Statistica Sinica Preprint No: SS-2018-0484

Title	Nonparametric density estimation for intentionally corrupted functional data
Manuscript ID	SS-2018-0484
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0484
Complete List of Authors	Aurore Delaigle and Alexander Meister
Corresponding Author	Alexander Meister
E-mail	alexander.meister@uni-rostock.de

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

Nonparametric density estimation for intentionally corrupted functional data

Aurore Delaigle and Alexander Meister

University of Melbourne, Australia and Universität Rostock, Germany

Abstract: We consider statistical models in which the functional data are artificially contaminated by independent Wiener processes in order to satisfy privacy constraints. We show that the corrupted observations have a Wiener density that uniquely determines the distribution of the original functional random variables masked near the origin, and construct a nonparametric estimator of that density. We derive an upper bound for its mean integrated squared error, which has a polynomial convergence rate, and establish an asymptotic lower bound on the minimax convergence rates that is close to the rate attained by our estimator. Our estimator requires choosing a basis and two smoothing parameters. We propose data-driven ways to do so and prove that the asymptotic quality of our estimator is not significantly affected by the empirical parameter selection. Lastly, we examine the numerical performance of our method using simulated examples.

Key words and phrases: classification, convergence rates, differential privacy, infinite-dimensional Gaussian mixtures, Wiener densities.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

1. Introduction

Data privacy is an important feature of a database, where the collected data are transformed and released so as to make it difficult to identify individuals participating in a study. Various privatization methods are available, resulting in different privacy constraints, such as differential privacy. Refer to Wasserman and Zhou (2010) for a statistical introduction to differential privacy. The privatization mechanism typically has an effect on the statistical analysis of the data; thus, one of the research directions in statistical privacy is to find ways of ensuring differential privacy, while keeping as much of the information as possible from the original database (see, e.g., Hall et al. (2013) in the functional data context, and Karwa and Slavkovi (2016) in the setting of synthetic graphs).

One simple way of ensuring differential privacy is to contaminate the data artificially with additive random noise; see, for example, Wasserman and Zhou (2010). In the functional data context, Hall et al. (2013) propose a data release mechanism where the observed functional data are contaminated by adding a random Gaussian process to each function (one per functional observation) that is independent of the original data. Proposition 3.3 in Hall et al. (2013) roughly states that the data can be made differentially private whenever the scaling noise factor of the Gaussian pro-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

cess is sufficiently large.

Here, we show that if the Gaussian process is a Wiener process and the values of the raw data are masked at the origin, then the contaminated data are differentially private and have a density. This contrasts with the usual functional data setting, where the assumption that all measures admitted to be the true image measure of the functional random variables are dominated by a known basic measure seems difficult to justify. There exists no canonical basic measure, such as the Lebesgue measure for finite-dimensional Euclidean data, or the Haar measure for data in general locally compact groups. As a result, inferences and descriptive summaries of functional data are often based on pseudo-densities; see, for example, Delaigle and Hall (2010) and Ciollaro et al. (2016). Recently, Lin et al. (2018) considered the estimation of densities for functions that lie in a dense subset S of the Hilbert space $L_2(D)$, where D is a finite interval. There, S is defined as the (non-closed) linear hull of an orthonormal basis of $L_2(D)$, and does not contain the functional data contaminated by Wiener processes that we consider. Privacy issues for functional data are also discussed by Mirshani et al. (2017). Although the authors deduce the existence of a Gaussian density for fixed functional observations, they do not examine a nonparametric estimation of that density.

In contrast, in the the proposed privatization process, the privatized functional data have a Radon–Nikodym derivative (and thus a true, nonpseudo density) with respect to the Wiener measure. Exploiting the fact that the contaminating distribution is usually known in this context, we consider statistical inferences from such data.

To the best of our knowledge, most existing nonparametric approaches for estimating a Wiener density are motivated by diffusion processes. Although these do not include the type of functional data we consider, some of these methods can be applied in our context. See, for example, Dabo-Niang (2004a), who suggests an orthogonal series estimator, Dabo-Niang (2002, 2004b) and Ferraty and Vieu (2006), who propose kernel density estimators (see also Prakasa Rao (2010a) for a generalization in the case of diffusion processes), and Prakasa Rao (2010b) and Chesneau et al. (2013), who construct wavelet estimators. See also Baíllo et al. (2011) for a parametric context in which the data and the reference measure are Gaussian. However, these methods either suffer from slow logarithmic convergence rates, or are derived under abstract assumptions that seem difficult to justify in our context or are difficult to implement in practice. We propose a fully data-driven estimator with fast polynomial convergence rates under simple conditions. Although our estimator is motivated by our privacy setting, our results can be extended to more general cases of functional data that have a Wiener density.

This remainder of the paper proceeds as follows. In Section 2, we introduce our statistical model, and show that the Wiener density exists and determines uniquely the image measures of the raw functional random variables masked near zero. Moreover, we prove that the privacy constraints are fulfilled when the noise level is sufficiently large. In Section 3, we construct a nonparametric orthonormal series estimator of the Wiener density, and propose data-driven procedures for choosing the basis (Section 3.4) and the smoothing parameters (Section 3.5). In Section 4, we derive an explicit upper bound for the mean integrated squared error of our estimator, and show that it achieves polynomial convergence rates under intuitive tail restrictions and metric entropy constraints on the measure of the original data. Functional data problems in which such fast rates are available are rare; usually, the achievable rates are only logarithmic or sub-polynomial; see, for example, Dabo-Niang (2004a), Mas (2012), and Meister (2016). Finally, we derive a lower bound on the mean integrated squared error under our intuitive conditions, and show that choosing the parameters in a data-driven way does not significantly deteriorate the asymptotic performance of our procedure (thus, we establish a weak adaptivity result). Numerical simulations are provided in Section 5. All proofs are deferred to the online Supplementary Material.

2. Model, data, and applications

2.1 Model and data

We observe functional data $Y_1, \ldots, Y_n$ defined on $[0, 1]$, without loss of generality, which, for reasons such as the differential privacy constraints discussed in Section 1, have been intentionally contaminated by additive random noise. Specifically, we assume that

$$Y_j = X_j + \sigma W_j \, , \qquad j = 1, \ldots, n \, ,$$

where the random functions X_j and W_j, for $j = 1, \ldots, n$, are totally independent. Here, X_j represents the jth function of interest, which is corrupted by a standard Wiener process W_j with a deterministic scaling factor $\sigma > 0$. Unlike typical measurement error problems, where contamination is due to imprecise measurement or unavoidable perturbation, our data are contaminated artificially; thus, we assume σ is known.

We assume X_j takes a value in $C_{0,0}([0, 1])$; where $C_{0,\ell}([0, 1])$ denotes the set of ℓ-times continuously differentiable (or just continuous when $\ell = 0$) functions f defined on $[0, 1]$, such that $f(0) = 0$. Here, X_j has an

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

unknown probability measure P_X on the Borel σ-field $\mathfrak{B}(C_{0,0}([0,1]))$ of $C_{0,0}([0,1])$, where we equip the space $C_{0,0}([0,1])$ with the supremum norm $\|\cdot\|_\infty$. Throughout, we use the notation $V_j = \sigma W_j$, and use V, W, X, and Y to denote a generic function that has the same distribution as V_j, W_j, X_j, and Y_j, respectively. Critically here, the functional data X_j are assumed to satisfy $X_j(0) = 0$. Indeed, because $W_j(0) = 0$, $Y_j(0) = X_j(0)$, and if the value of X_j at zero is not masked, then individuals can be identified from $Y_j(0)$. In practice, if the raw data do not satisfy $X_j(0) = 0$, they can be pre-masked at zero before the contamination step, for example, by replacing X_j with $\widetilde{X}_j = X_j - X_j(0)$ or $\widetilde{X}_j = X_j w$, where w is a smooth function such that $w(0) = 0$ and $w(1) = 1$.

2.2 Density of contaminated data and differential privacy

In this section, we show that Y_j has a well-defined density with respect to the scaled Wiener measure, and that this density characterizes the distribution of X_j uniquely. Finally, we show that the contamination process ensures differential privacy.

To ensure the existence of a density, we need the following assumption, which we assume throughout this work.

Assumption 1

$X \in C_{0,2}([0,1])$ a.s.

Under Assumption 1, using Girsanov's theorem (Girsanov, 1960), for any Borel measurable mapping φ from $C_{0,0}([0,1])$ to $[0,1]$, we have

$$\begin{aligned}
&E\{\varphi(Y)\} = E\{\varphi(X+\sigma W)\}\\
&\quad = E\Big\{\varphi(\sigma W)\exp\Big(\frac{1}{\sigma}\int_0^1 X'(t)dW(t)\Big)\exp\Big(-\frac{1}{2\sigma^2}\int_0^1 |X'(t)|^2 dt\Big)\Big\}\\
&\quad = E\Big[\varphi(V)E\Big\{\exp\Big(\frac{1}{\sigma^2}\int_0^1 X'(t)dV(t)\Big)\exp\Big(-\frac{1}{2\sigma^2}\int_0^1 |X'(t)|^2 dt\Big)\Big|V\Big\}\Big],
\end{aligned}$$

such that, by integration by parts, we have, almost surely,

$$\begin{aligned}
&\frac{dP_Y}{dP_V}(V) = E\Big[\exp\Big\{\frac{1}{\sigma^2}\int_0^1 X'(t)dV(t) - \frac{1}{2\sigma^2}\int_0^1 |X'(t)|^2 dt\Big\}\Big|V\Big]\\
&\quad = \int \exp\Big\{\frac{1}{\sigma^2}x'(1)V(1) - \frac{1}{\sigma^2}\int_0^1 x''(t)V(t)dt - \frac{1}{2\sigma^2}\int_0^1 |x'(t)|^2 dt\Big\}\, dP_X(x)\,.
\end{aligned} \tag{2.1}$$

Applying the factorization lemma to this conditional expectation, we deduce that there exists a Borel measurable mapping $f_Y\colon C_{0,0}([0,1]) \to \mathbb{R}$, such that $f_Y(V)$ is equal to the right-hand side of (2.1), almost surely. This implies that f_Y is the density of P_Y with respect to P_V. Thus, the contaminated Y_j has a density f_Y. The next theorem establishes its connection with the measure of X_j.

Theorem 1. *The functional density f_Y in* (2.1) *characterizes the probability measure P_X uniquely.*

We deduce from this theorem that inferences about P_X (e.g., goodness-of-fit tests or classification problems; see Section 2.3) can be performed via f_Y. To use this result in practice, it remains to see whether we can estimate f_Y nonparametrically using the data $Y_1, \ldots, Y_n$; see Section 3.

Throughout, we use the notation $\langle \cdot, \cdot \rangle$ for the inner product of $L_2([0,1])$ and $\| \cdot \|_2$ for the corresponding norm, and make the following assumption.

Assumption 2

For some constant $C_{X,1} \in (0, \infty)$, we have that $\|X'\|_2 \leq C_{X,1}, \qquad \text{a.s.}$

The following proposition shows that if the scaling factor σ is sufficiently large, the contaminated data are privatized. For the definition of (α, β)-privacy, refer to Hall et al. (2013); in our setting, this criterion means that

$$P[x + \sigma W \in B] \leq \exp(\alpha) \cdot P[\tilde{x} + \sigma W \in B] + \beta\,, \quad \forall B \in \mathfrak{B}(C_{0,0}([0,1]))\,,$$

for all $x, \tilde{x} \in C_{0,2}([0,1])$, with $\max\{\|x'\|_2, \|\tilde{x}'\|_2\} \leq C_{X,1}$.

Proposition 1. *For any $\alpha, \beta > 0$, choosing $\sigma > 2C_{X,1}\sqrt{2\log(2/\beta)}/\alpha$ guarantees (α, β)-privacy of the observation of $Y = X + \sigma W$ under Assumptions 1 and 2.*

2.3 Applications

The existence of a density for contaminated data has important practical applications. One of them is goodness-of-fit testing. Goodness-of-fit tests for functional data have been considered in, for example, Bugni et al. (2009). In our context, using the observed independent and identically distributed (i.i.d.) contaminated functional data $Y_1, \ldots, Y_n$, the problem consists of testing the null hypothesis $H_0 : X_1 \sim P_X$ versus the alternative $H_1 : X_1 \not\sim P_X$, for some fixed probability measure P_X on $\mathfrak{B}(C_{0,0}([0,1]))$. According to Theorem 1, H_0 is equivalent to the claim that Y_1 has the functional density $f_Y = d(P_X * P_V)/dP_V$. Using the estimator $\hat{f}_Y$ of f_Y introduced in Section 3, we can base a testing procedure on

$$T(Y_1, \ldots, Y_n) := \begin{cases} 1\,, & \text{for } \int \big|\hat{f}_Y(y) - f_Y(y)\big|^2 dP_V(y) \, > \, \rho\,, \\ 0\,, & \text{otherwise,} \end{cases}$$

where ρ is a threshold parameter. In Theorem 2, we derive an upper bound on the mean integrated squared error of our estimator $\hat{f}_Y$. Using the Markov inequality, we deduce that the test can attain any given significance level $\alpha > 0$ if we select ρ larger or equal to the ratio of this upper bound and α. While this gives some insights about ρ, this upper bound does not provide a data-driven rule for selecting ρ in practice. The latter is a difficult problem. For example, it requires deriving the asymptotic distribution of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

the fully data-driven estimator. Another possibility would be to select ρ using a bootstrap approach. However, such a technique would require careful theoretical considerations to ensure the validity of the bootstrap in this context. While these issues are interesting, they go beyond the scope of this study; thus, we leave the practical choice of ρ for future research.

Another interesting application is classification, which, in our context, can be expressed as follows. We observe training contaminated data pairs (Y_i, I_i), for $i = 1, \ldots, n$, where $Y_i = X_i + V_i$, the X_i come from two distinct populations Π_0 and Π_1, and the class label $I_i = k$ if X_i comes from population Π_k, for $k = 0, 1$. The V_i are Wiener processes independent of X_i, and are identically distributed within each population, but the scaling noise parameter σ need not be the same for the two populations. Using these data, the goal is to classify in Π_0 or Π_1 a new random curve $Y = X + V$, where X comes from either Π_0 or Π_1, but has an unknown class label.

It is well known in general classification problems that the optimal classifier is the Bayes classifier, which, adapted to our context, assigns a curve to Π_1 if $E(I|Y = y) > 1/2$, and to Π_0 otherwise. In the case where the probability measures $P_{Y,0}$ and $P_{Y,1}$ of the Y_i that originate from Π_0 and Π_1, respectively, have well-defined densities $f_{Y,0}$ and $f_{Y,1}$, respectively, the Bayes classifier can be expressed as follows: assign Y to Π_1 if $\pi_1 f_{Y,1}(Y) >$

$\pi_0 f_{Y,0}(Y)$, and to Π_0 otherwise, where $\pi_k = P(I = k)$. In the particular Gaussian case, Baíllo et al. (2011) showed that these densities are well defined and showed how to estimate them.

In our case, the Y_i are, in general, not Gaussian, but they have functional densities $f_{Y,k} = dP_{Y,k}/dP_V$, for $k = 0, 1$. Because $P_{X,0} \neq P_{X,1}$ implies that $f_{Y,0} \neq f_{Y,1}$ (see Theorem 1), these densities can be used to classify X from observations on Y in the optimal Bayes classifier. In practice, we classify Y in Π_1 if $\pi_1 \hat{f}_{Y,1}(Y) \geq \pi_0 \hat{f}_{Y,0}(Y)$, and in Π_0 otherwise, where for $k = 0, 1$, $\hat{f}_{Y,k}$ denotes the estimator of $f_{Y,k}$ from Section 3 constructed from the training data Y_i, for which $I_i = k$.

There exist many other classification procedures for functional data, often based on pseudo-densities or finite-dimensional approximations. However, Delaigle and Hall (2012) pointed that, except in the Gaussian case, such projections often do not ensure good finite-sample performance; see, for example, Hall et al. (2001), Ferraty and Vieu (2006), Escabias et al. (2007), Preda et al. (2007), and Shin (2008). See also Dai et al. (2017) for a recent example, where the authors approximate the densities in two populations using the finite-dimensional surrogate densities proposed in Delaigle and Hall (2010); see Delaigle and Hall (2013) for a related classifier.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

3. Methodology

In this section, we consider the problem of estimating the functional density f_Y nonparametrically.

3.1 Existing methods

Several authors have examined nonparametric estimations of a density for stochastic processes in which the probability measure has a Radon–Nikodym derivative with respect to the Wiener measure. In Dabo-Niang (2002, 2004b), the author proposes using a kernel density estimator; see also Prakasa Rao (2010a). This estimator is simple, but it suffers from slow logarithmic convergence rates, which are reflected in its practical performance. Wavelet estimators with polynomial convergence rates were proposed by Prakasa Rao (2010b) and Chesneau et al. (2013), but their conditions are quite technical, and it is not clear how their parameters can be chosen in practice. Moreover, their theory is derived under abstract high-level conditions that might not be satisfied easily in our context.

A simpler estimator is the orthogonal series estimator of Dabo-Niang (2004a), defined as follows. Let $\{\varphi_j\}_{j\in\mathbb{N}}$ denote an orthonormal basis of real-valued functions of $[0, 1]$, where each $\varphi_j \in L_2([0, 1])$, and let $(H_j)_{j\geq1}$ denote the scaled Hermite polynomials defined by $H_k(x) = (-1)^k\phi^{(k)}(x)/\{\phi(x)\sqrt{k!}\}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

for all integers $k \geq 0$, where $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$. In addition, for $x \in C_0([0,1])$, let

$$\beta'_{x,\ell} = \int_0^1 \varphi_\ell(t) dx(t) \,. \tag{3.1}$$

Using the results from Cameron and Martin (1947), as $K \to \infty$, the Fourier–Hermite series $(\Psi_{k_1,\ldots,k_K})_{0\leq k_1 \leq K,\ldots,0\leq k_K \leq K}$, where, for $x \in C_0([0,1])$,

$$\Psi_{k_1,\ldots,k_K}(x) \equiv H_{k_1,\ldots,k_K}(\beta'_{x,1},\ldots,\beta'_{x,K}) \equiv \prod_{\ell=1}^{K} H_{k_\ell}(\beta'_{x,\ell}) \,, \tag{3.2}$$

forms an orthonormal basis of the Hilbert space of all square-integrable $C_0([0,1])$-valued random variables with respect to the Wiener measure. Motivated by this, the author proposes to estimate the Wiener density f_T of the functional data $T_1,\ldots,T_n$ (that have a Wiener density) as

$$\hat{f}_T^K(x) = \sum_{k_1,\ldots,k_K=0}^{K} \frac{1}{n}\sum_{j=1}^{n} H_{k_1,\ldots,k_K}(\beta'_{T_j,1},\ldots,\beta'_{T_j,K}) \cdot H_{k_1,\ldots,k_K}(\beta'_{x,1},\ldots,\beta'_{x,K}) \,, \tag{3.3}$$

where K is a smoothing parameter. This estimator is attractive for its simplicity, but has the drawback that the rates derived by Dabo-Niang (2004a) are logarithmic. In the next two sections, we use a two-stage approximation approach (first a sieve approximation of f_Y, followed by an estimator of the approximation) to introduce a different regularization scheme that involves two parameters. This increases the flexibility of the estimator, which, as we shall see, enables us to obtain polynomial convergence rates. Moreover

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

we provide data-driven choices of the basis and the threshold parameters.

3.2 Finite-dimensional approximation of f_Y

Recall from (2.1) that for $V = \sigma W$ with W a standard Wiener process, we have

$$f_Y(V) = E\Big[\exp\Big\{\frac{1}{\sigma^2}\int_0^1 X'(t)dV(t) - \frac{1}{2\sigma^2}\int_0^1 |X'(t)|^2 dt\Big\}\Big| V\Big], \qquad \text{a.s.},$$

and that our goal is to estimate f_Y from the data $Y_1, \ldots, Y_n$. Instead of directly expressing f_Y in the Fourier–Hermite basis in (3.2), we first construct a sieve approximation of f_Y. Then, we express our sieve approximation in the Fourier–Hermite basis (see Section 3.3).

Using the notation $\beta'_{x,\ell} = \int_0^1 \varphi_\ell(t)dx(t)$ from Equation (3.1), where $\{\varphi_j\}_{j\in\mathbb{N}}$ is a real-valued orthonormal basis of $L_2([0,1])$, we can write

$$\int_0^1 X'(t)\, dV(t) - \frac{1}{2}\int_0^1 |X'(t)|^2\, dt = \sum_{j=1}^{\infty} \beta'_{X,j} \cdot \beta'_{V,j} - \frac{1}{2}\sum_{j=1}^{\infty} {\beta'_{X,j}}^2,$$

where the infinite sums should be understood as mean squared limits. Truncating the sums to m terms, with $m \geq 1$ an integer, this suggests that we can approximate $f_Y(V)$ by $f_Y^{[m]}(\beta'_{V,1},\ldots,\beta'_{V,m})$, where, for all $s_1,\ldots,s_m \in \mathbb{R}$,

$$f_Y^{[m]}(s_1,\ldots,s_m) = E\Big\{\exp\Big(\frac{1}{\sigma^2}\sum_{j=1}^m \beta'_{X,j}\cdot s_j - \frac{1}{2\sigma^2}\sum_{j=1}^m {\beta'_{X,j}}^2\Big)\Big\}$$
$$= \exp\Big(\frac{1}{2\sigma^2}\sum_{j=1}^m s_j^2\Big)\int \exp\Big\{-\frac{1}{2\sigma^2}\sum_{j=1}^m \big(s_j - x_j\big)^2\Big\} dP_{X,m}(x_1,\ldots,x_m)\,, \tag{3.4}$$

and $P_{X,m}$ denotes the measure of $(\beta'_{X,1},\ldots,\beta'_{X,m})$.

The following lemma shows that, as long as m is sufficiently large, $f_Y^{[m]}(\beta'_{V,1},\ldots,\beta'_{V,m})$ is a good approximation to $f_Y(V)$, where V denotes a generic $V_i \sim P_V$.

Lemma 1. *Let $\mathfrak{A}_m$ denote the σ-field generated by $\beta'_{V_1,1},\ldots,\beta'_{V_1,m}$. Under Assumptions 1 and 2,*

(a) $f_Y^{[m]}(\beta'_{V_1,1},\ldots,\beta'_{V_1,m}) = E\{f_Y(V_1)|\mathfrak{A}_m\}$ *a.s.*,

(b) we have

$$E\big|f_Y^{[m]}(\beta'_{V_1,1},\ldots,\beta'_{V_1,m}) - f_Y(V_1)\big|^2$$
$$\le \frac{1}{\sigma^2}\cdot\exp\big(C_{X,1}^2/\sigma^2\big)\cdot\Big(\sum_{j,j'>m}\big|\langle\varphi_j,\Gamma_X\varphi_{j'}\rangle\big|^2\Big)^{1/2},$$

where the linear operator $\Gamma_X : L_2([0,1]) \to L_2([0,1])$ is defined by

$$\big(\Gamma_X f\big)(t) = E\Big\{X'(t)\int_0^1 X'(s)f(s)\,ds\Big\}, \qquad t\in[0,1],\ f\in L_2([0,1])\,.$$

Because Γ_X is a self-adjoint and positive-semidefinite Hilbert–Schmidt operator, the upper bound in Lemma 1(b) is finite for any orthonormal basis

$\{\varphi_j\}_j$ of $L_2([0,1])$, and converges to zero as $m \to \infty$. Indeed, Assumption 2 guarantees that $\sum_{j,j'} \big|\langle \varphi_j, \Gamma_X \varphi_{j'}\rangle\big|^2 \le E\|X_1'\|_2^4 \le C_{X,1}^4 < \infty$. If X (and hence X') is centered, then Γ_X coincides with the covariance operator of X'.

3.3 Estimating the sieve approximation of f_Y

Next, we estimate $f_Y^{[m]}$ using a Fourier–Hermite series. For this, let $P_{Y,m}$ and $f_{Y,m}$ denote, respectively, the measure and the m-dimensional Lebesgue density of the observed random vector $(\beta'_{Y_j,1}, \ldots, \beta'_{Y_j,m})$, where

$$\beta'_{Y_j,k} \;=\; \int_0^1 \varphi_k(t)\, dY_j(t) \;=\; \beta'_{X_j,k} + \beta'_{V_j,k}\,, \quad j = 1, \ldots, n;\; k = 1, \ldots, m\,.$$

Let g_σ denote the $N(0, \sigma^2 I_m)$-density, with I_m the $m \times m$-identity matrix, let $L_{2,g_\sigma}(\mathbb{R}^m)$ denote the Hilbert space of Borel measurable functions $f : \mathbb{R}^m \to \mathbb{R}$ that satisfy $\|f\|_{g_\sigma}^2 \equiv \int |f(t)|^2 g_\sigma(t) dt < \infty$, and let $\langle \cdot, \cdot \rangle_{g_\sigma}$ denote the inner product of $L_{2,g_\sigma}(\mathbb{R}^m)$.

It is easy to deduce from (3.4) that

$$f_Y^{[m]}(s_1, \ldots, s_m) \;=\; f_{Y,m}(s_1, \ldots, s_m)/g_\sigma(s_1, \ldots, s_m)\,, \tag{3.5}$$

and it can be proved that $f_Y^{[m]} \in L_{2,g_\sigma}(\mathbb{R}^m)$. Therefore, if $\Psi_1, \Psi_2, \ldots$ is an orthonormal basis of $L_{2,g_\sigma}(\mathbb{R}^m)$, we can write

$$f_Y^{[m]} = \sum_{k=1}^{\infty} \alpha_k \, \Psi_k,$$

$$\alpha_k = \langle \Psi_k, f_Y^{[m]} \rangle_{g_\sigma} = \int \Psi_k(y) f_{Y,m}(y) \, dy = E\{\Psi_k(\beta'_{Y,1}, \ldots, \beta'_{Y,m})\} .$$

Now, the sequence $(H_{k_1,\ldots,k_m})_{k_1,\ldots,k_m \geq 0}$ of functions $H_{k_1,\ldots,k_m}(x_1,\ldots,x_m) = \prod_{j=1}^m H_{k_j}(x_j)$ defined in (3.2) forms an orthonormal basis of $L_{2,g_1}(\mathbb{R}^m)$. Thus, we can take $\Psi_k(\cdot) = H_{k_1,\ldots,k_m}(\cdot/\sigma)$. To estimate $f_Y^{[m]}$, we replace α_k with $\hat\alpha_k = n^{-1} \sum_{j=1}^n \Psi_k(\beta'_{Y_j,1}, \ldots, \beta'_{Y_j,m})$.

Finally, for U a functional random variable independent of $Y_1, \ldots, Y_n$ that has a density with respect to P_V, we define our estimator of $f_Y(U)$ as

$$\begin{aligned} &\hat f_Y^{[m,K]}(U) \\ &= \sum_{k_1,\ldots,k_m \geq 0} \frac{1}{n} \sum_{j=1}^n H_{k_1,\ldots,k_m}(\beta'_{Y_j,1}/\sigma, \ldots, \beta'_{Y_j,m}/\sigma) H_{k_1,\ldots,k_m}\big(\beta'_{U,1}/\sigma, \ldots, \beta'_{U,m}/\sigma\big) \\ &\qquad \times \omega_K(k_1 + \cdots + k_m) \, 1\{k_1 + \cdots + k_m \leq K\} , \quad (3.6) \end{aligned}$$

where $K \geq 0$ is a truncation parameter, and $0 \leq \omega_K(x) \leq 1$ is a continuous function defined on $[0, K]$. The term $\omega_K(k_1 + \cdots + k_m) \, 1\{k_1 + \cdots + k_m \leq K\}$ prevents k_i from being too large, which controls the variability of the estimator. Using wavelet terminology, the function ω_K dictates whether k_i is chosen using a soft or a hard rule. Specifically, a hard rule corresponds to $\omega_K \equiv 1$: here, all k_i summing to at most K are given equal weights, and as K increases, new indices appear and play as big a role as older ones do. For a soft rule, $\omega_K(x)$ is taken to be a smooth decreasing function of x, for example, $\omega_K(x) = 1 - x/(K+1)$; as K increases, new indices start playing a role, but have less weight than the former ones.

A major difference between (3.6) and Dabo-Niang's (2004a) estimator in (3.3) is our regularization scheme: because of the two-step construction of our estimator (a sieve approximation followed by a basis expansion), we do not use all of the indices $(k_1, \ldots, k_K) \in \{0, \ldots, K\}^K$. Instead, we use $(k_1, \ldots, k_m) \in \{0, \ldots, K\}^m$, such that $k_1 + \ldots + k_m \leq K$, and we assign a weight $\omega_K(k_1 + \ldots + k_m)$ to each group of m indices. As shown in the next sections, our use of a second parameter m and the restriction we put on $k_1 + \ldots + k_m$ drastically improve the quality of the estimator, both theoretically and practically. Moreover, in Section 3.4, we introduce a data-driven way of choosing the basis $\{\varphi_j\}_{j\in\mathbb{N}}$ used to construct the coefficients $\beta'_{Y_j,k}$ and $\beta'_{U,k}$.

3.4 Choosing φ_j

To compute our estimator in practice, we need to choose the basis $\{\varphi_j\}_j$ used in (3.1). Lemma 1(b) implies that if we take φ_j equal to the eigenfunctions of Γ_X, ordered such that the sequence of corresponding eigenvalues

$(\lambda_j)_j$ decreases monotonically, then

$$E\big|f_Y^{[m]}(\beta'_{V_1,1},\ldots,\beta'_{V_1,m})-f_Y(V_1)\big|^2 \le \frac{1}{\sigma^2}\cdot\exp\left(C_{X,1}^2/\sigma^2\right)\cdot\Big(\sum_{j>m}\lambda_j^2\Big)^{1/2}.$$

This bound decreases monotonically as m increases, indicating that the first m terms of the basis capture some of the main characteristics of f_Y.

Of course, in practice, Γ_X is unknown, and thus φ_j is unknown. Therefore, we need to estimate Γ_X, but a priori, this does not seem to be an easy task because, up to some mean terms, Γ_X is the covariance function of the first derivative X' of X. If we could observe $X'_1,\ldots,X'_n$, we could use standard covariance estimation techniques, such as those of Hall and Hosseini-Nasab (2006), Mas and Ruymgaard (2015), and Jirak (2016). However, we observe only the contaminated Y_j. If Y_j were differentiable, we could take its derivative and estimate Γ_X and its eigenfunctions, as in the references just cited. However, it is not differentiable, and we cannot take such a simple approach.

Instead, we propose the following approximation procedure. Let $\{\psi_j\}_j$ denote an orthonormal basis of $L_2([0,1])$, and recall that φ_ℓ denotes the eigenfunction of Γ_X with eigenvalue λ_ℓ, where $\lambda_1\ge\lambda_2\ge\cdots$. In the Supplementary Material, we show that, for all $k\ge 1$,

$$\sum_{j=1}^{\infty}\varphi_{\ell,j}\,\langle\psi_k,\Gamma_X\psi_j\rangle \;=\; \lambda_\ell\,\varphi_{\ell,k}\,, \tag{3.7}$$

where $\varphi_{\ell,j} = \langle \varphi_\ell, \psi_j \rangle$. If we take ψ_j as continuously differentiable such that $\psi_j(0) = \psi_j(1) = 0$, for example, if $\{\psi_j\}_j$ is the Fourier sine basis, then for $j, k = 1, 2, \ldots$, we have

$$\langle \psi_k, \Gamma_X \psi_j \rangle = \mathcal{M}_{j,k} - \sigma^2 \cdot 1\{j = k\}\,, \tag{3.8}$$

where $\mathcal{M}_{j,k} = \int_0^1 \psi_j'(t) \int_0^1 E\big\{Y(t)Y(s)\big\}\psi_k'(s)\,ds\,dt$ (see the proof in the Supplementary Material). We propose approximating φ_ℓ using $\sum_{j=1}^{M} \hat{\varphi}_{\ell,j}\psi_j$, with M a large positive integer, where $\hat{\varphi}_{\ell,j}$ denotes an estimator of $\varphi_{\ell,j}$. Next, we show how to compute $\hat{\varphi}_{\ell,1}, \ldots, \hat{\varphi}_{\ell,M}$ from our data. First, combining (3.7) and (3.8), we have $\sum_{j=1}^{\infty} \varphi_{\ell,j}\,\big(\mathcal{M}_{j,k} - \sigma^2 \cdot 1\{j = k\}\big) = \lambda_\ell\,\varphi_{\ell,k}$, such that

$$\sum_{j=1}^{M} \varphi_{\ell,j}\,\big(\mathcal{M}_{j,k} - \sigma^2 \cdot 1\{j = k\}\big) = \lambda_\ell\,\varphi_{\ell,k} + R_{k,\ell}\,, \tag{3.9}$$

where $R_{k,\ell}$ is a remainder term resulting from the truncation of the sum to M terms. Let I_M and $\mathcal{M}$ denote the $M \times M$-identity matrix and the $M \times M$-matrix with components defined by $\mathcal{M}_{j,k}$, for $j, k = 1, \ldots, M$, respectively, and let $\mathit{\Phi}_\ell = (\varphi_{\ell,1}, \ldots, \varphi_{\ell,M})^T$ and $R_\ell = (R_{1,\ell}, \ldots, R_{M,\ell})^T$. Then, (3.9) implies that $(\mathcal{M} - \sigma^2 I_M)\,\mathit{\Phi}_\ell \;=\; \lambda\,\mathit{\Phi}_\ell + R_\ell$.

Note that $|R_\ell|$ shrinks to zero as $M \to \infty$ because $|R_\ell|^2 \le C_{X,1}^4 \sum_{j>M} |\varphi_{\ell,j}|^2$. Thus, $\big(\mathcal{M} - \sigma^2 I_M\big)\mathit{\Phi}_\ell \approx \lambda_\ell \mathit{\Phi}_\ell$, which motivates us to approximate $\mathit{\Phi}_\ell$ using the unit eigenvector v_ℓ of the matrix $\mathcal{M} - \sigma^2 I_M$ correponding to the

ℓth largest eigenvalue. Now, $\big(\mathcal{M} - \sigma^2 I_M)v_\ell = \lambda_\ell v_\ell$ implies that $\mathcal{M}v_\ell = (\lambda_\ell + \sigma^2)v_\ell$. Thus, v_ℓ is also the eigenvector of $\mathcal{M}$ corresponding to its ℓth largest eigenvalue. Of course, $\mathcal{M}$ is unknown, but it can be estimated as

$$\hat{\mathcal{M}} = \frac{1}{n}\sum_{\ell=1}^{n}\Big\{\int_0^1\int_0^1 \psi_j'(t)Y_\ell(t)Y_\ell(s)\psi_k'(s)ds\,dt\Big\}_{j,k=1,\ldots,M}.$$

For $\ell = 1, \ldots, M$, let $\hat{v}_\ell$ denote the M unit eigenvectors of $\hat{\mathcal{M}}$ (ordered so that the corresponding eigenvalues decrease monotonically). We propose estimating Φ_ℓ as $\hat{\Phi}_\ell = (\hat{\varphi}_{\ell,1}, \ldots, \hat{\varphi}_{\ell,M})^T = \hat{v}_\ell$. Finally, we estimate φ_ℓ as $\hat{\varphi}_\ell = \sum_{j=1}^{M}\hat{\varphi}_{\ell,j}\,\psi_j$.

3.5 Choosing the parameters M, m, and K

To compute the estimator in (3.6) in practice, we need to choose three parameters: M, the parameter used in Section 3.4 to construct the basis functions φ_j employed to compute the projections in (3.1); m, which dictates the dimension of our approximation of f_Y using $f_Y^{[m]}$ in (3.4); and K, the truncation parameter of our orthogonal series expansion in (3.6). Having $\hat{\varphi}_j$ close to the eigenfunctions of Γ_X is likely to give better practical performance, but it is not necessary for the consistency of our estimator. This suggests that the choice of M is not crucial, and we take $M = 20$. In contrast, m and K are important smoothing parameters that influence consistency, and need to be chosen with care. We suggest choosing (m, K)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

by minimizing the cross-validation (CV) criterion

$$\mathrm{CV}(m, K) = \int \left|\hat{f}_Y(v)\right|^2 dP_V(v) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_Y^{(-i)}(Y_i)\,, \tag{3.10}$$

with $\hat{f}_Y^{(-i)}$ defined in the same way as in the estimator in (3.6), except that it uses only the data $Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n$. To compute the integral at (3.10), we generate a large sample (10000, in our numerical work) of V_j from P_V, and approximate the integral using the mean of $|\hat{f}_Y(V_j)|^2$.

As in standard nonparametric density estimation problems, our CV criterion can have multiple local minima, and the global minimum is not necessarily a good choice. In the case of multiple local minima, we choose the one that produces the smallest value of $m + K$. Moreover, when minimizing $CV(K, m)$, we discard all pairs of values (K, m) for which more than 50% of $\hat{f}_Y^{(-i)}$ or $\hat{f}_Y$ are negative. For the (K, m) that remain, we replace each negative $\hat{f}_Y^{(-i)}(Y_i)$ and $\hat{f}_Y(V_j)$ by recomputing these estimators. To do so, we repeatedly replace K with $K-1$ and m with $m-1$ until the negative estimators become positive.

4. Theoretical properties

In this section, we derive the theoretical properties of our estimator. For simplicity, we derive our results in the case where the weight function ω_K in (3.6) is equal to one. Similar results can be established for a more

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

general weight function, but at the expense of more technical proofs. In Section 4.1, we derive an upper bound on the mean integrated squared error of our estimator that is valid for all n. Next, in Section 4.2, we derive the asymptotic properties of our estimator.

4.1 Finite-sample properties

In the next theorem, we give an upper bound on the mean integrated squared error,

$$\mathcal{R}(\hat{f}_Y^{[m,K]}, f_Y) = E \int \left| \hat{f}_Y^{[m,K]}(v) - f_Y(v) \right|^2 dP_V(v),$$

of the estimator in (3.6) when the orthonormal basis $\{\varphi_j\}_j$ and the parameters m and K are deterministic. Our result is nonasymptotic and is valid for all n.

Theorem 2. *Under Assumptions 1 and 2 and the selection $\omega_K \equiv 1$, we have $\mathcal{R}(\hat{f}_Y^{[m,K]}, f_Y) \leq \mathcal{V} + \mathcal{B} + \mathcal{D}$, where*

$$\mathcal{V} = \frac{1}{n} \exp\left(K C_{X,1}^2/\sigma^2\right) \cdot \binom{K+m}{K}, \quad \mathcal{B} = \inf_{h \in \mathcal{H}_{m,K}} \left\| f_Y^{[m]}(\sigma \cdot) - h \right\|_{g_1}^2,$$
$$\mathcal{D} = \frac{1}{\sigma^2} \cdot \exp\left(C_{X,1}^2/\sigma^2\right) \cdot \Big(\sum_{j,j'>m} \left| \langle \varphi_j, \Gamma_X \varphi_{j'} \rangle \right|^2 \Big)^{1/2},$$

and $\mathcal{H}_{m,K}$ denotes the linear hull of $H_{k_1,\ldots,k_m}$ for which $k_1 + \cdots + k_m \leq K$.

In Theorem 2, $\mathcal{V}$ represents a variance term, and $\mathcal{B}$ represents a bias term that depends on the smoothness properties of $f_Y^{[m]}$. Both are typical of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

nonparametric estimators, but the term $\mathcal{D}$ is of a different type. It reflects the error of the finite-dimensional approximation of the density f_Y using the function $f_Y^{[m]}$.

4.2 Asymptotic properties

Next, we derive the asymptotic properties of our density estimator. For this, we need an additional assumption, which we use when dealing with the term $\mathcal{D}$ from Theorem 2:

Assumption 3

There exist constants $C_{X,2}, C_{X,3} \in (0,\infty)$ and $\gamma > 0$, such that

$$\sum_{j,j'>m} \Big| \int_0^1 \varphi_j(s)\big(\Gamma_X \varphi_{j'}\big)(s)ds\Big|^2 \leq C_{X,2} \cdot \exp\big(-C_{X,3} m^\gamma\big)\,, \quad \forall m \in \mathbb{N}\,.$$

For example, if X_1 is centered and $\{\varphi_j\}_j$ is the principal component basis with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ discussed in Section 3.4, then Assumption 3 is satisfied as soon as $\sum_{j=1}^\infty \exp(C'_{X,3} j^\gamma) \cdot \lambda_j^2 < \infty$, for some $C'_{X,3} > C_{X,3}$. In this case, Assumption 3 can be interpreted as an exponential decay of the eigenvalues of Γ_X; specifically, Assumption 3 is satisfied if there exist some $C''_{X,3} > C'_{X,3} > C_{X,3}$ and some $C'''_{X,3} > 0$ such that $\lambda_j \leq C'''_{X,3} \exp(-C''_{X,3} j^\gamma/2)$, for all integer $j \geq 1$.

The next theorem establishes an upper bound on the convergence rates of the mean integrated squared error of our estimator $\hat{f}_Y^{[m,K]}$ as the sample size n tends to infinity. We establish the upper bound uniformly over the class $\mathcal{F}_X = \mathcal{F}_X\big(C_{X,1}, C_{X,2}, C_{X,3}, \gamma, \{\varphi_j\}_j\big)$ of all admitted image measures of X_1, such that Assumptions 1 to 3 are satisfied for some deterministic orthonormal basis $\{\varphi_j\}_j$ of $L_2([0,1])$. The next three theorems consider functions in this class, which implies that they are derived under Assumptions 1 to 3.

Theorem 3. *Assume that $\gamma \in (0,1)$, and select the weight function $\omega_K \equiv 1$ and the parameters K and m, such that $K = K_n = \lfloor \gamma(\log n)/\log(\log n)\rfloor$, $m = m_n = \lfloor \big(C_M \cdot \log n\big)^{1/\gamma}\rfloor$, for some finite constant $C_M > 2/C_{X,3}$. Then, our estimator $\hat{f}^{[m,K]}$ satisfies*

$$\limsup_{n\to\infty} \sup_{P_X \in \mathcal{F}_X} \log\big\{\mathcal{R}\big(\hat{f}_Y^{[m,K]}, f_Y\big)\big\}/\log n \leq -\gamma\,.$$

Theorem 3 shows that the risk of our estimator converges to zero faster than $\mathcal{O}(n^{-\gamma'})$, for any $\gamma' < \gamma < 1$. In particular, our estimator achieves polynomial convergence rates, which is usually impossible in problems of nonparametric functional regression or density estimation. In

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

standard problems of that type, where the data range over an infinite-dimensional space, only logarithmic or sub-algebraic rates can usually be achieved (e.g., Mas, 2012, Chagny and Roche, 2014, and Meister, 2016). In our case, the dimension of the data is infinite as well; however, the density f_Y forms an infinite-dimensional Gaussian mixture, and its smoothness degree is sufficiently high to overcome the difficulty caused by the high dimensionality.

The next theorem provides an asymptotic lower bound for the problem of estimating f_Y nonparametrically. For simplicity, we restrict this to the case where $C_{X,1} = 1$.

Theorem 4. *Assume that $\gamma \in (0,1)$, and let $C_{X,1} = 1$ in Assumption 2. Moreover, assume that the orthonormal basis $\{\varphi_j\}_j$ of $L_2([0,1])$ is such that all φ_j are continuously differentiable. Then, for any sequence $(\hat{f}_n)_n$ of estimators of f_Y computed from the data $Y_1, \ldots, Y_n$, we have*

$$\liminf_{n\to\infty} \sup_{P_X \in \mathcal{F}_X} \log\big\{\mathcal{R}\big(\hat{f}_n, f_Y\big)\big\}/\log n \geq -\gamma + (\gamma-1)^2/(\gamma-2)\,.$$

We learn from the theorem that, in this problem, no nonparametric estimator can reach the parametric squared convergence rate n^{-1}. This is significantly different from the simpler problem of nonparametric estimation of one-dimensional Gaussian mixtures, where the parametric rates are

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

achievable up to a logarithmic factor (see Kim, 2014). Note that the upper bound in Theorem 3 is usually larger than the lower bound in Theorem 4, although the two bounds are very close to each other for γ close to one. Rather than our estimator being suboptimal, we suspect that our lower bound is not sufficiently sharp. Deriving the exact minimax rates seems a very challenging problem, and thus is left to future research.

As is standard in nonparametric estimation problems requiring the choice of smoothing parameters, Theorem 3 is derived under deterministic choices of m and K. Next, using CV, we establish an asymptotic result when $(\hat{m}, \hat{K})$ is chosen using CV as in (3.10), where the minimization is performed over the mesh

$$G = \left\{\lfloor \log n \rfloor, \ldots, \lfloor (\log n)^{1/\gamma_0} \rfloor\right\} \times \left\{1, \ldots, \lfloor (\log n)/\log(\log n) \rfloor\right\}, \quad (4.1)$$

for some constant $\gamma_0 \in (0, \gamma)$. The following theorem shows that the convergence rates from Theorem 3 can be maintained, at least in a weak sense.

Theorem 5. *Our estimator $\hat{f}_Y^{[\hat{m},\hat{K}]}$, where $\omega_K \equiv 1$ and $(\hat{m}, \hat{K})$ is selected using CV over the mesh G in* (4.1)*, satisfies*

$$\lim_{n\to\infty} \sup_{P_X \in \mathcal{F}_X} P\left\{n^\gamma \int \left|\hat{f}_Y^{[\hat{m},\hat{K}]}(x) - f_Y(x)\right|^2 dP_V(x) \geq n^d\right\} = 0\,,$$

for all $\gamma \in [\gamma_0, 1)$ and $d > 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

5. Simulation results

To illustrate the performance of our density estimation procedure, we performed simulations in different settings. For a grid of $T = 101$ points $0 = t_0 < t_1 < \ldots < t_T = 1$ equispaced by $\Delta t = 1/(T-1)$, we generated data $Y_i(t_k) = \sum_{j=1}^{J} \sqrt{\lambda}_j Z_{ik}\ \phi_j(t_k) + \sigma\, W_i(t_k)$, where Z_{ik} is i.i.d., each Z_{ik} is the average of the two independent $U[-.1, .1]$ random variables, $W_i(t_0) = 0$, and, for $k = 1, \ldots, T$, $W_i(t_k) = W_i(t_{k-1}) + \epsilon_{ik}$, where ϵ_{ik} is i.i.d. $\sim N(0, \Delta t)$. We considered five settings: (i) $J = 20$, $\sigma = 0.1$, $\lambda_j = \exp(-j)$, and $\phi_j(t) = \sqrt{2}\sin(\pi t j)$; (ii) as in (i), but with $J = 40$; (iii) as in (ii), but with $\sigma = 0.075$; (iv) as in (i), but with $\sigma = 0.075$, $\phi_j(t) = \sqrt{2}\cos(\pi t j)\kappa(t)$, $\kappa(t) = 2\exp(10t)/\{1 + \exp(10t)\} - 1$; (v) as in (i), but with $\sigma = 0.075$, $\phi_j(t) = \sqrt{2}\sin(\pi t j)\kappa(t)$.

In each case, we generated $B = 200$ samples of $Y_i(t_k)$ of sizes $n = 500$, 1000, 2000 and 5000. Then, for $b = 1, \ldots, B$, using the bth sample of $Y_i(t_k)$, we computed our density estimator $\hat{f}_Y^{[m,K]}(V)$ in (3.6) for 10^4 functions V generated from the same distribution as σW, where m and K were chosen using CV by minimizing (3.10), and where we took the weight function $\omega_K(x) = 1 - x/(K+1)$. The basis functions φ_j were computed as in Section 3.4, with $M = 20$ and $\psi_j(t) = \sqrt{2}\sin(\pi t j)$; we denote by DM the resulting estimator. Each time the m and K selected using CV produced a negative estimator $\hat{f}_Y(v)$ for a new data curve v, we repeatedly replaced, K with $K-1$ and m with $m-1$ for the new curve until the resulting value of (m, K) was such that $\hat{f}_Y(v) > 0$.

In each case, we also computed the estimator of Dabo-Niang (2004a) with our adaptive basis of φ_j, which we denote by DN. We chose K by minimizing the CV criterion in (3.10), replacing our estimator with this estimator and (m, K) with K. As for our estimator, each time the selected value of K produced a negative estimator for a new curve v, for that curve v, we replaced K with the largest value smaller or equal to K that produced a positive estimator.

We also considered the kernel density estimator of Dabo-Niang (2004b), which requires choosing a bandwidth. We considered several versions of CV and a nearest-neighbour bandwidth version of the estimator. However, we encountered major numerical issues with denominators getting too close to zero, and did not manage to obtain reasonable results. Therefore, we do not consider this estimator in our numerical work.

The results of our simulations are summarized in Table 1, where, for each case and each sample size n, we present 10^4 times the median and the first and third quartiles of the squared error SE $= \{\hat{f}_Y(V) - f_Y(V)\}^2$ computed for the 200×10^4 V values. As expected by the theory, both es-

Table 1: Simulation results for density estimation: $10^4\times$ median [first quartile, second quartile] of 2×10^6 values of the SE.

Model	Method	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
(i)	DM	635[145,2242]	492[120,1660]	395[103,1252]	316[86,953]
	DN	891[171,4122]	800[166,3439]	664[125,2970]	527[100,2271]
(ii)	DM	683[152,2427]	506[123,1732]	409[108,1293]	343[94,1051]
	DN	911[179,4133]	823[168,3568]	659[124,2990]	544[101,2420]
(iii)	DM	1134[237,4538]	898[188,3529]	813[175,3237]	784[165,3197]
	DN	1375[209,8046]	1200[186,7325]	1081[174,6611]	1025[177,5574]
(iv)	DM	908[194,3788]	801[172,3158]	744[154,3135]	590[124,2399]
	DN	1468[232,8351]	1151[183,6878]	1097[190,6514]	1052[196,5460]
(v)	DM	849[187,3287]	751[163,2812]	654[143,2500]	565[122,2273]
	DN	1097[170,6389]	1024[172,5817]	914[160,5133]	865[160,4309]

timators improved as the sample size increased and, overall, our estimator significantly outperformed that of Dabo-Niang (2004a). In Table 2, for our estimator and that of Dabo-Niang (2004a), we also show the average time (in seconds and averaged over 10 simulated examples) required to compute one density estimator and its associated data-driven smoothing parameters on a Windows computer with Intel Xeon processor E5-2643 v4 and 32 GB memory. Recall that our estimator requires choosing two smoothing

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

Table 2: Average computational time (in seconds) to compute one density estimator (including the CV choice of smoothing parameters).

Model	Method	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
(i)	DM	94	114	130	198
	DN	42	46	54	77
(ii)	DM	95	113	135	200
	DN	49	55	68	96
(iii)	DM	102	116	138	218
	DN	50	53	71	97
(iv)	DM	104	110	127	191
	DN	46	47	59	82
(v)	DM	91	130	125	182
	DN	41	47	65	100

parameters m and K using CV, whereas that of Dabo-Niang (2004a) requires choosing one smoothing parameter K. It is unsurprising then that our estimator requires a longer computational time: this is the price for the additional accuracy brought by choosing, in a data-driven way, two parameters instead of one.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

Supplementary Material

The online Supplementary Material provides side results and all proofs.

Acknowledgments

Delaigle's research was supported by a grant and a fellowship from the Australian Research Council. The authors are grateful to the editors and two referees for their helpful and valuable comments.

References

Appell, P. (1988). Sur une classe de polynômes. *Annales scientifiques de l'École Normale Supérieure 2me série* **9**, 119–144.

Baíllo, A., Cuevas, A. and Cuesta-Albertos, J.A. (2011). Supervised classification for a family of Gaussian functional models. *Scand. J. Statist.* **38**, 480–498.

Bugni, F.A., Hall, P., Horowitz, J.L. and Neumann, G.R. (2009). Goodness-of-fit tests for functional data. *Econometrics J.* **12**, 1–18.

Cameron, R. H. and Martin, W. T. (1947). The orthogonal development of nonlinear functionals in series FourierHermite functionals. *Ann. Math.*, **48**, 385–392.

Chagny, G. and Roche, A. (2014). Adaptive and minimax estimation of the cumulative distribution function given a functional covariate. *Elec. J. Statist.* **8**, 2352–2404.

Chesneau, C., Kachour, M. and Maillot, B. (2013). Nonparametric estimation for functional data by wavelet thresholding. *REVSTAT – Statistical Journal* **11**, 211–230.

Ciollaro, M., Genovese, Ch.R. and Wang, D. (2016). Nonparametric clustering of functional data using pseudo-densities. *Elec. J. Statist.* **10**, 2922-2972.

Dabo-Niang, S. (2002). Estimation de la densité en dimension infinie: Application aux processus de type diffusion. *C.R. Acad. Sci. Paris I*, **334**, 213-216.

Dabo-Niang, S. (2003). Density estimation in a separable metric space. *Pub. Inst. Stat. Univ. Paris*, **47**, fasc. 1-2, 3–21.

Dabo-Niang, S. (2004a). Density estimation by orthogonal series in an infinite dimensional space: application to processes of diffusion type I. *J. Nonparametric Statistics*, **16**, 171–186.

Dabo-Niang, S. (2004b). Kernel density estimator in an infinite-dimensional space with a rate of convergence in the case of diffusion process. *Applied Mathematics Letters* **17**, 381-386.

Dabo-Niang, S. and Yao, A.-F. (2013). Kernel spatial density estimation in infinite dimension space. *Metrika* **76**, 19–52.

Dai, X., Müller, H.-G. and Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika* **104**, 545–560.

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *Ann. Statist.* **38**, 1171–1193.

Delaigle, A. and Hall, P. (2012). Achieving near-perfect classification for functional data. *J. Roy. Statist. Soc., Ser. B* **74**, 267–286.

Delaigle, A., Hall, P. and Bathia, N. (2012). Componentwise classification and clustering of functional data. *Biometrika*, **99**, 299–313.

Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *J. Amer. Statist. Assoc.* **108**, 1269–1283.

Escabias, M., Aguilera, A.M. and Valderrama,M.J. (2007). Functional PLS logit regression model. *Comput. Statist. Data Anal.* **51**, 4891–4902.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*, Springer.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

Girsanov, I. V. (1960). On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theo. Probab. Appl.* **5**, 285-301.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. Roy. Statist. Soc., Ser B* **68**,109–126.

Hall, P., Poskitt, D. and Presnell, B. (2001). A functional data-analytic approach to signal discrimination. *Technometrics* **43**, 1–9.

Hall, R., Rinaldo, A. and Wasserman, L. (2013). Differential privacy for functions and functional data. *J. Mach. Learn. Research* **14**, 703–727.

Jirak, M. (2016). Optimal eigen expansions and uniform bounds. *Probab. Theo. Rel. Fields* **166**, 753–799.

Karwa, V. and Slavkovi, A. (2016). Inference using noisy degrees: differentially private β-model and synthetic graphs. *Ann. Statist.* **44**, 87–112.

Kim, A.K.H. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli* **20**, 1802–1818.

Lin, Z., Müller, H.-G. and Yao, F. (2018). Mixture inner product spaces and their application to functional data analysis. *Ann. Statist.* **46**, 370–400.

Mas, A. (2012). Lower bound in regression for functional data by small ball probability representation in Hilbert space. *Elec. J. Statist.* **6**, 1745–1778.

Mas A. and Ruymgaart, F. (2015). High dimensional principal projections. *Complex Anal. Operator Theo.* **9**, 35–63.

Meister, A. (2016). Optimal classification and nonparametric regression for functional data. *Bernoulli* **22**, 1729–1744.

Mirshani, A., Reimherr, M. and Slavkovic, A. (2017). Establishing statistical privacy for functional data via functional densities. *arXiv:1711.06660.*

Prakasa Rao, B.L.S. (2010a). Nonparametric density estimation for functional data by delta sequences. *Braz. J. Probab. Stat.* **24**, 468–478.

Prakasa Rao, B.L.S. (2010b). Nonparametric density estimation for functional data via wavelets. *Comm. Stat. – Theo. Meth.* **39**, 1608–1618.

Preda, C., Saporta, G. and Leveder, C. (2007). PLS classification of functional data. *Comput. Statist.* **22**, 223–235.

Shin, H. (2008). An extension of Fisher's discriminant analysis for stochastic processes. *J. Mult. Anal.* **99**, 1191–1216.

Wasserman, L., and Zhou, S. (2010). A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* **105**, 375–389.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0484

Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W. and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin* **38**, 531–538.

Aurore Delaigle

School of Mathematics and Statistics and Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers,

University of Melbourne, Australia.

E-mail: aurored@unimelb.edu.au

Alexander Meister

Institut für Mathematik,

Universität Rostock,

D-18051 Rostock, Germany.

E-mail: alexander.meister@uni-rostock.de