (C3). (ii) is used for adopting Bernstein’s inequality for an uniform almost sure rate of the average deviation.

24
Condition (C3) requires \( h \to 0 \) slowly. In addition, in the super dense case where \( m \to \infty \) much faster than \( n \to \infty \), (C3)(ii) might be replaced to allow smaller \( h \) for a better rate. Though this situation would be much alike that without presmoothing, and of less interest.

The asymptotic are based on two key properties for pre-smoothing errors. For ease of reference, we list them below and will proof them in Proposition S5.2.

(I) There exists positive \( b_n = o(1) \) such that

\[
\| \hat{\mu} - \tilde{\mu} \|_2 = O_p(b_n), \quad \| \hat{G} - \tilde{G} \|_2 = O_p(b_n).
\]

(II) \( \sup_{1 \leq i \leq n} \| \hat{f}_i - f_i \|_\infty = o(1) \), and there exists positive \( a_n = o(1) \) such that

\[
\| \hat{\mu} - \tilde{\mu} \|_\infty = O(a_n), \quad \| \hat{G} - \tilde{G} \|_\infty = O(a_n)
\]

almost surely.

The overall structure of the theoretical framework is as follow. Using KDE as pre-smoother, under (A1), (K1), and (C3), we can proof (I) and (II) characterizing the presmoothing errors, as in Subsection S5.1. Further, in combine with functional central limit theorem, we provide the asymptotic of sample mean and covariance, as in Proposition S5.3. Hence, with perturbation theory, we obtain errors of estimated eigenvalues and eigenfunctions, which dominate the projection errors of the log-densities to \( F_K \), as shown in Lemma S5.3 and Lemma S5.4. After these, we use an established theorem (Theorem S6.1, or theorem 3 of Barron and Sheu (1991)) to arrive at our main result of Theorem S6.2.

S4 Some Lemmas and Proofs

S4.1 Identifiability

Lemma S4.1. An exponential family with density \( p_\theta(x) = h(x) \exp(\theta^T \varphi(x) - B(\theta)) \), \( x \in \mathcal{X} \) on some domain \( \mathcal{X} \) is non-identifiable iff \( P_{\theta_0}(v^T \varphi(X) = a) = 1 \) for some \( v \neq 0, v \in \mathbb{R}^d, a \in \mathbb{R} \) and \( \theta_0 \in \mathbb{R}^d \) with \( B(\theta_0) < \infty \).

Proof. Note that the family is non-identifiable if and only if there exists distinct \( \theta_1, \theta_2 \) in the natural parameter space \( \{ \theta : B(\theta) < \infty \} \) such that \( p_{\theta_1} = p_{\theta_2} \) on \( \mathcal{X} \), which holds if and only if \( (\theta_1 - \theta_2)^T \varphi(x) = B(\theta_1) - B(\theta_2) \) on \( \{ x \in \mathcal{X} : h(x) \neq 0 \} \), we have the “only if” part.

Conversely for all \( s \in \mathbb{R} \) with \( \theta_0 + sv \) inside the natural parameter space, we have \( B(\theta_0 + sv) = \log \int_\mathcal{X} p_{\theta_0}(x) \exp(sv^T \varphi(x) + B(\theta_0)) dx = sa + B(\theta_0) \), implying \( p_{\theta_0 + sv} = p_{\theta_0} \) on \( \mathcal{X} \). \( \square \)

Now, we show that the approximating families \( P_K, \hat{P}_K, \) and \( \hat{P}_K \) as defined in (2.3), (2.17), and (2.12) are identifiable due to the orthogonality
of the sufficient statistics. Given any covariance function \( G \) supported on \( T \times T \), we define the associated integral operator \( \hat{G} : L^2(T) \to L^2(T) \) as \((\hat{G}g)(s) = \int_T G(s,t)g(t)dt, s \in T\), and define its rank to be the dimension of its image \( \{ \hat{G}g : g \in L^2(T) \} \), denoted as rank \( \hat{G} \), which may be infinity.

**Proposition S4.1.** Under (A1), \( \mathcal{P}_K \) is identifiable for \( K = 1, \ldots, \text{rank } \hat{G} \) and \( \tilde{\mathcal{P}}_K \) is identifiable for \( K = 1, \ldots, \text{rank } \hat{G} \). Moreover, \( \mathcal{P}_K \) is identifiable for \( K = 1, 2, \ldots \), \( \text{rank } \hat{G} \) as long as \( \hat{p}_1, \ldots, \hat{p}_n \) are strictly positive on \( T \).

**Proof of Proposition S4.1.** Firstly let 1 denote the constant function on \( T \) with value 1, then \( \langle \psi p, 1 \rangle = 0 \) for all \( p \in \mathcal{P} \). Hence the space \( \psi \mathcal{P} \) is perpendicular to constant functions on \( T \), and so are the mean \( \mu \) and eigenfunctions \( \varphi_k \) for \( k \leq \text{rank } G \). Indeed, as the transformed trajectories \( f = \psi p \) integrate to zero, i.e., \( \langle f, 1 \rangle = 0 \), where 1 denotes the constant function on \( T \), its expectation \( \mu \) will also satisfy such condition by definition of expectation of random elements in Hilbert space. Similar argument applies to the eigenfunctions, as \( \langle \varphi_k, 1 \rangle = \lambda_k^{-1} \langle \hat{G}\varphi_k, 1 \rangle \), and that

\[
\langle \hat{G}\varphi_k, 1 \rangle = \langle (Ef \otimes f)\varphi_k, 1 \rangle - \langle (\mu \otimes \mu)\varphi_k, 1 \rangle = E \langle f, 1 \rangle \langle f, \varphi_k \rangle - \langle \mu, 1 \rangle \langle \mu, \varphi_k \rangle = 0.
\]

Now, since the sufficient statistics \( \{ \varphi_k \}_k \) are orthonormal and \( \langle \varphi_k, 1 \rangle = 0 \), if the family is non-identifiable, under (A1) that \( p > 0 \) on \( T \), by Lemma S4.1, exists \( v = (v_1, \ldots, v_{K-1}, 1) \neq 0 \) and \( a \in \mathbb{R} \) such that

\[
v_1\varphi_1(t) + v_2\varphi_2(t) + \cdots + v_{K-1}\varphi_{K-1}(t) + \varphi_K(t) = a
\]

for all \( t \in T \) outside a null set, implying

\[
\langle \varphi_K, \varphi_K \rangle = \langle \varphi_K, a - v_1\varphi_1 - \cdots - v_{K-1}\varphi_{K-1} \rangle = 0,
\]

a contradict. Thus, \( \mathcal{P}_K \) is identifiable.

As for \( \tilde{\mathcal{P}}_K \) and \( \tilde{\mathcal{P}}_K \), it suffices to note that we always have \( \langle \psi \hat{p}_i, 1 \rangle = \langle \psi \hat{p}_i, 1 \rangle = 0 \) for all \( i = 1, \ldots, n \) whenever \( \| \psi \hat{p}_i \|_\infty < \infty \). Denote \( \hat{G} \) as the operator associated with the estimated covariance function, we have

\[
\hat{\lambda}_k \langle \hat{\varphi}_k, 1 \rangle = \langle \hat{G}\hat{\varphi}_k, 1 \rangle = \int_T \hat{G}(s,t)\hat{\varphi}_k(s)dsdt
\]

\[
= \frac{1}{n} \sum_{i=1}^n \int_T (\psi p_i(s) - \hat{\mu}(s))(\psi p_i(t) - \hat{\mu}(t))\hat{\varphi}_k(s)dsdt
\]

\[
= \frac{1}{n} \sum_{i=1}^n \langle (\psi p_i - \hat{\mu}, 1) \langle \psi p_i - \hat{\mu}, \hat{\varphi}_k \rangle \rangle = 0.
\]

Similarly \( \langle \hat{\varphi}_k, 1 \rangle = 0 \), implying the estimated sufficient statistics are also orthogonal to constant functions. Analogous to the proof for \( \mathcal{P}_K \), we shall see \( \tilde{\mathcal{P}}_K \) and \( \tilde{\mathcal{P}}_K \) are also identifiable. \( \square \)
S4.2 Equivalence of $L^2$ and $L^\infty$ norm in $S_K$

**Proposition S4.2.** Consider two symmetric and non-negative definite integral operator $G$ and $\tilde{G}$ on $L^2(T)$ with continuous kernel and eigenfunctions $(G, \varphi_k)$ and $(\tilde{G}, \tilde{\varphi}_k)$. Assume that $G$ is continuously differentiable on $T \times T$, then for all $g \in S_K$, $\tilde{g} \in \tilde{S}_K$, we have

$$\|g\|_\infty \leq A_K \|g\|_2, \quad \|\tilde{g}\|_\infty \leq (A_K + \sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_\infty) \|\tilde{g}\|_2,$$

where $S_K$ and $\tilde{S}_K$ are the spaces spanned by the first $K$ eigenfunctions, and $A_K$ defined in (S3.1).

**Proof.** For all $a \in \mathbb{R}^K$ and $a \neq 0$, let $g = \sum_{j \leq K} a_j \varphi_j$ and $\tilde{g} = \sum_{j \leq K} a_j \tilde{\varphi}_j$. Then $\|g\|_\infty \leq A_K \|g\|_2$. In detail, $\|g'\|_\infty \leq \max_{k \leq K} |a_k| \sum_{k \leq K} \|\varphi'_k\|_\infty$ where the left-hand-side of the inequality is a Lipschitz constant for $g$, hence by lemma 1 in the supplementary of Petersen and M"uller (2016),

$$\|g\|_\infty / \|g\|_2 \leq 2 \max \left( |T|^{-1/2}, (\|g'\|_\infty / \|g\|_2)^{1/3} \right),$$

where the continuous differentiability of $g$ is granted by that of eigenfunctions inherited from the covariance function. Then note that $\|g\|_2 = \|a\|_2$, $\|g'\|_\infty \leq \|a\|_\infty \sum_{k \leq K} \|\varphi'_k\|_\infty$, and that $\|a\|_\infty \leq \|a\|_2$, we have $\|g\|_\infty \leq A_K \|g\|_2$.

Further, noting $\|g\|_2 = \|\tilde{g}\|_2 = \|a\|_2 \geq \max_k |a_k|$, $\|\tilde{g} - g\|_\infty = \|\sum_k a_k (\tilde{\varphi}_k - \varphi_k)\|_\infty \leq \max_k |a_k| \sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_\infty$, we have

$$\|\tilde{g}\|_\infty \leq \|g\|_\infty + \|\tilde{g} - g\|_\infty \leq A_K \|\tilde{g}\|_2 + \|\tilde{g}\|_2 \sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_\infty.$$

☐

S5 Asymptotic Behaviors of Sample Mean, Covariance, Eigenvalues, Eigenfunctions, and Component Scores

S5.1 Pre-smoothing

We now study the errors due to pre-smoothing, beginning with introduction to some results from empirical process. The framework we use is closely related to probability inequalities in empirical processes.

Given $X_1, \ldots, X_N$ i.i.d. observations from density $q : \mathcal{X} \to \mathbb{R}^+$, define

$$q_{h,N}(t) = \frac{1}{Nh} \sum_{j=1}^N \kappa \left( \frac{t - X_j}{h} \right), \; t \in \mathbb{R} \quad (S5.1)$$
to be a KDE with bandwidth $h$.

Note that the $L^\infty$ deviation of $q_{h,N}$ to its expectation is

$$\|q_{h,N} - E q_{h,N}\|_\infty \leq \frac{1}{Nh} \sup_{g \in K} \sum_{j=1}^{N} (g(X_j) - E g(X_1)),$$

for which Talagrand (1994) provides an exponential tail inequality. We present the version of Giné and Guillou (2002) in the following lemma.

For kernel function $\kappa$, define a class of functions mapping $\mathcal{X} \subseteq \mathbb{R}$ to $\mathbb{R}$ as

$$K = \{ \kappa((t-x)/h) : t \in \mathbb{R}, h > 0 \}.$$

For example, for some fixed $t$ and $h$, the corresponding element in $K$ is $x \mapsto \kappa((t-x)/h), x \in \mathcal{X}$. Define the envelope function of the class as

$$F(x) = \sup_{g \in K} |g(x)|, x \in \mathcal{X}.$$

Suppose we endorse $K$ with some norm $\|\cdot\|$, e.g. $L^2$ norm, then we can define its covering number $N(\varepsilon, K, \|\cdot\|)$ to be the minimal number of $\varepsilon$-radius balls needed to cover $K$. We say that $K$ is a bounded measurable Vapnik–Chervonenkis (VC) class if it is separable and there exist positive constants $A, \nu$ such that for all probability measure $P$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ and $0 < \tau < 1$,

$$N]\left(\tau \|F\|_{L^2(P)}, K, L^2(P)\right) \leq \left(\frac{A}{\tau}\right)^\nu.$$

The $A$ and $\nu$ are called VC characteristics of the class $K$.

**Lemma S5.1** (Talagrand inequality). Given kernel function $\kappa$ supported on $\mathbb{R}$, suppose $\|\kappa\|_\infty < \infty$ and $K$ forms a VC-class with characteristics $A, \nu$ with envelope function $F$. Then for sequence $h \rightarrow 0, a \rightarrow 0$ in a way such that

$$\frac{[\log h]}{Nh} = O(a^2),$$

then there exists constants $N_0 \in \mathbb{N}, a_0 > 0, h_0 > 0; c_1 > 0$ only depending on $A, \nu; c_2 > 0$ only depending on $A, \nu, \|\kappa\|_\infty$, such that for all $N > N_0, a < a_0, h < h_0$, we have

$$P(\|q_{h,N} - E q_{h,N}\|_\infty > a) \leq c_1 \exp\left(-c_2 N h a^2\right). \quad (S5.2)$$

**Proof.** Inequality (S5.2) is a direct implication of corollary 2.2 of Giné and Guillou (2002), which is presented as proposition 9 of Rinaldo and Wasserman (2010), whose requirement (16) seems to be a typo, which could be replaced by

$$\sup_{\kappa \leq \kappa_0} \sup_{v \in \mathbb{R}} \int \kappa\left(\frac{t-x}{h}\right)^2 q(x)dx < \infty$$

for an arbitrary constant $c_k > 0$. \qed
The previous Talagrand inequality applies to any density \( q \in \mathcal{P}_e \), next we turn to more specific results given \( n \) i.i.d. random densities \( p_{e,1}, \ldots, p_{e,n} \) on the extended domain \( \mathcal{T}_e \). The goal of this subsection is to study how pre-smoothing would affect the estimation of mean function \( \mu \) and covariance function \( G \) of the centered log-ratio \( \psi_p \). Since \( p_i(t) = p_{e,i}(t)/\int_\mathcal{T} p_{e,i}(s)ds \) for all \( t \in \mathcal{T} \), we have identical centered log-ratios as \( \psi_{p,e,i} = \psi_{p,i} \) on \( \mathcal{T} \), and hence it suffices to work with \( p_{e,i} \). As stated previously, the purpose of introducing \( p_{e,i} \) is just to achieve better boundary performance, otherwise KDE would be biased near the boundary.

**Lemma S5.2.** Let \( g_i : \mathcal{T} \times \mathcal{T} \to \mathbb{R} \) be a sequence of differentiable random functions satisfying \( E g_i \mid p_{e,i} = g_i \) for all \( i \geq 1 \), \( \sup_n \| g_n \|_\infty \leq \infty \), and \( \sup_n \| g_n' \|_\infty \leq \infty \). Then as \( n \to \infty \),

\[
\int_\mathcal{T} \left( \frac{1}{n} \sum_{i=1}^{n} g_i(s,t) \left( \hat{p}_{e,i}(t) - E_\mathcal{T} \hat{p}_{e,i}(t) | p_{e,i} \right) \right)^2 dsdt = O_p \left( \frac{1}{nmh} \right),
\]

where

\[
\hat{p}_{e,i}(t) = \frac{1}{N_i h} \sum_{j=1}^{N_i} \kappa \left( \frac{t - X_{ij}}{h} \right) \quad (S5.3)
\]

is KDE of i.i.d. sample \( X_{i1}, \ldots, X_{iN_i} \) from density \( p_{e,i} \) with kernel \( \kappa \) which is finite and Lipschitz.

Moreover, if \( n \to \infty, m \to \infty, \) and \( h \to 0 \) in a way satisfying (C3). (ii), we have

\[
\sup_{s,t \in \mathcal{T}} | \frac{1}{n} \sum_{i=1}^{n} g_i(s,t) \left( \hat{p}_{e,i}(t) - E_\mathcal{T} \hat{p}_{e,i}(t) | p_{e,i} \right) | = O \left( \sqrt{\left( \frac{1}{n} + \frac{1}{nmh} \right) \log n} \right),
\]

almost surely.

**Proof.** We first show the uniform part, which is an application of Bernstein inequality and Lipschitz continuity. Similar idea can be found in Appendix D of Zhang and Wang (2016).

Denote

\[
V_i(s,t) = \sum_{j=1}^{N_i} \frac{g_i(s,t)}{n_i h} U_{ij}(t),
\]

where

\[
U_{ij}(t) = \kappa \left( \frac{t - X_{ij}}{h} \right) - E_\mathcal{T} \kappa \left( \frac{t - X_{ij}}{h} \right) | p_{e,i} |,
\]

and let \( V(s,t) = \sum_{i \leq n} V_i(s,t) \). Let \( \Pi = \{t_1 < t_2 < \ldots\} \) denote a partition of \( \mathcal{T} \) and its gap \( \| \Pi \| = \inf \{ |s - t| : s, t \in \Pi, s \neq t \} \). We can see that

\[
\sup_{s,t \in \mathcal{T}} |V(s,t)|
\]
\[
\leq \sup_{s, t \in \Pi} |V(s, t)| + \sup_{|s_1 - s_2| \vee |t_1 - t_2| \leq \|\Pi\|} |V(s_1, t_1) - V(s_2, t_2)|. \tag{S5.4}
\]

Notice for \(s_1, s_2, t_1, t_2 \in T\) with \(|s_1 - s_2| \vee |t_1 - t_2| \leq \|\Pi\|\),
\[
|V_i(s_1, t_1) - V_i(s_2, t_2)| = \left| \frac{1}{nN_i h} \sum_{j=1}^{N_i} \left( g_i(s_1, t_1)(U_{ij}(t_1) - U_{ij}(t_2)) + U_{ij}(t_2)(g_i(s_1, t_1) - g_i(s_2, t_2)) \right) \right| \leq \frac{2}{nh} \left( \sup_{i \leq n} \|g_i\|_\infty \|\kappa'\|_\infty \frac{|t_1 - t_2|}{h} + \sup_{i \leq n} \|g_i'\|_\infty \|\kappa\|_\infty (|s_1 - s_2| + |t_1 - t_2|) \right),
\]
where with some abuse of notation, \(\|\kappa'\|_\infty\) is the Lipschitz constant for \(\kappa\). Therefore the second term on the RHS of (S5.4) is \(O\left(\frac{\|\Pi\|}{nh^2}\right)\) almost surely.

As for the first term on the RHS of (S5.4), note that \(EV_i(s, t) = EE(V_i(s, t)p_{e,i}) = 0\), and that
\[
|V_i(s, t)| \leq 2 \sum_{j=1}^{N_i} \frac{\|g_i\|_\infty \|\kappa\|_\infty}{nN_i h} = \frac{2M \|\kappa\|_\infty}{nh},
\]
where \(M\) is a constant with \(\sup_{n} \|g_n\|_\infty \leq M\). Moreover, by independence,
\[
EV_i(s, t)^2 = E \left( \frac{g_i(s, t)}{nN_i h} \right)^2 \sum_j \var \left[ \kappa \left( \frac{t - X_{ij}}{h} \right) \right] p_{e,i} \\
\leq E \left( \frac{g_i(s, t)}{nN_i h} \right)^2 \sum_j \int \kappa \left( \frac{t - x}{h} \right)^2 p_{e,i}(x) dx \\
= E \left( \frac{g_i(s, t)}{nN_i h} \right)^2 h \sum_j \int \kappa(u)^2 p_{e,i}(t - uh) du \\
\leq \frac{M^2 \|\kappa\|_\infty^2}{n^2 mh}. \tag{S5.5}
\]
Therefore, for any \(s, t \in T\), by Bernstein’s inequality, for any \(\gamma > 0\), we have
\[
P \left( \left| \sum_{i=1}^{n} V_i(s, t) \right| > \gamma \right) \leq 2 \exp \left( -\frac{1}{2} \gamma^2 \left( \frac{c_1}{nmh} + \frac{c_2}{nh} \gamma \right) \right),
\]
where \(c_1 = M^2 \|\kappa\|_\infty^2, c_2 = 2M \|\kappa\|_\infty / 3\). Take
\[
\gamma = \gamma_n = c_3 \sqrt{\frac{1}{n} + \frac{1}{nmh}} \log n,
\]
30
for arbitrary $c_3 > 0$, notice that given any sufficiently large $n, m$ and small $h$, (C3). (ii) leads to $1/(nh) \leq c_0 \gamma/(c_3 \log n)$ for some constant $c_0 > 0$, we have

$$
P \left( \left| \sum_{i=1}^{n} V_i(s, t) \right| > \gamma \right) \leq 2 \exp \left( -\frac{1}{2} \gamma^2 / \left( \frac{c_1}{nmh} + \frac{c_0 c_2 \gamma^2}{c_3 \log n} \right) \right)
$$

$$
\leq 2 \exp \left( -\frac{1}{2} \frac{c_1^2 \log n}{c_3} / (c_1 + c_0 c_2 c_3) \right).
$$

Then if we take $\|\Pi_n\| \asymp n^{-c_4}$, we have

$$
P \left( \sup_{s, t \in \Pi_n} \left| \sum_{i=1}^{n} V_i(s, t) \right| > \gamma \right) \leq 2 \exp \left( 2c_4 \log n - \frac{c_3^2}{2c_2} - \frac{c_3}{2c_2 + 2c_0 c_2 c_3} \log n \right),
$$

which is summable w.r.t. $n$ for sufficiently large $c_3$ according to $c_4$. Thus by Borel–Cantelli lemma,

$$
\sup_{s, t \in \Pi_n} \left| \sum_{i=1}^{n} V_i(s, t) \right| = O \left( \sqrt{\left( \frac{1}{n} + \frac{1}{nmh} \right) \log n} \right)
$$

almost surely. Meanwhile, we can take $c_4$ large enough so that $\|\Pi\|/h^2 = O(\gamma)$, making the RHS of (S5.4) dominated by its first terms, which lead to the desired result.

As for the $L^2$ result, note that by independence, $EV_i = 0$, and (S5.5), we have

$$
E \left\| \sum_{i=1}^{n} V_i \right\|_2^2 = \sum_{i=1}^{n} \|EV_i\|_2^2 \leq \frac{c_1}{nmh},
$$

then Markov’s inequality leads to the desired result. $\square$

To bridge $p_t$ restricted on $\mathcal{T}$ and $p_{c,i}$ on $\mathcal{T}_c$, recall that the centered log-ratio $\psi$ is defined on $\mathcal{T}$ to be invariant to multiplying constants, and that $p_t(t) = p_{c,i}(t) / \int_{\mathcal{T}} p_{c,i}(s) ds$ for all $t \in \mathcal{T}$. Therefore, the trajectories after transformation are identical as $f_i = \psi p_t = \psi p_{c,i}$ and $\tilde{f}_i = \psi \tilde{p}_t = \psi \tilde{p}_{c,i}$. In particular, we write $\tilde{p}_t(t) = \tilde{p}_{c,i}(t) / \int_{\mathcal{T}} \tilde{p}_{c,i}(s) ds$, which is equivalent to (2.11) of main text applied to $X_1, \ldots, X_N$, drawn from $p_{c,i}$, justifying reusing its notation.

**Proposition S5.1.** Under (A1), (K1), (C3). (i), and conditions of Lemma S5.2, we have

$$
\int_{\mathcal{T}} \left( \frac{1}{n} \sum_{i=1}^{n} g_i(s, t) (\tilde{f}_i(t) - f_i(t)) \right)^2 ds dt = O_p \left( h^4 + \frac{1}{nmh} + \frac{1}{m^2h^2} \right).
$$

If in addition, (C3). (ii) also holds, then

$$
\sup_{s, t \in \mathcal{T}} \left| \frac{1}{n} \sum_{i=1}^{n} g_i(s, t) (\tilde{f}_i(t) - f_i(t)) \right| = O \left( h^2 + \sqrt{\frac{1}{n} + \frac{1}{nmh}} \log n + \frac{\log n}{mh} \right)
$$

almost surely, where $\tilde{f}_i(t) = \psi \tilde{p}_t(t) = \log \tilde{p}_t(t) - \int_{\mathcal{T}} \log \tilde{p}_i(s) ds$, $\tilde{p}_t(t) = \tilde{p}_{c,i}(t) / \int_{\mathcal{T}} \tilde{p}_{c,i}(s) ds$, and $\tilde{p}_{c,i}$ is defined in (S5.3), for $i = 1, \ldots, n$. 

31
Proof. Note that \( \tilde{T}_i = \psi \hat{p}_{e,i} \) by the nature of centered log-ratio transformation \( \psi \). Further, since \( T \) is a compact set, it suffices to consider the differences in log-densities. Note that by Taylor expansion,

\[
\frac{1}{n} \sum_{i=1}^{n} g_i(s, t) (\log \hat{p}_{e,i}(t) - \log p_{e,i}(t)) = \frac{1}{n} \sum_{i=1}^{n} \frac{g_i(s, t)}{\hat{p}_{e,i}(t)} (\hat{p}_{e,i}(t) - p_{e,i}(t)) - \frac{1}{2n} \sum_{i=1}^{n} \frac{g_i(s, t)}{\xi_i(t)} (\hat{p}_{e,i}(t) - p_{e,i}(t))^2 \\
= \frac{1}{n} \sum_{i=1}^{n} \frac{g_i(s, t)}{p_{e,i}(t)} (\hat{p}_{e,i}(t) - p_{e,i}(t)) + \frac{1}{n} \sum_{i=1}^{n} \frac{g_i(s, t)}{p_{e,i}(t)} (E \hat{p}_{e,i}(t)|p_{e,i}) (\hat{p}_{e,i}(t) - p_{e,i}(t)) \\
- \frac{1}{2n} \sum_{i=1}^{n} \frac{g_i(s, t)}{\xi_i(t)} (\hat{p}_{e,i}(t) - p_{e,i}(t))^2 \\
=: T_{n1}(s, t) + T_{n2}(s, t) + T_{n3}(s, t)
\]

for some \( \xi_i(t) \) lying between \( \hat{p}_{e,i}(t) \) and \( p_{e,i}(t) \) for all \( i = 1, \ldots, n \).

Under boundedness of \( \sup_i \|1/p_{e,i}\|_\infty \) from (A1) and Lemma S5.2, we have

\[
\sup_{s, t \in T} |T_{n1}(s, t)| = O \left( \sqrt{\left( \frac{1}{n} + \frac{1}{nmh} \right) \log n} \right) \quad \text{(S5.6)}
\]

almost surely under (C3). (ii).

Moreover, note that because density \( p_{e,i} \) are assumed to be supported on a slightly larger domain \( (a - \varepsilon, b + \varepsilon) \), then for all \( t \in T = [a, b] \), write \( \text{bias } \hat{p}_{e,i}(t) = E [\hat{p}_{e,i}(t)|p_{e,i}] - p_{e,i}(t) \), we have

\[
- \text{bias } \hat{p}_{e,i}(t) = p_{e,i}(t) - \frac{1}{\hbar} \int_{a-\varepsilon}^{b+\varepsilon} \kappa \left( \frac{t - x}{\hbar} \right) p_{e,i}(x) dx \\
= p_{e,i}(t) - \int I_h(u) \kappa(u) p_{e,i}(t - uh) du \\
= p_{e,i}(t) - p_{e,i}(t) \int I_h(u) \kappa(u) du - h \int I_h(u) \kappa(u) du + \frac{1}{2} \int I_h(u) u^2 \kappa(u) du,
\]

where

\[
I_h(u) = 1 \{ t - b - \varepsilon \leq uh \leq t - a + \varepsilon \}
\]

is the indicator function for integration domain. Note that given sufficiently small \( h > 0 \), \( I_h(u) = 1 \) over the support of \( \kappa \). Therefore combining (A1) and (K1), \( \| \text{bias } \hat{p}_{e,i} \|_\infty \leq h^2 M \int u^2 \kappa(u) du / 2 \) for all \( i = 1, \ldots, n \), which implies \( \sup_{1 \leq i \leq n} \| \text{bias } \hat{p}_{e,i} \|_\infty = O(h^2) \) almost surely. Hence

\[
\sup_{s, t \in T} |T_{n2}(s, t)| = O(h^2) \quad \text{(S5.7)}
\]

32
almost surely.

To handle $T_{n3}$, note that under (K1), conditions for Lemma S5.1 are satisfied (see e.g., lemma 22 of Nolan and Pollard (1987); condition (K.iii) of Einmahl and Mason (2005), the bounded variation ensures $\{\kappa((t - \cdot)/h) : t \in \mathbb{R}, h > 0\}$ forms a VC class), and thus if we take

$$a = c_3 \sqrt{\log n / mh}$$

for an arbitrary constant $c_3$, and set $|\log h| = O(\log n)$ as assumed in (C3). (i), then Lemma S5.1 implies

$$P( \sup_{1 \leq i \leq n} \|\hat{p}_{e,i} - E[\hat{p}_{e,i}|p_{e,i}]\|_\infty > a) \leq c_1 \exp ((1 - c_2 c_3^2) \log n),$$

which is summable w.r.t. $n$ with a large enough $c_3$. Thus by Borel–Cantelli lemma,

$$\sup_{1 \leq i \leq n} \sup_{t \in T} |\hat{p}_{e,i}(t) - p_{e,i}(t)| = O\left( h^2 + \sqrt{\log n / mh} \right).$$

This also implies that $\sup_{1 \leq i \leq n} \|1/\hat{p}_{e,i}\|_\infty = O(1)$ almost surely thanks to (A1), hence $\sup_{1 \leq i \leq n} \|1/\xi_i\|_\infty = O(1)$ almost surely as well. In combine,

$$\sup_{s,t \in T} \|T_{n3}(s, t)\| = O\left( h^4 + \log n / mh \right).$$

Combining the rates from (S5.6), (S5.7), and (S5.10) leads to the $L^\infty$ rate as claimed under (C3). (i) and (C3). (ii).

Next, for the $L^2$ rate, note that Lemma S5.2 indicates $\|T_{n1}\|_2 = O_p(1/\sqrt{nmh})$, thus it suffices to show $\|T_{n3}\|_2 = O_p(h^4 + 1/mh)$.

Following the previous $L^\infty$ result for $T_{n2}$, the bias is uniformly small, hence we have $\sup_{1 \leq i \leq n} \|g_i/\xi_i\|_\infty = O(1)$ almost surely, and that for any $s \in T$, we have

$$\|T_{n3}(s, \cdot)\|_2 \leq \left\| \frac{1}{n} \sum_{i=1}^{n} g_i(s, \cdot) / \xi_i \right\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \|g_i/\xi_i\|_\infty \sup_{1 \leq i \leq n} \|\text{bias} \hat{p}_{e,i}\|_2^2 + O(h^4).$$
Also note that by triangular inequality and Jensen’s inequality,

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{g_i(s, \cdot)}{\xi_i}(\hat{p}_{e,i} - E\hat{p}_{e,i} | p_{e,i})^2 \right\|_2 \leq \left( \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{g_i(s, \cdot)}{\xi_i}(\hat{p}_{e,i} - E\hat{p}_{e,i} | p_{e,i})^2 \right\|_2 \right)^2 \leq \frac{1}{n} \sum_{i=1}^{n} \left\| \frac{g_i(s, \cdot)}{\xi_i}(\hat{p}_{e,i} - E\hat{p}_{e,i} | p_{e,i})^2 \right\|_2^2 \leq \frac{1}{n} \sup_{1 \leq i \leq n} \|g_i/\xi_i\|_\infty^2 \sum_{i=1}^{n} \|\hat{p}_{e,i} - E\hat{p}_{e,i} | p_{e,i}\|_4^4.
\]

Therefore, for any \( R > 0 \) and sufficiently large \( B \),

\[
P \left( \int \left( \frac{1}{n} \sum_{i=1}^{n} \frac{g_i(s, t)}{\xi_i(t)}(\hat{p}_{e,i}(t) - E\hat{p}_{e,i}(t) | p_{e,i})^2 \right)^2 ds dt > \frac{R^2}{m^2 h^2} \right) \leq P \left( \frac{1}{n} \sup_{1 \leq i \leq n} \|g_i/\xi_i\|_\infty^2 \sum_{i=1}^{n} \|\hat{p}_{e,i} - E\hat{p}_{e,i} | p_{e,i}\|_4^4 > \frac{R^2}{m^2 h^2} \right) \leq P \left( \sum_{i=1}^{n} \|\hat{p}_{e,i} - E\hat{p}_{e,i} | p_{e,i}\|_4^4 > \frac{nR^2}{m^2 h^2 B^2} \right) + P \left( \sup_{1 \leq i \leq n} \|g_i/\xi_i\|_\infty > B \right) \leq \frac{m^2 h^2 B^2}{nR^2} \sum_{i=1}^{n} E \|\hat{p}_{e,i} - E\hat{p}_{e,i} | p_{e,i}\|_4^4 + o(1),
\]

where the last line is due to Markov’s inequality and independence of \( \hat{p}_{e,i} \). Under (K1), \( \kappa \) is bounded, thus for any \( t \in T \), we have

\[
E \left[ (\hat{p}_{e,i}(t) - E[\hat{p}_{e,i}(t)|p_{e,i}])^4 | p_{e,i} \right]
= E \left[ (N_i h)^{-4} N_i \left( \frac{t - X_{i1}}{h} \right)^4 \right] E \left[ \kappa \left( \frac{t - X_{i1}}{h} \right) | p_{e,i} \right] + \frac{N_i(N_i - 1)}{N_i^4 h^4} \left( \text{var} \left[ \kappa \left( \frac{t - X_{i1}}{h} \right) | p_{e,i} \right] \right)^2 \leq N_i^{-3} h^{-4} (2 \|\kappa\|_\infty)^2 \text{var} \left[ \kappa \left( \frac{t - X_{i1}}{h} \right) | p_{e,i} \right] + \frac{N_i - 1}{N_i^4 h^4} \left( \text{var} \left[ \kappa \left( \frac{t - X_{i1}}{h} \right) | p_{e,i} \right] \right)^2.
\]

Note that for sufficiently small \( h > 0 \),

\[
\sup_{1 \leq i \leq n} \sup_{t \in T} \text{var} \left[ \kappa \left( \frac{t - X_{i1}}{h} \right) | p_{e,i} \right] \leq h \left( \|p\|_\infty^2 + h \|p'\|_\infty \int u \kappa(u)^2 du \right)
\]

34
= O(h)

almost surely, thus

$$\sum_{i=1}^{n} E \left\| \tilde{p}_{e,i} - E \tilde{p}_{e,i} \right\|_4^4 = \sum_{i=1}^{n} \int \left( E \left( (\tilde{p}_{e,i}(t) - E \tilde{p}_{e,i}(t)) \left\| \tilde{p}_{e,i} \right\|_4^4 \right) \right) dt$$

$$\lesssim \frac{n}{m^2 h^2},$$

implying that $$\|T_n\|_2 = O_p(h^4 + 1/mh)$$, which concludes the proof.

The previous proposition gives the following convergence results for the sample mean and covariance listed as (I) and (II), on which all the asymptotic involving pre-smoothed trajectories are based.

**Proposition S5.2.** Under (A1), (K1), and (C3). (i),

(I) we have positive $$b_n = o(1)$$ such that $$\|\hat{\mu} - \tilde{\mu}\|_2 = O_p(b_n)$$, and $$\|\hat{G} - \tilde{G}\|_2 = O_p(b_n)$$.

If in additional (C3). (ii) also holds, then

(II) we have $$\sup_{1 \leq i \leq n} \left\| \tilde{f}_i - f_i \right\|_\infty = O \left( h^2 + \sqrt{\log(n)/(mh)} \right) = o(1)$$, and $$a_n = o(1)$$ such that $$\|\hat{\mu} - \tilde{\mu}\|_\infty = O(a_n)$$, and $$\|\hat{G} - \tilde{G}\|_\infty = O(a_n)$$ almost surely,

where

$$a_n = a_n(m, h) = h^2 + \sqrt{\frac{1}{n} + \frac{1}{nmh}} \log n + \frac{\log n}{mh},$$

$$b_n = b_n(m, h) = h^2 + 1/\sqrt{nmh} + 1/mh.$$

**Proof.** Under (A1) we have

$$\sup_{1 \leq i \leq n} \sup_{t \in T} |\tilde{f}_i - f_i| \asymp \sup_{1 \leq i \leq n} \sup_{t \in T} |\tilde{p}_{e,i} - p_{e,i}|$$

almost surely, then use (S5.9) to see

$$\sup_{1 \leq i \leq n} \left\| \tilde{f}_i - f_i \right\|_\infty = O \left( h^2 + \sqrt{\log(n/mh)} \right) = o(1)$$

almost surely. Also, Proposition S5.1 implies $$\|\hat{\mu} - \tilde{\mu}\|_2 = O_p(b_n)$$ and $$\|\hat{\mu} - \tilde{\mu}\|_\infty = O(a_n)$$ almost surely. For the covariance function, note that

$$\hat{G}(s, t) - \tilde{G}(s, t)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left( \tilde{f}_i(s)\tilde{f}_i(t) - f_i(s)f_i(t) \right) - (\hat{\mu}(s)\hat{\mu}(t) - \tilde{\mu}(s)\tilde{\mu}(t))$$
\[
\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^{n} (\tilde{f}_i(s) - f_i(s)) (\tilde{f}_i(t) - f_i(t)) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} f_i(t) (\tilde{f}_i(s) - f_i(s)) + \frac{1}{n} \sum_{i=1}^{n} f_i(s) (\tilde{f}_i(t) - f_i(t)) \\
&\quad - \mu(s) (\tilde{\mu}(t) - \tilde{\mu}(t)) - \tilde{\mu}(t) (\tilde{\mu}(s) - \tilde{\mu}(s)).
\end{aligned}
\]

Use \( Q_{n1}, Q_{n2}, \) and \( Q_{n3} \) to denote the terms in each row of the last equality. Following steps for \( T_{n3} \) in the proof of Proposition S5.1, we have

\[
\sup_{s,t \in T} |Q_{n1}(s,t)| \leq \sup_{1 \leq i \leq n, t \in T} |\tilde{f}_i(t) - f_i(t)|^2 \\
\quad \leq \sup_{1 \leq i \leq n, t \in T} |\tilde{p}_{e,i}(t) - p_{e,i}(t)|^2 \\
= O \left( h^4 + \frac{\log n}{mh} \right)
\]

almost surely. Also, by Proposition S5.1,

\[
\sup_{s,t \in T} |Q_{n2}(s,t)| = O(a_n)
\]

almost surely. In the end, note that \( \|	ilde{\mu}\|_{\infty} \approx \|\tilde{\mu}\|_{\infty} \approx 1 \) almost surely, applying Proposition S5.1 again gives the desired result for the \( L^\infty \) error.

As for \( \|\tilde{G} - G\|_2 \), Proposition S5.1 implies \( \|Q_{n2} + Q_{n3}\|_2 = O_p(h^2 + 1/\sqrt{nmh} + 1/(mh)) \), while

\[
\|Q_{n1}\|_2 \leq \frac{1}{n} \sum_{i=1}^{n} \| (\tilde{f}_i - f_i) \otimes (\tilde{f}_i - f_i) \|_2 \\
\quad \leq \frac{1}{n} \sum_{i=1}^{n} \| \tilde{p}_{e,i} - p_{e,i} \|_2^2 \\
\quad \leq \sup_{1 \leq i \leq n} \| \text{bias } \tilde{p}_{e,i} \|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \| \tilde{p}_{e,i} - E[\tilde{p}_{e,i}] \|_2^2 \\
\quad \leq \| \text{bias } \tilde{p}_{e,i} \|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \| \tilde{p}_{e,i} - E[\tilde{p}_{e,i}] \|_2^2 \\
\quad \leq \| \text{bias } \tilde{p}_{e,i} \|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \| \tilde{p}_{e,i} - E[\tilde{p}_{e,i}] \|_2^2 \\
\quad \leq \| \text{bias } \tilde{p}_{e,i} \|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \| \tilde{p}_{e,i} - E[\tilde{p}_{e,i}] \|_2^2 \\
\quad \leq \| \text{bias } \tilde{p}_{e,i} \|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \| \tilde{p}_{e,i} - E[\tilde{p}_{e,i}] \|_2^2 \\
\quad \leq \| \text{bias } \tilde{p}_{e,i} \|_2^2 + \frac{1}{n} \sum_{i=1}^{n} \| \tilde{p}_{e,i} - E[\tilde{p}_{e,i}] \|_2^2 \leq h^4 + 1/(mh),
\]

where the last line is due to Markov’s inequality. In combine we have the desired results. \( \square \)

**S5.2 Limiting Behaviors**

We have the following asymptotic characterization of the sample mean and covariance, with or without pre-smoothing.

**Proposition S5.3.** Under (A1), we have

\[
\|\tilde{\mu} - \mu\|_2 = O_p(1/\sqrt{n}), \quad \|\tilde{G} - G\|_2 = O_p(1/\sqrt{n}),
\]

36
and that
\[
\|\hat{\mu} - \mu\|_\infty = O\left(\sqrt{\log n \over n}\right) \text{ a.s.,} \quad \|\hat{G} - G\|_\infty = O\left(\sqrt{\log n \over n}\right) \text{ a.s.,}
\]
\[
\|\hat{\mu} - \mu\|_\infty = O_p\left(1/\sqrt{n}\right), \quad \|\hat{G} - G\|_\infty = O_p\left(1/\sqrt{n}\right).
\]
as \(n \to \infty\).

Moreover, under additional assumptions (K1) and (C3). (i), using the centered log-ratio transformation \(\psi\), we have
\[
\|\hat{\mu} - \mu\| = O_p\left(1/\sqrt{n} + b_n \right), \quad \|\hat{G} - G\| = O_p\left(1/\sqrt{n} + b_n \right).
\]

If in addition, (C3). (ii) also holds, then
\[
\|\hat{\mu} - \mu\|_\infty = O\left(\sqrt{\log n \over n} + a_n\right) \text{ a.s.,} \quad \|\hat{G} - G\|_\infty = O\left(\sqrt{\log n \over n} + a_n\right) \text{ a.s.,}
\]
\[
\|\hat{\mu} - \mu\|_\infty = O_p\left(1/\sqrt{n} + a_n\right), \quad \|\hat{G} - G\|_\infty = O_p\left(1/\sqrt{n} + a_n\right),
\]
where
\[
a_n = a_{n,m,h} = h^2 + \sqrt{\left(1 + 1/nm\right) \log n + \log n \over mh},
\]
\[
b_n = b_{n,m,h} = h^2 + 1/\sqrt{nmh} + 1/mh.
\]

**Proof.** To show the results for sample mean \(\hat{\mu}\) and covariance \(\hat{G}\) without pre-smoothing, follow similar steps as in Lemma S5.2. Note that the transformed densities \(f_i\) are uniformly bounded below and above, and being Lipschitz continuous. Without loss of generality, we write
\[
\|f_i\|_\infty \leq M
\]
for all \(s, t \in T, i = 1, \ldots, n\) and some constant \(M > 0\). Moreover, such boundedness and Lipschitz continuity will apply to \(\hat{\mu}\) and \(\mu\).

Let \(\Pi = \{t_1 < t_2 < \ldots\}\) denote a partition of \(T\) and its gap \(\|\Pi\| = \inf \{|s - t| : s, t \in \Pi, s \neq t\}\). To proof the asymptotic rate for sample mean, note that
\[
\sup_{t \in T} |\hat{\mu}(t) - \mu(t)| \leq \sup_{t \in \Pi} |\hat{\mu}(t) - \mu(t)| + \sup_{|t_1 - t_2| \leq \|\Pi\|} |\mu(t_1) - \mu(t_2)| + \sup_{|t_1 - t_2| \leq \|\Pi\|} |\mu(t_1) - \mu(t_2)|,
\]
\[37\]
where the last two terms will be $O(\|\Pi\|)$ almost surely due to Lipschitz continuity. As for the first term, by Bernstein’s inequality, for any $\gamma > 0$, we have

$$\mathbb{P}\left(\sup_{t \in \Pi} |\hat{\mu}(t) - \mu(t)| > \gamma \right) \leq \|\Pi\| \cdot \sup_{t \in \Pi} \mathbb{P}(\sum_{i=1}^{n} f_i(t) - Ef_i(t) > \gamma) \leq \|\Pi\| \cdot 2 \exp\left(-\frac{\gamma\gamma^2}{2} / \left(M^2 + \frac{2}{3} M\gamma\right)\right).$$

Taking $\gamma = c_3 \sqrt{\log n}/\sqrt{n}$, and finer partition so that $\|\Pi_n\| \approx n^{-c_4}$ for some constant $c_3, c_4 > 0$, as $n \to \infty$, we have

$$\mathbb{P}\left(\sup_{t \in \Pi} |\hat{\mu}(t) - \mu(t)| > \gamma \right) \lesssim \exp\left(c_4 \log n - \frac{c_3^2 \log n}{M^2 + 2c_3 M/3}\right),$$

which is summable w.r.t. $n \to \infty$ given proper $c_3, c_4$. Thus by Borel–Cantelli lemma,

$$\sup_{t \in \Pi_n} |\hat{\mu}(t) - \mu(t)| = O\left(\sqrt{\log n}/n\right)$$

almost surely as $n \to \infty$. Combining with the rate derived from Lipschitz continuity leads to the desired result for sample mean. The asymptotic for sample covariance can be obtained using similar steps.

As for the $L^2$ error, apply continuous mapping theorem and functional C.L.T. on random elements in Hilbert space $L^2(T)$ or $L^2(T^2)$, see section 8.1 of Hsing and Eubank (2015). The results related to pre-smoothed trajectories are direct implication of (I) and (II) as in Proposition S5.2.

**Remark.** If an almost sure rate is desired for the $L^2$ error, it suffice to treat $f_i$ as random elements in Hilbert space $L^2(T)$ and apply Bernstein inequality (e.g. Bosq (2000) theorem 2.6, (2)), and then $\|\hat{\mu} - \mu\|_2 \approx \|\tilde{G} - G\|_2 = O(\sqrt{\log n}/\sqrt{n})$ almost surely. Also, notice that (C3) requires $mh \to \infty$, and hence $1/(nhm) = o(1/n)$, thus we can drop the $\sqrt{\log n/(nmh)}$ and $1/\sqrt{nmh}$ terms in $a_n$ and $b_n$.

S5.3 Perturbation Theory

With the asymptotics of the mean function, to assess the perturbation of eigenvalues and eigenfunctions while altering the operator, some results from perturbation theory for self-adjoint compact operators on Hilbert space are necessary and presented below. A detailed treatment can be found in Chapter 4, 5 of.
Hsing and Eubank (2015). It also turns out to be useful while handling the approximating errors \( \| f - f_K \|_d \) for \( d = 2, \infty \) with \( f_K(x) = \mu(x) + \eta^T \varphi(x) \).

Under (A1), \( \sup_{q \in P} E \| \psi q \|_2^2 < \infty \), hence the covariance operator for the random element \( \psi p \) is a non-negative definite self-adjoint Hilbert–Schmidt operator (Hsing and Eubank 2015, section 7.2), and hence is compact (Hsing and Eubank 2015, theorem 4.4.3). By theorem 4.2.8 of Hsing and Eubank (2015), for two compact self-adjoint operator \( G \) and \( \tilde{G} \) on some Hilbert space with eigenvalues \( \lambda_k \) and \( \tilde{\lambda}_k \), \( k = 1, 2, \ldots \), we have

\[
\sup_{k \geq 1} |\tilde{\lambda}_k - \lambda_k| \leq \| \Delta \|,
\]

where \( \Delta = \tilde{G} - G \) and \( \| \Delta \| \) is the induced operator norm. As for eigenfunctions, we have

**Theorem S5.1.** For compact self-adjoint operator \( G : L^2 \to L^2 \) on Hilbert space \( L^2 \), let \( \lambda_k \) be one of its eigenvalues with multiplicity one and corresponding eigenvector \( \varphi_k \). Let \( \tilde{G} : L^2 \to L^2 \) be another compact self-adjoint operator and suppose \( (\tilde{\lambda}_k, \tilde{\varphi}_k) \) correspond to \( (\lambda_k, \varphi_k) \) with \( \langle \tilde{\varphi}_k, \varphi_k \rangle \geq 0 \). Denote \( \Delta = \tilde{G} - G \), \( \delta_k = \inf_{l \neq k} |\lambda_k - \lambda_l|/2 \).

(i). \( \text{Bosq 2000, Lemma 4.3} \)

\[
\| \tilde{\varphi}_k - \varphi_k \|_2 \leq \sqrt{2} \delta_k^{-1} \| \Delta \|.
\]

(ii). \( \text{Hsing and Eubank 2015, Theorem 5.1.8} \)

\[
\left\| \tilde{\varphi}_k - \varphi_k \right\|_2 \leq \| \Delta \| \left( 2 \delta_k \right)^{-1} + r_p(\zeta_k) \zeta_k^2,
\]

for some non-decreasing continuous function \( r_p \) on \([0, 1/2]\).

(iii). If the associated kernel function \( G \) is continuously differentiable, then

\[
\left\| \tilde{\varphi}_k - \varphi_k \right\|_\infty \leq \frac{\hat{\lambda}^{-1}}{T} \left\| G \right\|_\infty \left\| \tilde{\varphi}_k - \varphi_k \right\|_2 \left( \left\| G \right\|_\infty \left\| \tilde{\varphi}_k - \varphi_k \right\|_2 + \left\| G - \tilde{G} \right\|_\infty \right)
\]

\[
+ \frac{\hat{\lambda}^{-1}}{T} \left\| \varphi_k \right\|_\infty \left\| \Delta \right\|.
\]

Proof. Equation (S5.12) is a direct result of Lemma 4.3 in Bosq (2000). For the second result in \( L^2 \) norm, use Theorem 5.1.8 of Hsing and Eubank (2015), note that in (5.27) of Hsing and Eubank (2015), \( \| \delta_k \| = \left\| \sum_{l \neq k} \mathcal{P}_l/(\lambda_l - \lambda_k) \right\| \leq (2\delta_k)^{-1} \) where \( \mathcal{P}_l \) is the projection to the span of \( \varphi_l \) with operator norm 1.

For the \( L^\infty \) norm, following the proof of Lemma 1 in the supplementary of Müller and Yao (2008), for all \( t \in T \), we have

\[
|\tilde{\varphi}_k(t) - \varphi_k(t)|
\]
we thus have the desired result.

Note that with additional requirement, (S5.14) gives a tighter bound than (S5.12), though asymptotically they are equivalent, at least in the situation discussed in this paper. Moreover, note that by Proposition S4.2, \(\|\varphi_k\|_\infty \leq A_k \|\varphi_k\|_2 = A_k \leq A_K\), we can replace \(\|\varphi_k\|_\infty\) by \(A_K\) for all \(k \leq K\). Thus, we have the following result.

**Corollary S5.1.** For compact self-adjoint operator \(G\), \(\hat{G}\) associated with kernel functions \(G\), \(\hat{G}\) on some compact domain, assuming the kernel functions are continuously differentiable, and the first \(K\) eigenvalues of \(G\) all have multiplicity one, \(\min_{k \leq K} \delta_k > 0\), and \(\langle \varphi_k, \hat{\varphi}_k \rangle \geq 0\), for \(k = 1, \ldots, K\). Then we have

\[
\sum_{k \leq K} \|\hat{\varphi}_k - \varphi_k\|_2 \leq \sqrt{2} \left\| \hat{G} - G \right\|_2 \sum_{k \leq K} \delta_k^{-1}.
\]

If (S5.14) is satisfied for \(k = 1, \ldots, K\), then

\[
\sum_{k \leq K} \|\hat{\varphi}_k - \varphi_k\|_2 \leq \left\| \hat{G} - G \right\|_2 \sum_{k \leq K} (2\delta_k)^{-1} + \left\| \hat{G} - G \right\|_2^2 \sum_{k \leq K} r_p(\|\Delta\|/\delta_k),
\]

and

\[
\sum_{k \leq K} \|\hat{\varphi}_k - \varphi_k\|_\infty \\
\leq \sum_{k \leq K} \hat{\lambda}_k^{-1} \left( |T|^{1/2} \left\| \hat{G} - G \right\|_\infty + \left\| \hat{G} - G \right\|_2 A_K \right) \\
+ |T|^{1/2} \left\| G \right\|_\infty \left\| \hat{G} - G \right\|_2 \sum_{k \leq K} (2\delta_k \hat{\lambda}_k)^{-1} \\
+ |T|^{1/2} \left\| G \right\|_\infty \left\| \hat{G} - G \right\|_2^2 \sum_{k \leq K} r_p(\|\Delta\|/\delta_k),
\]

for \(A_K = 2 \max \left(1/\sqrt{|T|}, \left(\sum_{k \leq K} \|\varphi_k\|_\infty\right)^{1/3}\right)\) as defined in (S3.1).

**Proof.** Trivial by (S5.11) and that \(\|\Delta\| \leq \left\| \hat{G} - G \right\|_2\) (use Cauchy–Schwarz inequality to see). \(\square\)

**Corollary S5.2.** Under (A1), (C2), and (I)- (II) of Proposition S5.2,

\[
\sup_{k \geq 1} |\hat{\lambda}_k - \lambda_k| = O_p(1/\sqrt{n}), \quad \sup_{k \geq 1} |\hat{\lambda}_k - \lambda_k| = O_p(1/\sqrt{n} + b_n); \\
\|\hat{\varphi}_k - \varphi_k\|_2 = \delta_k^{-1} O_p(1/\sqrt{n}), \quad \|\hat{\varphi}_k - \varphi_k\|_2 = \delta_k^{-1} O_p(1/\sqrt{n} + b_n); \\
\]
and
\[ \| \hat{\varphi} - \varphi \|_\infty = \tilde{\lambda}_k^{-1} O_p \left( \max(1, \delta_k^{-1}, \| \varphi_k \|_\infty) / \sqrt{n} \right), \]
\[ \| \hat{\varphi} - \varphi \|_\infty = \tilde{\lambda}_k^{-1} O_p \left( \max(1 / \sqrt{n} + a_n, \delta_k^{-1} \vee \| \varphi_k \|_\infty)(1 / \sqrt{n} + b_n) \right), \]
where \( \| \varphi_k \|_\infty \) can be replaced by \( A_K \) for \( k \leq K \), and the \( O_p \)-rates are uniform over \( k = 1, 2, \ldots \).

**Proof.** Combine (S5.11), Theorem S5.1, and Proposition S4.2 then follow the steps of corollary 1 in the supplementary of Petersen and Müller (2016). \( \square \)

### S5.4 Projection Error

Recall that, in light of the Karhunen–Loève expansion, given \( f \in L^2(T) \), we write
\[ f_K \overset{\triangle}{=} \mu + \sum_{k=1}^{K} \eta_k \varphi_k, \quad \tilde{f}_K \overset{\triangle}{=} \tilde{\mu} + \sum_{k=1}^{K} \tilde{\eta}_k \tilde{\varphi}_k, \quad \hat{f}_K \overset{\triangle}{=} \hat{\mu} + \sum_{k=1}^{K} \hat{\eta}_k \hat{\varphi}_k, \]
for projection to \( S_K, \tilde{S}_K \) and \( \hat{S}_K \) respectively.

**Lemma S5.3.** Under (A1), for \( L^2(T) \)-valued random element \( f = \psi p \),

for all \( K \),
\[
E \| f - f_K \|_2^2 = \sum_{k > K} \lambda_k,
\]
\[
E \| f - f_K \|_\infty \leq 2 \left( L_K \sum_{k > K} \lambda_k \right)^{1/3} + 2 \left( \sum_{k > K} \lambda_k / |T| \right)^{1/2} ;
\]
\[
\| f - f_K \|_2 = O_p \left( \sqrt{\sum_{k > K} \lambda_k} \right), \quad \| f - f_K \|_\infty = O_p \left( \sqrt{\sum_{k \leq K} \| \varphi_k' \|_\infty \sum_{k > K} \lambda_k} \right),
\]
as \( K \to \infty \), where
\[
L_K = 2M \left( 1 + \sqrt{|T|} \sum_{k \leq K} \| \varphi_k' \|_\infty \right),
\]
with \( \| f' \|_\infty \leq M \).

**Proof.** As argued right after the assumption lists, \( f \) is a mean-square continuous process. Hence by the Karhunen–Loève theorem (c.f. Hsing and Eubank (2015), theorem 7.3.5), we have \( E|f(t) - f_K(t)|^2 = G(t, t) - \sum_{k \leq K} \lambda_k \varphi_k(t)^2 \to 0 \) uniformly over \( t \in T \) as \( K \to \infty \), implying
\[
E \| f - f_K \|_2^2 = E \int_T (f(t) - f_K(t))^2 dt
\]

41
\[
E(f(t) - f_K(t))^2 dt \\
= \int T G(t, t) - \sum_{k \leq K} \lambda_k \varphi_k(t)^2 dt \\
= \sum_{k > K} \lambda_k.
\]

Therefore as \( c \to \infty \),

\[
P \left( \frac{\|f - f_K\|_2}{\sqrt{\sum_{j > K} \lambda_j}} > c \right) \\
\leq \frac{1}{c^2} E \frac{\|f - f_K\|_2^2}{\sum_{j > K} \lambda_j} \\
= \frac{1}{c^2} \to 0
\]
gives the desired result for the \( L^2 \) norm.

As for \( \|\cdot\|_\infty \), by (4.13), (4.14) in Petersen and Müller (2016), for any \( K \), \( f - f_K \) is Lipschitz continuous with Lipschitz constant \( L_K \) and hence by Lemma 1 in Petersen and Müller (2016),

\[
E \|f - f_K\|_\infty \leq 2E \max \left( \frac{\|f\|_2}{|T|^{1/2}}, L_K \frac{\|f - f_K\|_2^{2/3}}{\|f - f_K\|_2} \right) \\
\leq 2 |T|^{-1/2} E \|f - f_K\|_2 + 2L_K^{1/3} E \|f - f_K\|_2^{2/3} \\
\leq 2 \sqrt{\sum_{k > K} \lambda_k / |T|} + 2 \left( L_K \sum_{k > K} \lambda_k \right)^{1/3} \\
= O \left( \sqrt{\sum_{k \leq K} \|\varphi_k\|_\infty \sum_{k > K} \lambda_k} \right)
\]
as \( K \to \infty \), where the last inequality from Jensen’s inequality noting \( x \mapsto x^{1/2} \) and \( x \mapsto x^{1/3} \) are concave; and that \( L_K = O(\sum_{k \leq K} \|\varphi_k\|_\infty) \) is bounded away from zero for all \( K \), since \( \langle \varphi_k, 1 \rangle = 0 \) implying eigenfunctions are not constant and hence \( \|\varphi_k\|_\infty > 0 \); and that \( \sum_{k > K} \lambda_k \to 0 \). Further applying Markov’s inequality gives the desired result.

Next the effect on the projections of replacing \( F_K \) by its estimated versions are studied. Recall in Section S3, \( \tilde{\eta}_k = \langle f - \tilde{\mu}, \varphi_k \rangle \) and \( \tilde{f}_K = \tilde{\mu} + \sum_{k=1}^K \tilde{\eta}_k \tilde{\varphi}_k \), similarly we define \( \hat{\eta}_k \) and \( \hat{f}_K \) by replacing \( \tilde{\mu} \) and \( \tilde{\varphi}_k \) by \( \hat{\mu} \) and \( \hat{\varphi}_k \).

**Lemma S5.4.** Under (A1), for all \( f \in F \) and \( K \leq \text{rank} \tilde{G} \wedge \text{rank} \hat{G} \), we have

\[
\sup_{k \leq K} (|\eta_k| \vee |\tilde{\eta}_k|) \leq 2M |T|^{1/2},
\]

\[
\sup_{k \leq K} (|\hat{\eta}_k| \vee |\tilde{\eta}_k|) \leq 2M |T|^{1/2},
\]

42
\[
\|\hat{f}_k - f_k\|_2 \leq (K + 1) \|\hat{\mu} - \mu\|_2 + 4M |\mathcal{T}|^{1/2} \sum_{k \leq K} \|\hat{\varphi}_k - \varphi_k\|_2,
\]

Moreover,

\[
\sup_{k \leq K} |\hat{\eta}_k - \eta_k| \leq 2M |\mathcal{T}|^{1/2} + \|\hat{\mu} - \mu\|_2,
\]

\[
\|\hat{f}_k - f_k\|_\infty \leq \|\hat{\mu} - \mu\|_\infty + KA_K \|\hat{\mu} - \mu\|_2 + 2M |\mathcal{T}|^{1/2} \left( A_K \sum_{k \leq K} \|\hat{\varphi}_k - \varphi_k\|_2 + \sum_{k \leq K} \|\hat{\varphi}_k - \varphi_k\|_\infty \right).
\]

Proof. Under (A1), by Hölder’s inequality, \(|\eta_k| = |f(\varphi_k - \varphi_k)| \leq \|f - \mu\|_2 \leq 2M |\mathcal{T}|^{1/2}\), similarly \(|\hat{\eta}_k| \leq \|f - \hat{\mu}\|_2 \leq 2M |\mathcal{T}|^{1/2}\). Also,

\[
|\hat{\eta}_k - \eta_k| = \left| \int f(\hat{\varphi}_k - \varphi_k) + \mu \varphi_k - \hat{\mu} \hat{\varphi}_k \right|
\]

\[
= \left| \int f(\hat{\varphi}_k - \varphi_k) - \mu (\hat{\varphi}_k - \varphi_k) - \hat{\varphi}_k (\hat{\mu} - \mu) \right|
\]

\[
\leq \|f - \mu\|_2 \|\hat{\varphi}_k - \varphi_k\|_2 + \|\hat{\mu} - \mu\|_2
\]

\[
\leq 2M |\mathcal{T}|^{1/2} \|\hat{\varphi}_k - \varphi_k\|_2 + \|\hat{\mu} - \mu\|_2.
\]

Therefore

\[
\|\hat{f}_k - f_k\|_2 \leq \|\hat{\mu} - \mu\|_2 + \sum_{k \leq K} \|\eta_k (\hat{\varphi}_k - \varphi_k) + \hat{\varphi}_k (\eta_k - \varphi_k)\|_2
\]

\[
\leq \|\hat{\mu} - \mu\|_2 + \sum_{k \leq K} \|\eta_k\| \|\hat{\varphi}_k - \varphi_k\|_2 + \sum_{k \leq K} |\eta_k - \varphi_k|
\]

\[
\leq \|\hat{\mu} - \mu\|_2 + \sum_{k \leq K} \left( |\eta_k| + \|f - \mu\|_2 \right) \|\hat{\varphi}_k - \varphi_k\|_2 + K \|\hat{\mu} - \mu\|_2
\]

\[
\leq (K + 1) \|\hat{\mu} - \mu\|_2 + 4M |\mathcal{T}|^{1/2} \sum_{k \leq K} \|\hat{\varphi}_k - \varphi_k\|_2.
\]
Similarly noting $\|s_k\|_\infty \leq A_K$, $k = 1, \ldots, K$, we have
\[
\begin{aligned}
\|\hat{f}_K - f_K\|_\infty & \leq \|\hat{\mu} - \mu\|_\infty + \sum_{k \leq K} \|\hat{\eta}_k (\hat{s}_k - s_k) + s_k (\hat{\eta}_k - \eta_k)\|_\infty \\
& \leq \|\hat{\mu} - \mu\|_\infty + \sum_{k \leq K} \|s_k\|_\infty |\hat{\eta}_k - \eta_k| + \sum_{k \leq K} \|\hat{\eta}_k\|_\infty \|\hat{s}_k - s_k\|_\infty \\
& \leq \|\hat{\mu} - \mu\|_\infty + \sum_{k \leq K} \|\hat{\eta}_k\|_\infty \|\hat{s}_k - s_k\|_\infty \\
& + \sum_{k \leq K} \|s_k\|_\infty (\|f - \mu\|_2 \|\hat{s}_k - s_k\|_2 + \|\hat{\mu} - \mu\|_2) \\
& \leq \|\hat{\mu} - \mu\|_\infty + KA_K \|\hat{\mu} - \mu\|_2 \\
& + 2M |\mathcal{T}|^{1/2} \left( A_K \sum_{k \leq K} \|\hat{s}_k - s_k\|_2 + \sum_{k \leq K} \|\hat{s}_k - s_k\|_\infty \right).
\end{aligned}
\]

Next, for $\hat{f}_K$, just replace all tilde by hat, and update the stochastic bound of $|\hat{\eta}_k|$ to arrive at a uniform bound for the errors of projecting to the estimated space $\hat{\mathcal{F}}_K$. Note that $\|\hat{\mu} - \mu\|_2$ comes into the upper bound of $|\hat{\eta}_k|$ because we are working with the pre-smoothed trajectories, which are not necessarily uniformly bounded by $M$. The derivation for difference in component scores and in $L_2$ norm are identical, obtaining
\[
\begin{aligned}
|\hat{\eta}_k - \eta_k| & \leq 2M |\mathcal{T}|^{1/2} \|\hat{s}_k - s_k\|_2 + \|\hat{\mu} - \mu\|_2, \\
\|\hat{f}_K - f_K\|_2 & \leq (K + 1) \|\hat{\mu} - \mu\|_2 + 4M |\mathcal{T}|^{1/2} \sum_{k \leq K} \|\hat{s}_k - s_k\|_2.
\end{aligned}
\]

For the $L_\infty$ norm, by Hölder’s inequality, $\sup_{k \leq K} |\hat{\eta}_k| \leq \|f - \mu\|_2 + \|\hat{\mu} - \mu\|_2 \leq 2M |\mathcal{T}|^{1/2} + \|\hat{\mu} - \mu\|_2$, where an additional term is introduced, and
\[
\begin{aligned}
\|\hat{f}_K - f_K\|_\infty & \leq \|\hat{\mu} - \mu\|_\infty + \sum_{k \leq K} \|\hat{\eta}_k (\hat{s}_k - s_k) + s_k (\hat{\eta}_k - \eta_k)\|_\infty \\
& \leq \|\hat{\mu} - \mu\|_\infty + \sum_{k \leq K} \|s_k\|_\infty |\hat{\eta}_k - \eta_k| + \sum_{k \leq K} \|\hat{\eta}_k\|_\infty \|\hat{s}_k - s_k\|_\infty \\
& \leq \|\hat{\mu} - \mu\|_\infty + \sum_{k \leq K} \|\hat{\eta}_k\|_\infty \|\hat{s}_k - s_k\|_\infty \\
& + \sum_{k \leq K} \|s_k\|_\infty (\|f - \mu\|_2 \|\hat{s}_k - s_k\|_2 + \|\hat{\mu} - \mu\|_2) \\
& \leq \|\hat{\mu} - \mu\|_\infty + \left( KA_K + \sum_{k \leq K} \|\hat{s}_k - s_k\|_\infty \right) \|\hat{\mu} - \mu\|_2.
\end{aligned}
\]
Firstly, for all $\eta$, we have a uniform bound for the difference between its projection to the actual low-dimensional family $F_K$ and that to the estimated family $\hat{F}_K$.

Combining this with Lemma S5.3 leads to the order of magnitude for $\|f - \hat{f}_K\|$ and $\|f - \hat{f}_K\|\infty$, namely just take expectation w.r.t. the new independent copy $f$.

Moreover, it is easy to arrive at the consistency result for component scores, as an extension to Proposition 5.2 in the main text. Note that $\tilde{\eta}_k = \langle \hat{f}_i - \hat{\mu}, \hat{\varphi}_k \rangle$ and $\hat{\eta}_k = \langle f_i - \mu, \varphi_k \rangle$ for $i = 1, \ldots, n, k = 1, \ldots, K$.

**Proposition S5.4.** Under (A1), (C2), and (I)-(II) of Proposition S5.2, then for any fixed $K < \infty$, as $n \to \infty$, we have

$$\sup_{i \leq n} \sup_{k \leq K} |\hat{\eta}_{ik} - \tilde{\eta}_{ik}| = O_p \left( \sqrt{\frac{\log n}{mh}} + (n^{-1/2} + b_n) \sup_{k \leq K} \delta_k^{-1} \right),$$

$$\sup_{i \leq n} \sup_{k \leq K} |\hat{\eta}_k - \tilde{\eta}_k| = O_p \left( n^{-1/2} \sup_{k \leq K} \delta_k^{-1} \right).$$

**Proof.** Firstly, for all $i = 1, \ldots, n$ and $k = 1, \ldots, \text{rank} \hat{G}$, $|\tilde{\eta}_{ik}| = |\hat{f}_i - \hat{\mu}| \hat{\varphi}_k| \leq \|\hat{f}_i - \hat{\mu}\|_2 + \|\hat{\mu} - \mu\|_2$, and

$$|\hat{\eta}_{ik} - \tilde{\eta}_{ik}| = \left| \int (f_i \hat{\varphi}_k - \hat{\mu} \hat{\varphi}_k - f_i \varphi_k + \mu \varphi_k) \right|$$

$$= \left| \int (f_i \hat{\varphi}_k - \varphi_k) + \hat{\mu} (f_i - f_i) - \mu (\hat{\varphi}_k - \varphi_k) - \varphi_k (\hat{\mu} - \mu) \right|$$

$$\leq \|f_i - \mu\|_2 \|\hat{\varphi}_k - \varphi_k\|_2 + \|f_i - f_i\|_2 + \|\hat{\mu} - \mu\|_2$$

$$\leq 2M |T|^{1/2} \|\hat{\varphi}_k - \varphi_k\|_2 + |T|^{1/2} \|f_i - f_i\|_\infty + \|\hat{\mu} - \mu\|_2.$$

Therefore by Theorem S5.1, Proposition S5.2, and Proposition S5.3, as $n$ diverges,

$$\sup_{i \leq n} |\hat{\eta}_{ik} - \tilde{\eta}_k| = O_p \left( (1 + \delta_k^{-1})(1/\sqrt{n} + b_n) + \sqrt{\log(n)/(mh)} \right),$$

which goes to zero uniformly for any finitely many $k \leq K$. Further noting $\delta_k^{-1} \geq 1$ as $k \to \infty$, we have the desired result.

For component scores $\tilde{\eta}_k$ without pre-smoothing, it suffices to combine Lemma S5.4, Proposition S5.3, and Theorem S5.1. \hfill \Box
S6 Results related to KL divergence

In this section we discuss the consistency of MLE estimators within low-dimensional approximating families $\mathcal{P}_K$, $\hat{\mathcal{P}}_K$ and $\tilde{\mathcal{P}}_K$ when provided an additional i.i.d. sample of size $N$ sampled from $p$, an independent copy of the random density.

For simplicity of notation, we use $p$ instead of $p_0$ as in our main text for density of the testing sample, and let $X_1, \ldots, X_N$ be the i.i.d. observations within $\mathcal{T}$. To distinguish the approximating families used, we use ring overscript to denote the MLE within $\mathcal{P}_K$, tilde for that with $\tilde{\mathcal{P}}_K$, and hat for $\hat{\mathcal{P}}_K$; more precisely, let

$$\hat{\theta} = \arg \max_{\theta} \hat{i}_K(\theta), \quad \tilde{\theta} = \arg \max_{\theta} \tilde{i}_K(\theta), \quad \theta = \arg \max_{\theta} \theta_K(\theta),$$

where

$$\hat{i}_K(\theta) = \theta^T \varphi_{\text{ave}} - B_K(\theta),$$
$$\tilde{i}_K(\theta) = \theta^T \tilde{\varphi}_{\text{ave}} - \tilde{B}_K(\theta),$$
$$i_K(\theta) = \theta^T \varphi_{\text{ave}} - \hat{B}_K(\theta),$$

in which $\varphi_{\text{ave}}, \tilde{\varphi}_{\text{ave}}$ and $\varphi_{\text{ave}} \in \mathbb{R}^K$ are sufficient statistics w.r.t. $\mathcal{P}_K$, $\tilde{\mathcal{P}}_K$ and $\hat{\mathcal{P}}_K$; for example, the $k$th element of $\varphi_{\text{ave}}$ is $\sum_{j=1}^N \varphi_k(X_j)/N$ for $k = 1, \ldots, K$. Hence, correspondingly $\hat{p} = \hat{p}_{N,K}$ is the estimated density function in $\mathcal{P}_K$ with natural parameter $\hat{\theta} = \hat{\theta}_{N,K}$. Also, we use star for information projection

$$\hat{p}^* = \hat{p}_K^* = \arg \min_{q \in \mathcal{P}_K} D(p \| q),$$

as the element in low-dimensional approximation $\mathcal{P}_K$ that minimizes the KL divergence from the true density $p$ with parameter $\hat{\theta}^*$. Similarly, we write the information projection of $p$ to $\hat{\mathcal{P}}_K$ and $\mathcal{P}_K$ as $\tilde{p}^*$ and $\hat{p}^*$ with corresponding parameters $\tilde{\theta}^*$ and $\hat{\theta}^*$.

We use the following result adapted from theorem 3 of Barron and Sheu (1991) for bounding the KL divergence while fitting densities using an exponential approximating family.

**Theorem S6.1** (Theorem 3, Barron and Sheu (1991)). On $L^2(\mathcal{T})$ with Lebesgue measure and a closed and compact set $\mathcal{T} \subset \mathbb{R}$, let $A_K$ be such that $A_K |\mathcal{T}|^{1/2} \geq 1$ and $\|g\|_\infty \leq A_K \|g\|_2$ for all $g \in \text{span}\{\phi_1, \ldots, \phi_K\}$, where $\phi_k \in L^2(\mathcal{T})$ are orthonormal and $\int_\mathcal{T} \phi_k(t) dt = 0$ for $k = 1, \ldots, K$. Given any density $q$ and measurable finite function $\nu$ on $\mathcal{T}$, satisfying $1/M \leq \|q\|_\infty, \|1/q\|_\infty \leq M$ for some $M > 1$, let $f = \psi q$ be the centered log-density as defined in (2.1) of the main text, and $\Delta_K = \|f - f_K\|_2, \gamma_K = \|f - f_K\|_\infty$, where $f_K(t) = \nu(t) + \sum_{k \leq K} \zeta_k \phi_k(t), t \in \mathcal{T}$ with $\zeta_k = \int_\mathcal{T} (f(x) - \nu(x)) \phi_k(x) dx$. Denote

$$\varepsilon_K = 4M \exp(4\gamma_K + 1)A_K \Delta_K,$$
\[ \delta_{K,N} = 4M^{1/2} \exp(2\gamma_K + 2)A_K \sqrt{K/N}. \]

Let \( \mathcal{P}_K \) be the set of densities on \( \mathcal{T} \), in which the densities are, given natural parameters \( \theta_1, \ldots, \theta_K \), proportional to \( \exp(\nu(t) + \sum_{k \leq K} \theta_k \phi_k(t)), t \in \mathcal{T} \).

If \( \varepsilon_K \leq 1 \), the information projection \( p^*_K \) exists and satisfies
\[ D(q \| q^*_K) = \inf_{q_K \in \mathcal{P}_K} D(q \| q_K) \leq C_1 \Delta_{K}, \]
where \( C_1 = \frac{1}{2} e^{\gamma_K} M \).

Moreover, if \( \delta_{K,N} \leq 1 \), then for all \( \mathcal{K} \leq \delta_{K,N}^{-2} \), there is a set with probability less than \( 1/\mathcal{K} \), outside which the maximum likelihood estimator in \( \mathcal{P}_K \) with i.i.d. sample from \( q \) of size \( N \), denoted as \( \hat{q}_{N,K} \), exists and satisfies
\[ D(q_K \| \hat{q}_{N,K}) \leq C_2 \frac{K}{N} \mathcal{K}, \]
\[ D(q \| \hat{q}_{N,K}) \leq C_1 \Delta_{K} + C_2 \frac{K}{N} \mathcal{K}, \]
where \( C_2 = 2 \exp(2\gamma_K + 1 + \varepsilon_K) \leq 2 \exp(2\gamma_K + 2) \).

**Proof.** It suffices to verify the conditions in Theorem 3 of Barron and Sheu (1991). This version has two major differences to theirs: the baseline measure and the log-densities, as their discussion is based on a baseline measure and targeting the log-densities without centralizing. To apply their result, it suffice to verify the conditions in Theorem 3 of Barron and Sheu (1991) up to multiplying a constant due to different baseline measure. Therefore, the desired result follows easily with that \( q \) being lower and upper bounded, which implies \( 1/\sqrt{M} \| q \|_2 \leq \| q \|_{L_2(q)} \leq \sqrt{M} \| q \|_2 \) for all \( q \in L^2(\mathcal{T}) \), and that KL divergence is invariant to changing baseline measure as long as the baselines are equivalent, because \( D(q_1 \| q_2) = \int_T q_1(t) (\log q_1(t) - \log q_2(t)) dt = \int_T q_1(t)/e^{\nu(t)} (\log (q_1(t)/e^{\nu(t)}) - \log (q_2(t)/e^{\nu(t)})) e^{\nu(t)} dt. \)

The last inequality in the theorem is due to the Pythagorean-like relation
\[ D(q \| \hat{q}_{N,K}) = D(q \| q^*_K) + D(q^*_K \| \hat{q}_{N,K}) \]
as in (2.6) of Barron and Sheu (1991). \( \square \)
S6.1 A Uniform Convergence Result in KL Divergence

**Proposition S6.1.** Under (A1), (C1), and (C2), suppose \( A_K \sqrt{K/N} \to 0 \), then for all \( \varepsilon > 0 \), there exists \( B = B_\varepsilon \) and \( K_\varepsilon \) such that for all \( K > K_\varepsilon \), we have

\[
P \left( D(p \parallel \hat{p}_K^*) > B \sum_{k > K} \lambda_k \right) < \varepsilon,
\]

\[
P \left( D(p \parallel \hat{p}_{N,K}) > B \left( \sum_{k > K} \lambda_k + \frac{K}{N} \right) \right) < \varepsilon.
\]

Moreover, let \( R_K \triangleq \left( \sum_{k \leq K} \| \varphi'_k \|_\infty \sum_{k > K} \lambda_k \right)^{1/3} \), as \( K \to \infty, N \to \infty \), we have

\[
D(p \parallel \hat{p}_K^*) = O_P \left( e^{2R_k} \sum_{k > K} \lambda_k \right),
\]

\[
D(p \parallel \hat{p}_{N,K}) = O_P \left( e^{2R_k} \frac{K}{N} \right),
\]

\[
D(p \parallel \hat{p}_{N,K}) = O_P \left( e^{2R_k} \frac{K}{N} \right).
\]

**Proof.** Following notation in Theorem S6.1, let \( \varepsilon_K \triangleq 4M \exp(4\gamma_K + 1)A_K \Delta_K \) where \( \gamma_K = \| f - f_K \|_\infty \) and \( \Delta_K = \| f - f_K \|_2 \), then for \( B > 0 \),

\[
P \left( D(p \parallel \hat{p}_K^*) > B^2 \sum_{k > K} \lambda_k \right)
\]

\[
\leq P \left( D(p \parallel \hat{p}_K^*) > \frac{1}{2} M \exp(\gamma_K) \Delta_K^2 \right) + P \left( \frac{1}{2} M \exp(\gamma_K) \Delta_K^2 > B^2 \sum_{k > K} \lambda_k \right)
\]

\[
\leq P (\varepsilon_K > 1) + P \left( \frac{1}{2} M \exp(\gamma_K) \Delta_K^2 > B^2 \sum_{k > K} \lambda_k \right)
\]

by Theorem S6.1. Further by Lemma S5.3, for \( R > 0 \),

\[
P(\varepsilon_K > 1)
\]

\[
= P(\gamma_K > R, \varepsilon_K > 1) + P(\gamma_K \leq R, \varepsilon_K > 1)
\]

\[
\leq P(\gamma_K > R) + P \left( A_K \Delta_K > (4M)^{-1} e^{-4R-1} \right)
\]

\[
\leq \frac{1}{R} E\gamma_K + (4MA_K e^{4R+1})^2 E\Delta_K^2
\]

\[
\leq \frac{2}{R} \left( L_K \sum_{k > K} \lambda_k \right)^{1/3} + \left( \sum_{k > K} \lambda_k / |T| \right)^{1/2} + (4A_K e^{4R+1})^2 \sum_{k > K} \lambda_k.
\]

The last quantity can be made arbitrarily small under (C1) for all \( K > K_\varepsilon \) with sufficiently large \( R = R_\varepsilon \) and \( K_\varepsilon \).
Similarly, for $B$ large such that $2B/M > 1$,

$$P\left(\frac{1}{2} M \exp(\gamma_K) \Delta_K^2 > B^2 \sum_{k>K} \lambda_k\right)$$

$$\leq P(\gamma_K > \log(2B/M)) + P\left(\Delta_K^2 > B \sum_{k>K} \lambda_k\right)$$

$$\leq \frac{1}{\log(2B/M)} E_{\gamma_K} + \left(B \sum_{k>K} \lambda_k\right)^{-1} E\Delta_K^2$$

$$\leq \frac{2}{\log(2B/M)} \left((L_K \sum_{k>K} \lambda_k)^{1/3} + (\sum_{k>K} \lambda_k/|T|)^{1/2}\right) + B^{-1},$$

which can be made arbitrarily small under (C1) with a sufficiently large $B$.

Following a similar but more tedious derivation, we can derive the rate for the MLE $\hat{p}_{N,K}$. The only tricky part would be handling the joint distribution of the stochastic process and the actual observations with conditional probability.

In the end, note that $\gamma_K = \|f - f_K\|_{\infty} \lesssim_p R_K$ from Lemma S5.3, we have the desired result. \hfill $\Box$

We actually have a uniform convergence result for a group of random densities. Note that in RHS of the inequalities in the previous proof, the quantities are irrelevant to any specific $p$ as long as it belongs to $\mathcal{P}$ satisfying conditions (A1) and (C1). Thus one can take supremum over all such $p$ for an uniform result.

### S6.2 Main Result for MLE Consistency

**Theorem S6.2.** For $N$ i.i.d. observations from an independent copy $p$ of the $\mathcal{P}$-valued random element, under (A1), (C1), and (C2),

(i) As $K \to \infty$, with probability tending to 1, the information projection $\hat{p}^*$ of $p$ to $\mathcal{P}_K$ exists and

$$D(p \parallel \hat{p}^*_K) = O_p \left(\sum_{k>K} \lambda_k\right).$$

Moreover, if $N \to \infty$ such that $A_K \sqrt{K/N} = o(1)$, then with probability trending to 1, the maximum likelihood estimator $\hat{p}_{N,K}$ based on family $\mathcal{P}_K$ with $N$ i.i.d. observations exists, and

$$D(p \parallel \hat{p}_{N,K}) = O_p \left(\sum_{k>K} \lambda_k + K/N\right).$$
(ii) If \( n, N \) and \( K \) increases to infinity in such a way that
\[
A_K \sqrt{K/N} \to 0, \quad KA_K/\sqrt{n} \to 0, \quad \sum_{k \leq K} \lambda_k^{-1}/\sqrt{n} = O(1),
\]
\[
A_K \sum_{k \leq K} \delta_k^{-1}/\sqrt{n} \to 0, \quad \sum_{k \leq K} (\delta_k \lambda_k)^{-1}/\sqrt{n} = O(1),
\]
then MLE \( \hat{p}_{N,K} \) within \( \hat{P}_K \) deduced from \( n \) completely observed training subpopulations exists with probability tending to 1, and
\[
D(p \parallel \hat{p}_{N,K}) = O_p \left( \| f_K - f_K \|_2^2 + \sum_{k > K} \lambda_k + \frac{K}{N} \right)
\]
\[
= O_p \left( \frac{1}{n} \left( K + \sum_{k \leq K} \delta_k^{-1} \right)^2 + \sum_{k > K} \lambda_k + \frac{K}{N} \right).
\]

(iii) Under (I)–(II) of Proposition S5.2, if \( n, N \) and \( K \) increases to infinity in such a way that
\[
A_K \sqrt{K/N} \to 0, \quad KA_K(1/\sqrt{n} + b_n) \to 0, \quad \sum_{k \leq K} \lambda_k^{-1}(1/\sqrt{n} + a_n) = O(1),
\]
\[
A_K \sum_{k \leq K} \delta_k^{-1}(1/\sqrt{n} + b_n) \to 0, \quad \sum_{k \leq K} (\delta_k \lambda_k)^{-1}(1/\sqrt{n} + b_n) = O(1),
\]
then MLE \( \hat{p}_{N,K} \) within \( \hat{P}_K \) deduced from \( n \) sparsely observed training subpopulations exists with probability tending to 1, and
\[
D(p \parallel \hat{p}_{N,K}) = O_p \left( \| f_K - f_K \|_2^2 + \sum_{k > K} \lambda_k + \frac{K}{N} \right)
\]
\[
= O_p \left( \left( \frac{1}{n} + b_n^2 \right) \left( K + \sum_{k \leq K} \delta_k^{-1} \right)^2 + \sum_{k > K} \lambda_k + \frac{K}{N} \right).
\]

Proof of Theorem S6.2. To prove (i), denote \( \Delta_K \overset{\Delta}{=} \| f - f_K \|_2, \gamma_K \overset{\Delta}{=} \| f - f_K \|_\infty \), then under the condition of this theorem, by Proposition S4.2, \( \| g \|_\infty \leq A_K \| g \|_2 \) for all \( g \in \mathcal{S}_K \) with \( A_K |T|^{1/2} \geq 1 \) by (S3.1), and by Lemma S5.3, \( \gamma_K = O_p(1) \), \( A_K \Delta_K = o_p(1) \), noting \( A_K = O \left( \left( \sum_{k \leq K} \| \varphi_k \|_\infty \right)^{1/3} \right) \) as \( K \to \infty \). Thus the conditions of Theorem S6.1 is verified with a sufficiently large \( K \). Hence we have \( D(p \parallel \hat{p}_K) = O_p(\| f - f_K \|_2^2) = O_p(\sum_{k > K} \lambda_k) \). The result for MLE \( \hat{p}_{N,K} \) follows similarly.

To prove (ii), it suffices to combine Proposition S4.2, Proposition S5.3, Lemma S5.4, Corollary S5.1, and Theorem S6.1. To give details, when using the
estimated family $\tilde{P}_K$ derived with $n$ training samples, denote $\tilde{\gamma}_K = \|f - \tilde{f}_K\|_{\infty}$
and $\tilde{\Delta}_K = \|f - \tilde{f}_K\|_2$. It is desired that $n, N, K$ increase to infinity in a way such that $\tilde{\gamma}_K = O_p(1)$, $\tilde{A}_K \tilde{\Delta}_K = o_p(1)$ and $\tilde{A}_K \sqrt{K/N} = o_p(1)$, where $\tilde{A}_K$ is the Lipschitz constant for functions in $\tilde{S}_K$. In particular, with $\tilde{\gamma}_K \leq \gamma_K + \|\tilde{f}_K - f_K\|_{\infty}$, $\tilde{\Delta}_K \leq \Delta_K + \|\tilde{f}_K - f_K\|_2$, and Proposition S4.2 giving $\tilde{A}_K = A_K + \sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_{\infty}$, if

1. $N$ increases fast such that:

$$\sqrt{\frac{K}{N}} \left( A_K + \sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_{\infty} \right) = o_p(1);$$

(S6.1)

2. $K$ increases slowly in terms of number of training samples $n$ such that:

$$\sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_{\infty} = O_p(1),$$

(S6.2)

$$A_K \|\tilde{f}_K - f_K\|_2 = o_p(1),$$

(S6.3)

$$\|\tilde{f}_K - f_K\|_{\infty} = O_p(1),$$

(S6.4)

then with probability trending to 1, the MLE $\hat{p}_{N,K}$ within $\tilde{P}_K$ exists and

$$D(p \parallel \tilde{p}_{N,K}) = O_p \left( \left( \|f - f_K\|_2 + \|\tilde{f}_K - f_K\|_2 \right)^2 + K/N \right).$$

To replace the last two requirements (S6.3) and (S6.4) with something irrelevant to an individual trajectory $f$, for example, to replace by requirements upon the estimated families being well-approximating, in terms of their mean and covariance functions, note that under (S6.2), by Lemma S5.4, equation (S6.3) is satisfied if

$$A_K K \|\tilde{\mu} - \mu\|_2 = o_p(1),$$

(S6.5)

$$A_K \sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_2 = o_p(1);$$

(S6.6)

equation (S6.4) is satisfied if

$$\|\tilde{\mu} - \mu\|_2 = O_p(1), \quad A_K K \|\tilde{\mu} - \mu\|_2 = O_p(1), \quad A_K \sum_{k \leq K} \|\tilde{\varphi}_k - \varphi_k\|_2 = O_p(1),$$

where the first equation is verified by Proposition S5.3, the second and the last by (S6.5) and (S6.6); equation (S6.1) is satisfied under (S6.2) if $A_K \sqrt{K/N} \to 0$, noting $A_K \geq 1$. In the end, we only need to verify (S6.2), (S6.5), and (S6.6).

Using Corollary S5.1, one shall see (S6.2) and (S6.6) are satisfied if
where the second equation is implied by the last one, noting that \( \delta_k < \lambda_{k-1} \) for all \( k \) large. Note that compared to the \( L^\infty \) result in Corollary S5.1, we are replacing \( \tilde{\lambda}_k \) by \( \lambda_k \). The last equation, with \( A_K \gtrsim 1 \), implies that for sufficiently large \( n \) such that \( \| \tilde{G} - G \|_2 \) is small, we have \( \| \tilde{G} - G \|_2 \leq \inf_{k \leq K} \delta_k/2 \). Further, by (S5.11), we have \( \lambda_{k+1} < \tilde{\lambda}_k < \lambda_{k-1} \), justifying using \( \lambda_k \) in place of \( \tilde{\lambda}_k \). In the end, plugging-in the rate for errors in mean and covariance functions obtained in Proposition S5.3 leads to the required conditions.

The same argument applies to the situation when pre-smoothing is involved by replacing all tilde quantities with the hat ones, where the only potential difference occurs while handling \( \| \hat{f}_K - f_K \|_{\infty} = O_p(1) \), the hat-version of (S6.4). Notice that according to Lemma S5.4, one additional requirement

\[
\| \hat{\mu} - \mu \|_2 \sum_{k \leq K} \| \hat{\varphi}_k - \varphi_k \|_{\infty} = O_p(1)
\]

is needed. This additional requirement is already granted by Proposition S5.2 and \( \sum_{k \leq K} \| \hat{\varphi}_k - \varphi_k \|_{\infty} = O_p(1) \), the hat-version of (S6.2), therefore the same argument applies with no additional modification other than replacing all tildes with hats. In the end, use the error rates for the pre-smoothed mean and covariance to arrive at the stated results.

\( \square \)

If \( \mathcal{P} \) itself is a finite-dimensional exponential family, then we have the following corollary in completion to Corollary 5.2 in the main text. Note that \( \lambda_k \) and \( \varphi_k \) for \( k > K_0 \) are not involved, we can set them to be zero.

**Corollary S6.1.** Under (A1) and (C2), if \( \mathcal{P} \) is a \( K_0 \)-dimensional exponential family for some finite \( K_0 \), then for any fixed \( K \geq K_0 \),

\[
D(p \parallel \hat{p}_{N,K}) = O_p \left( \frac{1}{N} \right),
\]

as \( N \to \infty \), and

\[
D(p \parallel \hat{p}_{N,K}) = O_p \left( \frac{1}{n} + \frac{1}{N} \right),
\]
as \( n, N \to \infty \).

Moreover, if (I)-(II) of Proposition S5.2 are also satisfied, then
\[
D(p \parallel \hat{p}_{N,K}) = O_p \left( \frac{1}{n} + b_n^2 + \frac{1}{N} \right) ,
\]
as \( n, N \to \infty \).

Proof of Corollary S6.1. The result for \( D(p \parallel \hat{p}_{N,K}) \) is trivial from Theorem S6.1. For the sample version, it suffices to note that when \( n \) is large so that there are at least \( K \) sample eigenfunctions with \( K > K_0 \), for \( k \leq K_0 < l \leq K \), \( \langle \varphi_k, \hat{\varphi}_l \rangle = \langle \varphi_k - \hat{\varphi}_k, \hat{\varphi}_l \rangle \leq \| \varphi_k - \hat{\varphi}_k \|_2 + \sum_{k \leq K_0} \eta_k \langle \hat{\varphi}_l, \varphi_k \rangle \leq \| \hat{\mu} - \mu \|_2 + 2M \| T \|^{1/2} \sum_{k \leq K_0} \| \varphi_k - \hat{\varphi}_k \|_2 \), implying \( \hat{\Delta}_K = \| \hat{f}_K - f_K \|_2 \leq \| \hat{f}_{K_0} - f_{K_0} \|_2 + \sum_{K_0 < l \leq K} | \hat{\varphi}_l | \to 0 \) as \( n \to \infty \). Together with \( \hat{\gamma}_K = \| \hat{f}_K - f_K \|_\infty \leq \| \hat{f}_{K_0} - f_{K_0} \|_\infty + \sup_{K_0 < l \leq K} | \hat{\varphi}_l | \sum_{K_0 < l \leq K} \| \hat{\varphi}_l \|_\infty = O_p(1) \) as \( n \to \infty \), we have the result for \( D(p \parallel \hat{p}_{N,K}) \). Similar derivation also applies when pre-smoothing is involved.

\( \square \)

S6.3 Optimal \( K \)

Here we discuss the selection of the number of components \( K \) while fitting the proposed MLE using \( N \) testing observations, with a minimum size \( m = m(n) \) observations from each of the \( n \) training samples. We have the following proposition to complete Proposition 5.1 in the main text.

Proposition S6.2. If \( \| \varphi \|_\infty \approx k \) and \( \lambda_k \approx k^{-r} \) for some \( r \geq 3 \) when \( k \to \infty \), then (C1) is satisfied. Moreover, under (A1) and (C2), as \( n, N \to \infty \),

(i) setting \( K = \hat{K}^* \approx N^{1/r} \), the conditions in Theorem S6.2.(i) are satisfied, achieving optimal rate of convergence \( D(p \parallel \hat{p}_{N,K^*}) \lesssim_p N^{-1+1/r} \);

(ii) setting \( K = \hat{K}^* = \min \left( n^{1/(4r+4)}, N^{1/r} \right) \), the conditions in Theorem S6.2. (ii) are satisfied, achieving optimal rate of convergence \( D(p \parallel \hat{p}_{N,K^*}) \lesssim_p n^{-1/4+1/(2r+2)} + N^{-1+1/r} \).

(iii) Under additional assumptions (K1) and (C3),

\[
\hat{K}^* \approx \min \left( \left( \frac{m h}{\log n} \right)^{1/(r+1)}, \left( h^4 + \frac{1}{n} + \frac{1}{m^2 h^2} \right)^{-1/(4r+4)}, N^{1/r} \right) ,
\]

the conditions of Theorem S6.2. (iii) are satisfied, achieving optimal rate
\[
D(p \parallel \hat{p}_{N,K^*}) \lesssim_p \left( h^4 + \frac{1}{n} + \frac{1}{m^2 h^2} \right)^{-(r-1)/(4r+4)} + \left( \frac{\log n}{mh} \right)^{(r-1)/(r+1)} + N^{-1+1/r} .
\]

In addition, if
\[
N \lesssim \left( h^4 + \frac{1}{n} + \frac{1}{m^2 h^2} + \left( \frac{\log n}{mh} \right)^4 \right)^{-r/(4r+4)}
\]

53
then
\[ D(p \parallel \hat{p}_{N,K^*}) \lesssim_p N^{-1+1/r}. \]

**Proof.** If \( \|\varphi_k\|_\infty \approx k \) while \( k \) large, then \( A_k^2 \approx \sum_{k \leq K} \|\varphi_k\|_\infty \approx K^2 \), hence for (C1) to hold, we need \( \sum_{k>N} \lambda_k = O(K^{-2}) \). With \( \lambda_k \approx k^{-r} \), we have \( \sum_{k>N} \lambda_k \approx K^{1-r} \), and thus \( r \geq 3 \) implies (C1).

For (i), to apply Theorem S6.2. (i), we will also need \( K^{7/6}/\sqrt{N} \to 0 \), as \( K, N, n \to \infty \), under which we have \( D(p \parallel \hat{p}_{N,K}) = O_p(K^{1-r} + K/N) \) for MLE within \( \mathcal{P}_K \). The \( K \) minimizing the rate is \( \hat{K} \approx N^{1/r} \approx N^{1/3} \), with which the optimal rate would be \( D(p \parallel \hat{p}_{N,K^*}) \lesssim_p N^{-1+1/r} \leq N^{-2/3} \) as \( N \to \infty \).

For (ii), if using estimated approximating family \( \hat{\mathcal{P}}_K \), then to accommodate errors from the estimated mean and covariance, note that \( 2\delta_k \approx rk^{-1-r} \), implying that as \( K \to \infty \), we have
\[
\sum_{k \leq K} \lambda_k \approx K^{r+1}, \quad \sum_{k \leq K} \delta_k \approx K^{r+2}, \quad \sum_{k \leq K} (\delta_k \lambda_k)^{-1} \approx K^{2r+2}.
\]

Thus, to apply Theorem S6.2. (ii), in addition to the previously required conditions (i.e., \( K^{7/6}/\sqrt{N} \to 0 \)), we also need \( K^{2r+2}/\sqrt{n} = O(1) \), as \( K, N, n \to \infty \), in which case
\[
D(p \parallel \hat{p}_{N,K}) = O_p\left( \frac{1}{n} K^{2r+4} + K^{1-r} + \frac{K}{N} \right) = o_p(1).
\]

Seeing the rate as function of \( K \), we shall observe the asymptotic rate is equal to the slower one of \( R_1(K) = cK^{2r+4} + K^{1-r} \) for some \( c \to 0^+ \) and \( R_2(K) = K/N + K^{1-r} \). One can see \( R_1 \) as the trade-off between using more \( K \) for better approximating family \( \hat{\mathcal{P}}_K \) (as in the \( K^{1-r} \) term) and the errors for estimating more components (as in the \( cK^{2r+4} \) term.) Similarly, \( R_2 \) is the trade-off between better \( \hat{\mathcal{P}}_K \) and difficulty in fitting \( K \) parameters with \( N \) observations (as in the \( K/N \) term). Observe that
\[
\arg\min_K R_1(K) \approx c^{-1/(3r+3)}, \quad \arg\min_K R_2(K) \approx N^{1/r},
\]
and that \( R_2 > R_1 \) when \( K \) small and \( R_1 > R_2 \) when \( K \) large. By comparing the unique positive solution \( K_0 \approx (cN)^{-1/(2r+3)} \) of \( R_1(K) = R_2(K) \) to their minima respectively, we shall observe that
\[
\arg\min_{K>0} (R_1(K) \lor R_2(K)) \approx \arg\min_K R_2(K) \text{ if } N \lesssim \exp\left(-\frac{r}{3r+3} \log c\right),
\]
\[
\arg\min_{K>0} (R_1(K) \lor R_2(K)) \approx \arg\min_K R_1(K) \text{ if } N \gtrsim \exp\left(-\frac{r}{3r+3} \log c\right).
\]

Note that here we take \( c = 1/n \), the optimal rate would be constrained by \( R_2 \) if the new sample is insufficient; conversely by \( R_1 \) if training would be the bottleneck. Also accounting for the requirements upon \( K \), we have an optimal
\[ \hat{K}^* \approx \min \left( n^{1/(4r+4)}, N^{1/r} \right) \] as \( n, N \to \infty \), with which the rate \( D(p \parallel \hat{P}_{N, \hat{K}^*}) \lesssim_p n^{-1/4+1/(2r+2)} + N^{-1+1/r} \lesssim_p n^{-1/8} + N^{-2/3} \). In particular, if \( N^{1/r} \lesssim n^{1/(4r+4)} \), then \( D(p \parallel \hat{P}_{N, \hat{K}^*}) \lesssim_p N^{-1+1/r} \) as if we are working with \( \mathcal{P}_K \).

For (iii), for MLE to be consistent, in addition to the previous requirements from (i), (ii), i.e. \( K^{7/6}/\sqrt{N} \to 0 \) and \( K^{2r+2}/\sqrt{n} = O(1) \), we also need

\[ a_n K^{r+1} = O(1), \quad b_n K^{2r+2} = O(1). \]

Therefore, the feasible domain for \( K \) to apply Theorem S6.2. (iii) is \( K = o \left( N^{3/r} \right) \), and that

\[ K^{r+1} \frac{\log n}{mh} = O(1), \quad K \left( h^4 + \frac{1}{n} + \frac{1}{m^2 h^2} \right)^{1/(4r+4)} = O(1). \]

Similar to what we did in (ii), take \( c = h^4 + 1/n + 1/(mh)^2 \approx 1/n + b_n^2 \), we have \( D(p \parallel \hat{P}_{N, K}) \approx_p R_1(K) + R_2(K) \), and the unconstrained minimum is at \( c^{-1/(3r+3)} \) or \( N^{1/r} \) depending on whether \( N \) is large or not. Though, accounting for the restriction to apply Theorem S6.2. (iii), we set

\[ \hat{K}^* \approx \min \left( \left( \frac{mh}{\log n} \right)^{(r+1)/(r+1)}, \left( h^4 + \frac{1}{n} + \frac{1}{m^2 h^2} \right)^{-1/(4r+4)} \right), \quad N^{1/r} \]

After some tedious comparison according to which term is most restricting in the definition of \( \hat{K}^* \), we have,

\[ D(p \parallel \hat{P}_{N, \hat{K}^*}) \lesssim_p \begin{cases} \left( \frac{\log n}{mn} \right)^{(r-1)/(r+1)}, & \text{if } \hat{K}^* \approx \left( \frac{mh}{\log n} \right)^{1/(r+1)}, \\ c^{(r-1)/(4r+4)}, & \text{if } \hat{K}^* \approx c^{-1/(4r+4)}, \\ N^{-1+1/r}, & \text{otherwise } \hat{K}^* \approx N^{1/r}. \end{cases} \]

In particular, when \( N^{1/r} \lesssim \min \left( (mh)^{1/(r+1)}, (\log n)^{1/(r+1)}, c^{-1/(4r+4)} \right) \), which means that the fitting sample is so small that the errors from estimating \( \hat{P}_K \) is dominated by the errors of fitting \( K \) parameters with \( N \) observations, and the resulting rate for KL divergence is \( N^{-1+1/r} \), as if we are working with the family \( \mathcal{P}_K \).

\[ \square \]

### S6.4 Effect of Training

We provide the following result to further characterize the asymptotic behavior of the proposed MLE for a given finite new fitting sample with finitely many \( K \) while there would be ample training subpopulations. Essentially we show that fixing \( N \) and \( K \), as \( m, n \to \infty \), the log-likelihood under \( \hat{P}_K \) would converge to that under \( P_K \), so would the corresponding estimates. We only show the derivation for \( \theta \), the results for \( \hat{\theta} \) can be obtained similarly.

---

55
Proposition S6.3. Under (A1) and (I)-(II) of Proposition S5.2, and a fixed $K$, for a given sample $X_1, \ldots, X_N$ drawn from density $p \in \mathcal{P}$, assume the MLE $\hat{\theta}_{N,K} \in \mathcal{P}_K$ exists and unique with corresponding parameters $\theta \in \mathbb{R}^K$. Then the MLE $\hat{\theta}_{N,K} \in \hat{\mathcal{P}}_K$ exists and unique with probability tending to 1 as $n \to \infty$, where the $n$ is the number of training subpopulations used to construct $\hat{\mathcal{P}}_K$. The corresponding parameters $\hat{\theta}_n \in \mathbb{R}^K$ converges to $\theta$ in probability with
\[
\|\hat{\theta}_n - \theta\|_2 = O_p(1/\sqrt{n} + a_n + b_n).
\]

To proof Proposition S6.3, we first show the following lemmas.

Lemma S6.1. For a sequence of continuous random function $f_n$ on compact domain $\mathcal{T}$, suppose $\|f_n - f\|_{\infty} = \sup_{t \in \mathcal{T}} \|f_n(t) - f(t)\| = O_p(a_n)$ as $n \to \infty$ for some continuous deterministic function $f : \mathcal{T} \to \mathbb{R}^d$ and sequence $a_n \to 0$. Then for any continuously differentiable function $g : \mathcal{G} \to \mathbb{R}^d$, we have
\[
\|g \circ f_n - g \circ f\|_{\infty} = O_p(a_n),
\]
where $\mathcal{G} \subset \mathbb{R}^d$ is a compact set properly containing range of $f$, i.e., $\{f(t) : t \in \mathcal{T}\} \subset \mathcal{G}^\circ$, where $\mathcal{G}^\circ$ is the interior of $\mathcal{G}$.

Proof. Note that $\|f_n - f\|_{\infty} = o_p(1)$, and that $\{f(t) : t \in \mathcal{T}\} \subset \mathcal{G}^\circ$, we have $P(f_n(t) \notin \mathcal{G}$ for any $t \in \mathcal{T}) \to 0$. Hence with sufficiently large $n$ and probability tending to 1, by mean value theorem, there exists some function $f^* : \mathcal{T} \to \mathcal{G}$ such that for all $t \in \mathcal{T}$,
\[
|g \circ f_n(t) - g \circ f(t)| = |g' \circ f^*(t)(f_n(t) - f(t))| \\
\leq \sup_{x \in \mathcal{G}} \|g'\|_{\infty} \|f_n - f\|_{\infty}.
\]
Noting $\sup_{x \in \mathcal{G}} \|g'\|_{\infty}$ is finite, we have the desired result. \hfill $\Box$

Lemma S6.2. Under (A1), (I)-(II) of Proposition S5.2, and a fixed $K$, for any $\theta = (\theta_1, \ldots, \theta_K)$, we have
\[
\hat{B}_K(\theta) - B_K(\theta) = O_p \left( \frac{1}{\sqrt{n}} + a_n + b_n \right),
\]
\[
\|\hat{B}_K'(\theta) - B_K'(\theta)\|_{\infty} = O_p \left( \frac{1}{\sqrt{n}} + a_n + b_n \right),
\]
\[
\|\hat{B}_K''(\theta) - B_K''(\theta)\|_{\infty} = O_p \left( \frac{1}{\sqrt{n}} + a_n + b_n \right),
\]
\[
\|\hat{p}_K(\cdot|\theta) - p_K(\cdot|\theta)\|_{\infty} = O_p \left( \frac{1}{\sqrt{n}} + a_n + b_n \right),
\]
as $n \to \infty$, where $\hat{p}_K(\cdot|\theta) \in \hat{\mathcal{P}}_K$ is the density in $\hat{\mathcal{P}}_K$ with parameter $\theta$, and correspondingly $p_K(\cdot|\theta) \in \mathcal{P}_K$. 

56
Moreover, for any compact set $\mathcal{G}$, $\sup_{\theta \in \mathcal{G}} |\hat{B}_K(\theta)| = O_p(1)$, $\sup_{\theta \in \mathcal{G}} \left\| \hat{B}_K(\theta) \right\| = O_p(1)$, and $\sup_{\theta \in \mathcal{G}} \left\| \hat{B}_K''(\theta) \right\| = O_p(1)$, so for sufficiently large $n$, $\{\hat{B}_K|_{\mathcal{G}} : \mathcal{G} \to \mathbb{R}\}$ and $\{\hat{B}_K'|_{\mathcal{G}} : \mathcal{G} \to \mathbb{R}\}$ each forms a stochastic equicontinuous class of functions.

Proof. To show the stated four convergence rates, denote $\hat{h}_K(t|\theta) = \hat{\mu}(t) + \sum_{k \leq K} \theta_k \hat{\phi}_k(t)$ and $h_K(t|\theta) = \mu(t) + \sum_{k \leq K} \theta_k \phi_k(t)$, then by Corollary S5.2,

$$\left\| \hat{h}_K(\cdot|\theta) - h_K(\cdot|\theta) \right\|_{\infty} = O_p \left( \frac{1}{\sqrt{n}} + a_n + b_n \right),$$

hence using Lemma S6.1 lead to the first rate. Moreover, noting

$$p_K(t|\theta) = \exp \left( h_K(t|\theta) - B_K(\theta) \right),$$

and similarly for $\hat{p}_K(\cdot|\theta)$, claiming Lemma S6.1 again gives the last rate.

Now, consider the $k$th element of $\hat{B}_K'(\theta) - B_K'(\theta)$, as

$$\left| \frac{\partial}{\partial \theta_k} \hat{B}_K(\theta) - \frac{\partial}{\partial \theta_k} B_K(\theta) \right|$$

$$= \left| \int T^{k-1} \phi_k(t)(\hat{p}_K(t|\theta) - p_K(t|\theta)) + p_K(t|\theta)(\hat{\phi}_k(t) - \phi_k(t)) dt \right|$$

$$\leq \left\| \hat{p}_K(\cdot|\theta) - p_K(\cdot|\theta) \right\|_2 + \left\| p_K(\cdot|\theta) \right\|_2 \left\| \hat{\phi}_k - \phi_k \right\|_2$$

$$= O_p \left( \frac{1}{\sqrt{n}} + a_n + b_n \right).$$

This applies to all $k = 1, \ldots, K$, leading to the second rate.

Finally, since we are working in finite dimensional exponential families, thus let $\text{cov}_{\hat{p}_K(\cdot|\theta)}(\hat{\phi}_k, \hat{\phi}_l)$ denote the covariance of $\hat{\phi}_k(X)$ and $\hat{\phi}_l(X)$ under $X \sim \hat{p}_K(\cdot|\theta)$, similarly we write $\text{cov}_{p_K(\cdot|\theta)}(\phi_k, \phi_l)$ and $\text{cov}_{\hat{p}_K(\cdot|\theta)}(\phi_k, \phi_l)$. Therefore

$$\frac{\partial^2}{\partial \theta_k \partial \theta_l} \hat{B}_K(\theta) - \frac{\partial^2}{\partial \theta_k \partial \theta_l} B_K(\theta) = \text{cov}_{\hat{p}_K(\cdot|\theta)}(\hat{\phi}_k, \hat{\phi}_l) - \text{cov}_{\hat{p}_K(\cdot|\theta)}(\phi_k, \phi_l) + \text{cov}_{p_K(\cdot|\theta)}(\phi_k, \phi_l).$$

Note that

$$\left| \text{cov}_{\hat{p}_K(\cdot|\theta)}(\hat{\phi}_k, \hat{\phi}_l) - \text{cov}_{\hat{p}_K(\cdot|\theta)}(\phi_k, \phi_l) \right|$$

$$= \left| E \hat{\phi}_k \hat{\phi}_l - E \hat{\phi}_k E \hat{\phi}_l - E \phi_k \phi_l + E \phi_k E \phi_l \right|$$

$$= \left| E (\hat{\phi}_k \hat{\phi}_l - \phi_k \phi_l) + \phi_k (\hat{\phi}_l - \phi_l) - E \hat{\phi}_k E (\hat{\phi}_l - \phi_l) - E \phi_k E (\hat{\phi}_k - \phi_k) \right|$$

$$\leq 2 \left\| \hat{p}_K(\cdot|\theta) \right\|_2 \left( \left\| \hat{\phi}_k - \phi_k \right\|_\infty + \left\| \hat{\phi}_l - \phi_l \right\|_\infty \right),$$

where the expectations are w.r.t. density $\hat{p}_K(\cdot|\theta)$. Also

$$\left| \text{cov}_{\hat{p}_K(\cdot|\theta)}(\phi_k, \phi_l) - \text{cov}_{p_K(\cdot|\theta)}(\phi_k, \phi_l) \right|$$

$$= 57$$
\[
\begin{align*}
&\leq |\int \varphi_k \varphi_l (\hat{p}_K - p_K)| + |\int \varphi_k \hat{p}_K \int \varphi_l - \int \varphi_k p_K | \\
&= |\int \varphi_k \varphi_l (\hat{p}_K - p_K)| + |\int \varphi_k \hat{p}_K \int \varphi_l (\hat{p}_K - p_K) + \int \varphi_l p_K | \int \varphi_k (\hat{p}_K - p_K)| \\
&\leq \|\hat{p}_K - p_K\|_\infty + (\|\hat{p}_K\|_2 + \|p_K\|_2) \|\hat{p}_K - p_K\|_2,
\end{align*}
\]

where the integrals are on \( T \) and \( \hat{p}_K = \hat{p}_K(\cdot|\theta) \), \( p_K = p_K(\cdot|\theta) \) for simplification of notation. Combining these with Corollary S5.2 and the rate for \( \|\hat{p}_K(\cdot|\theta) - p_K(\cdot|\theta)\|_\infty \), we have the stated results.

Moreover, note that

\[
\sup_{\theta \in \mathcal{G}} |\bar{B}_K(\theta)| = \sup_{\theta \in \mathcal{G}} \log \int_T \exp(\bar{h}_K(t|\theta))dt| \\
\leq |T| \left( \|\bar{\mu}\|_\infty + \sup_{\theta \in \mathcal{G}} \|\theta\|_\infty \sum_{k=1}^K \|\hat{\varphi}_k\|_\infty \right) \\
= O_p(1),
\]

and that

\[
\sup_{\theta \in \mathcal{G}} \|\hat{p}_K(\cdot|\theta)\|_\infty = \sup_{\theta \in \mathcal{G}} \exp \left( \bar{\mu}(t) + \sum_{k=1}^K \theta_k \hat{\varphi}_k(t) - \bar{B}_K(\theta) \right) \\
= \exp \left( \|\bar{\mu}\|_\infty + \sup_{\theta \in \mathcal{G}} \|\theta\|_\infty \sum_{k=1}^K \|\hat{\varphi}_k\|_\infty + \sup_{\theta \in \mathcal{G}} \|\bar{B}_K(\theta)\|_\infty \right) \\
= O_p(1),
\]

we have that

\[
\sup_{\theta \in \mathcal{G}} \left\| \bar{B}_K'(\theta) \right\| \leq \sup_{\theta \in \mathcal{G}} \|\hat{p}_K(\cdot|\theta)\|_2 \leq \sup_{\theta \in \mathcal{G}} \|\hat{p}_K(\cdot|\theta)\|_\infty = O_p(1),
\]

where the first inequality is due to Hölder’s inequality and that \( \varphi_k \) are orthonormal. Also,

\[
\sup_{\theta \in \mathcal{G}} \left\| \bar{B}_K''(\cdot|\theta) \right\|_\infty = \sup_{\theta \in \mathcal{G}} \max_{1 \leq k, l \leq K'} |E \hat{\varphi}_{k\cdot} \hat{\varphi}_{l\cdot} - E \hat{\varphi}_k \hat{\varphi}_l| \\
\leq \sup_{\theta \in \mathcal{G}} \|\hat{p}_K(\cdot|\theta)\|_\infty + \sup_{\theta \in \mathcal{G}} \|\hat{p}_K(\cdot|\theta)\|_\infty^2 \\
= O_p(1).
\]

In the end, note that \( \sup_{\theta \in \mathcal{G}} \left\| \bar{B}_K'(\theta) \right\|_\infty = O_p(1) \) if and only if for any \( \varepsilon > 0 \), there exists \( n_\varepsilon \in \mathbb{N} \) and \( \mathcal{K} > 0 \) such that for all \( n > n_\varepsilon \),

\[
P \left( \sup_{\theta \in \mathcal{G}} \left\| \bar{B}_K'(\theta) \right\|_\infty > \mathcal{K} \right) < \varepsilon,
\]

58
which implies that
\[ P \left( | \hat{B}_K(\theta_1) - \hat{B}_K(\theta_2) | \leq \mathcal{K} \| \theta_1 - \theta_2 \| \right) \text{ for any } \theta_1, \theta_2 \in \mathcal{G} < \varepsilon. \]

In other words, \( \{ \hat{B}_K|_\mathcal{G} : \mathcal{G} \to \mathbb{R} \} \) forms a class of stochastic equicontinuous functions. Similar steps can be used to show the result for \( \{ \hat{B}_K'|_\mathcal{G} : \mathcal{G} \to \mathbb{R} \}. \)

**Proof of Proposition S6.3.** Let \( \tilde{l}(\theta) = \hat{l}_n(\theta) - \bar{l}(\theta) \) be the objective function for MLE as defined in the beginning of Section S6, where the index \( n \) is inherited from the estimated \( \hat{\mu} \) and \( \hat{\varphi}_1, \ldots, \hat{\varphi}_K \). The over all steps are, first we show that \( \hat{l}_n \to \bar{l} \) uniformly over some compact neighborhood of \( \hat{\theta} \) as number of training samples \( n \to \infty \), so that the local maximum of \( \hat{l}_n \) converge to that of \( \bar{l} \), i.e., \( \hat{\theta} \). Then using the convexity of \( \bar{l} \) to show the local maxima are global.

Denote \( W_n(\theta) = \hat{l}_n(\theta) - \bar{l}(\theta) \). By Corollary S5.2 and Lemma S6.2, \( |W_n(\theta)| = o_p(1) \) as \( n \to \infty \) for any \( \theta \). Further note that \( \{W_n\} \) is stochastic equicontinuous as functions of \( \theta \), thus the convergence is uniform over any given compact set of \( \theta \). In other words, for any \( \varepsilon > 0 \), we have \( \sup_{\theta \in \mathcal{I}_\varepsilon} | \hat{l}_n(\theta) - \bar{l}(\theta) | \xrightarrow{\text{P}} 0 \), where \( \mathcal{I}_\varepsilon = \left\{ \theta : \| \theta - \hat{\theta} \|_2 \leq \varepsilon \right\} \) is the closed ball of radius \( \varepsilon \) centered at \( \hat{\theta} \). Also denote the boundary as \( \partial \mathcal{I}_\varepsilon = \left\{ \theta : \| \theta - \hat{\theta} \|_2 = \varepsilon \right\} \). Since \( \hat{\theta} \) is the maximizer of \( \bar{l} \), there exists positive \( \delta < \varepsilon \) such that
\[
\sup_{\theta \in \partial \mathcal{I}_\varepsilon} \bar{l}(\theta) < \sup_{\theta \in \mathcal{I}_\delta} \bar{l}(\theta) < \sup_{\theta \in \mathcal{I}_\delta} \hat{l}(\theta) + o_p(1),
\]
where the last inequality is due to \( \hat{l}_n \xrightarrow{\text{P}} \bar{l} \) uniformly over \( \mathcal{I}_\delta \). Also noting \( \sup_{\theta \in \partial \mathcal{I}_\varepsilon} | \hat{l}_n(\theta) - \bar{l}(\theta) | \xrightarrow{\text{P}} 0 \), indicating that
\[
\sup_{\theta \in \partial \mathcal{I}_\varepsilon} \hat{l}_n(\theta) \lesssim_p \sup_{\theta \in \mathcal{I}_\delta} \hat{l}_n(\theta).
\]
Moreover, the MLE \( \hat{\theta} \) exists and unique, hence in a neighborhood of \( \hat{\theta} \), the score \( \hat{\varphi}_{\text{ave}} - \hat{B}'_K \) is arbitrarily small and the Hessian \( \hat{B}'_K \) is strictly positive definite, so are the \( \hat{\varphi}_{\text{ave}} - \hat{B}'_K \) and \( \hat{B}'_K \) with probability tending to 1 as \( n \to \infty \), thanks to Lemma S6.2. WLOG, for sufficiently large \( n \) and probability tending to 1, we say \( \hat{l}_n \) is strictly concave over \( \mathcal{I}_\varepsilon \) and that the local maximum \( \arg \max_{\theta \in \mathcal{I}_\varepsilon} \hat{l}_n(\theta) \) exists and unique. Further since the log-likelihood functions \( \hat{l}_n \) are all concave, the local maxima are the global maxima, i.e., \( \hat{\theta}_n = \arg \max_{\theta \in \mathcal{I}_\varepsilon} \hat{l}_n(\theta) \) with probability tending to 1 as \( n \to \infty \). In combine, \( \hat{\theta}_n \) must be within \( \mathcal{I}_\varepsilon \) with probability tending to 1, i.e.
\[
P \left( \| \hat{\theta}_n - \hat{\theta} \|_2 > \varepsilon \right) \to 0,
\]
or equivalently, \( \hat{\theta}_n \xrightarrow{\text{P}} \hat{\theta} \) as \( n \to \infty \).
Next, to study the rate of convergence, for any sequence \( \hat{\theta}_n = \hat{\theta} + o_p(1) \) as \( n \to \infty \), by Taylor expansion, we have

\[
\begin{align*}
\hat{B}_K(\hat{\theta}) - \hat{B}_K(\hat{\theta}) \\
= \left( \hat{\theta} - \hat{\theta} \right)^T \hat{B}'_K(\hat{\theta}) + \left( \hat{\theta} - \hat{\theta} \right)^T \hat{B}''_K(\theta^0) (\hat{\theta} - \hat{\theta}) \\
\leq_p \left\langle \hat{\theta} - \hat{\theta} \right\rangle^T \hat{B}'_K(\hat{\theta}) + \left\| \hat{\theta} - \hat{\theta} \right\|^2
\end{align*}
\]

where \( \theta^0 = \theta^0_n \) is a sequence between \( \hat{\theta} = \hat{\theta}_n \) and \( \hat{\theta} \), and that the last inequality is due to \( \sup_{\theta \in \mathcal{I}_e} \left\| \hat{B}''_K(\theta) \right\| = O_p(1) \). Combine this with the fact that \( B_K \) is twice continuously differentiable so \( \sup_{\theta \in \mathcal{I}_e} \left\| B''_K(\theta) \right\| < \infty \), we have

\[
W_n(\hat{\theta}) = W_n(\hat{\theta}) + o_p(1)
\]

indicating that

\[
\left\| \hat{\theta}_n - \hat{\theta} \right\|_2 = O_p\left( \frac{1}{\sqrt{n}} + a_n + b_n \right).
\]

With the proceeding proposition, it is not difficult to see the next corollary. Again, the index \( n \) is inherited from the \( \hat{\mathcal{P}}_K \).

**Corollary S6.2.** Under the conditions of Proposition S6.3, we have

\[
D(\hat{\phi}_{N,K} \| \hat{\phi}_{N,K}) = O_p\left( \frac{1}{\sqrt{n}} + a_n + b_n \right),
\]

as \( n \to \infty \).
References


