Quantifying the Closeness to a Set of Random Curves via the Mean Marginal Likelihood

Cédric Rommel, Frédéric Bonnans, Baptiste Gregorutti, Pierre Martinon

Quantifying the Closeness to a Set of Random Curves via the Mean Marginal Likelihood

Cédric Rommel[a,b,c,*], Frédéric Bonnans[a,b], Baptiste Gregorutti[c], Pierre Martinon[a,b]

[a] *INRIA Saclay, 1 Rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France*
[b] *CMAP, Ecole Polytechnique, route de Saclay, 91128 Palaiseau, France*
[c] *Safety Line, 130 rue de Lourmel, Paris 75015, France*

Abstract

In this paper, we tackle the problem of quantifying the closeness of a newly observed curve to a given sample of random functions, supposed to have been sampled from the same distribution. We define a probabilistic criterion for such a purpose, based on the marginal density functions of an underlying random process. For practical applications, a class of estimators based on the aggregation of multivariate density estimators is introduced and proved to be consistent. We illustrate the effectiveness of our estimators, as well as the practical usefulness of the proposed criterion, by applying our method to a dataset of real aircraft trajectories.

Keywords: density estimation, functional data analysis, trajectory discrimination

1. Introduction

Functional Data Analysis (FDA) has received an increasing amount of attention in the last years (e.g. Ramsay and Silverman, 2007), this kind of data being present in many fields of application, such as speech recognition (Ferraty and Vieu, 2003), radar waveforms classification (Dabo-Niang et al., 2007) and aircraft trajectories classification (Nicol, 2013; Gregorutti et al., 2015). In this paper we are interested in the general problem of quantifying how close some newly observed random curve is to a set of random functions, being mainly motivated by the practical task of assessing optimized aircraft trajectories. To our knowledge, this problem has not been studied in the literature.

We choose to adopt a probabilistic point of view, interpreting the original problem as the estimation of the likelihood of observing the new curve, given the sample of previously observed functions. This problem is hence related to the estimation of the probability density of a random variable valued on a function space.

Density estimation has been a longstanding problem in statistics and machine learning. Many parametric and nonparametric techniques have been proposed ever since to address

*Corresponding author
Email addresses: cedric.rommel@inria.fr (Cédric Rommel), frederic.bonnans@inria.fr (Frédéric Bonnans), baptiste.gregorutti@safety-line.fr (Baptiste Gregorutti), pierre.martinon@inria.fr (Pierre Martinon)

it in a finite-dimensional setting. For functional data, density estimation has been studied for example by Dabo-Niang (2004), who proposed an extension of the well-known kernel density estimator. Similarly, Prakasa Rao (2010a) developed a delta-sequence method for functional density estimation.

As the distribution of a functional random variable is hard to grasp and does not present good topological properties because it is defined on sets of a space which is "too large" (see e.g. Jacod (2007); Bosq (2012)), alternatives were proposed for casting this problem into a finite-dimensional setting. For example, Prakasa Rao (2010b) proposed to project the random curves on basis functions, while Hall and Heckman (2002) suggested to study the structure of the distribution of a functional random variable by estimating its modes and density ascent lines. We propose another finite-dimensional approach, by estimating an aggregation of the marginal densities, which reduces to a finite sequence of multivariate density estimation problems.

Core Contribution. After introducing our method in section 2.1, an empirical version of it is presented for practical applications (section 2.2). The obtained statistic is built using marginal density estimators which are shown to be consistent in section 2.3. In section 3, we propose an implementation of our approach using the self-consistent kernel density estimator from Bernacchia and Pigolotti (2011) and extending it to the functional data context. We illustrate the effectiveness of our method, as well as its usefulness as an exploratory analysis tool for functional data, on a dataset of real aircraft trajectories and compare it to more standard approaches (section 4).

2. Mean Marginal Likelihood Estimator

2.1. Mean Marginal Likelihood

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathbb{T} = [0; t_f]$ be an interval of $\mathbb{R}$. We denote by E a compact subset of $\mathbb{R}^d$, $d \in \mathbb{N}^*$, endowed with the Borel σ-field $\mathcal{B}$. Let $Z = (Z_t)_{t\in\mathbb{T}}$ be a random variable valued in $\mathcal{C}(\mathbb{T}, E)$, the set of continuous functions from $\mathbb{T}$ to E, and suppose that a training set of m observations of Z is available: $\mathcal{T} = \{\boldsymbol{z}^1, \dots, \boldsymbol{z}^m\} \subset \mathcal{C}(\mathbb{T}, E)$. We denote by μ_t the marginal distribution of Z_t for any $t \in \mathbb{T}$, and we assume that it has a density f_t relative to the Lebesgue measure on $\mathbb{R}^d$. We assume that $(t, z) \in \mathbb{T} \times E \mapsto f_t(z)$ is continuous. Let $\boldsymbol{y} \in \mathcal{C}(\mathbb{T}, E)$ be some arbitrary new curve that we would like to assess.

Given $t \in \mathbb{T}$, we can interpret the quantity $f_t(\boldsymbol{y}(t))$ as the likelihood of observing $Z_t = \boldsymbol{y}(t)$. By summarizing in some way the infinite collection $\{f_t(\boldsymbol{y}(t)) : t \in \mathbb{T}\}$, which we call the *marginal likelihoods* of $\boldsymbol{y}$ hereafter, we hope to build a global and simple likelihood indicator. The first idea for aggregating these quantities is to average them with respect to time:

$$\frac{1}{t_f}\int_0^{t_f} f_t(\boldsymbol{y}(t))dt. \tag{1}$$

The main problem with this criterion is that it mixes elements from densities which may have very different shapes. Indeed, density values of likely observations at two times $t_1, t_2 \in \mathbb{T}$ may have completely different orders of magnitude. For this reason, we propose to use some continuous scaling map $\psi : L^1(E, \mathbb{R}_+) \times E \to [0; 1]$ prior to averaging:

$$\mathrm{MML}(Z, \boldsymbol{y}) = \frac{1}{t_f}\int_0^{t_f} \psi\left[f_t, \boldsymbol{y}(t)\right] dt, \tag{2}$$

and we call the obtained quantity the *mean marginal likelihood* of $\boldsymbol{y}$ given Z.

A natural choice for ψ is simply the function that normalizes $f_t(\boldsymbol{y}(t))$ over E:

$$\psi[f_t, \boldsymbol{y}(t)] = \frac{f_t(\boldsymbol{y}(t))}{\max\limits_{z \in E} f_t(z)}. \tag{3}$$

However, we can also consider more meaningful scaling maps, such as the *confidence level* at $\boldsymbol{y}(t)$:

Definition 1. Let X be a continuous random variable on E of density function $f \in \mathcal{C}(E)$ and let $a \in \mathbb{R}_+$. We call the *confidence level* of f at a the probability that X lies in a region where its density is lower or equal to a:

$$\psi[f, a] = \int_E f(x) \mathbf{1}_{\{f(x) \leq f(a)\}} dx = \mathbb{P}\left(f(X) \leq f(a)\right). \tag{4}$$

In this case, $\psi[f_t, \boldsymbol{y}(t)]$ corresponds to the probability of Z_t falling outside the smallest confidence region containing $\boldsymbol{y}(t)$, as illustrated in figure 1.

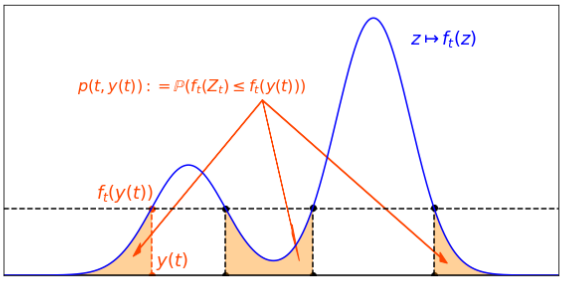


Figure 1: Illustration of the confidence level in the case of a univariate bimodal distribution

A numerical comparison of these two scalings can be found in section 4, while a class of estimators of the mean marginal likelihood is presented in the next section.

2.2. Empirical Version

Usually in FDA, one only has access to discrete observations of the random functions under study. We assume to be in this context: for $1 \leq r \leq m$, each path $\boldsymbol{z}^r$ of the training set $\mathcal{T}$ is assumed to be observed at $n \in \mathbb{N}^*$ discrete times $\{t_1^r < t_2^r < \cdots < t_n^r\} \subset \mathbb{T}$, drawn independently from some random variable T, supposed independent of Z. Hence, we denote by $\mathcal{T}^D$ the set of all discrete observations:

$$\mathcal{T}^D = \{(t_j^r, z_j^r)\}_{\substack{1 \leq j \leq n \\ 1 \leq r \leq m}} \subset \mathbb{T} \times E, \tag{5}$$

where $z_j^r = \boldsymbol{z}(t_j^r)$. Likewise, we assume that the new curve $\boldsymbol{y}$ is observed at $\tilde{n} \in \mathbb{N}$ discrete times $\{\tilde{t}_j\}_{j=1}^{\tilde{n}} \subset \mathbb{T}$ and we denote these observations by

$$\mathcal{Y} = \{(\tilde{t}_j, y_j)\}_{j=1}^{\tilde{n}} \subset \mathbb{T} \times E, \tag{6}$$

where $y_j = \boldsymbol{y}(\tilde{t}_j)$.

Had we enough observations of $\boldsymbol{y}$, we could approximate the integral in (2) by a Riemann sum:

$$\frac{1}{t_f} \sum_{j=1}^{\tilde{n}} \psi[f_j, y_j] \Delta \tilde{t}_j, \tag{7}$$

where $f_j := f_{\tilde{t}_j}$, $\Delta \tilde{t}_j := \tilde{t}_j - \tilde{t}_{j-1}$ and $\tilde{t}_0 = 0$. Yet, the marginal densities $\{f_j\}_{j=1}^{\tilde{n}}$ are unknown and need to be estimated for practical use.

Our approach is based on the idea of partitioning $\mathbb{T}$ into q_m intervals, or *bins*, of same length $b_m = t_f/q_m$. For $1 \leq \ell \leq q_m$, let $\tau_\ell := \tau_0 + \ell b_m$, where $\tau_0 = \inf \mathbb{T} = 0$. We denote by $B_\ell := [\tau_{\ell-1}; \tau_\ell)$ the ℓ^{th} bin for $\ell = 1, \ldots, q_m - 1$ and $B_{q_m} := [\tau_{q_m - 1}; \tau_{q_m}]$. Similarly, $\mathcal{T}_\ell = \{(t_j^r, z_j^r) : t_j^r \in B_\ell\} \subset \mathcal{T}^D$ denotes the set of observations whose sampling time fall into B_ℓ. For some m and $1 \leq j \leq \tilde{n}$, let ℓ be such that $\tilde{t}_j \in B_\ell$. For b_m small enough, we estimate f_j by building a density estimator with the partial data contained in $\mathcal{T}_\ell$. This is done by applying a common statistic $\Theta : \mathcal{S} \to L^1(E, \mathbb{R}_+)$ to $\mathcal{T}_\ell$, where $\mathcal{S}\{(z_k)_{k=1}^N \in E^N : N \in \mathbb{N}^*\}$ denotes the set of finite sequences valued on $E \subset \mathbb{R}^d$: $\hat{f}_j := \Theta[\mathcal{T}_\ell]$. Hence, we consider a single density estimator per bin, averaging along the times in it. We denote this estimated quantities $\{\hat{f}_j\}_{j=1}^{\tilde{n}}$ and by summing them we can build the following plug-in estimator, called the *Empirical Mean Marginal Likelihood* hereafter:

$$\mathrm{EMML}_m(Z, \boldsymbol{y}) := \frac{1}{t_f} \sum_{j=1}^{\tilde{n}} \psi[\hat{f}_j, y_j] \Delta \tilde{t}_j. \tag{8}$$

In the following subsection 2.3, sufficient conditions are given for the consistent estimation of the marginal densities f_t, while section 3 presents a possible class of kernel density estimators to compute $\{\hat{f}_j\}_{j=1}^{\tilde{n}}$.

2.3. Consistency of the marginal density estimations

General case. In this section we state that by using some well chosen statistic to build density estimators in the bins described in section 2.2 we obtain pointwise consistent estimations of the marginal densities of Z. We describe the main ideas of the proof here, while the technical details can be found in the supplementary material. Our consistency result is summarized in theorem 1 and relies on the following 4 assumptions:

Assumption 1. The random variable T is absolutely continuous and $\nu \in L^\infty(E, \mathbb{R}_+)$, its density relative to the Lebesgue measure, satisfies:

$$\nu_+ := \operatorname*{ess\,sup}_{t \in \mathbb{T}} \nu(t) < \infty, \qquad \nu_- := \operatorname*{ess\,inf}_{t \in \mathbb{T}} \nu(t) > 0. \tag{9}$$

Assumption 2. The function defined by

$$(t, z) \in \mathbb{T} \times E \mapsto f_t(z) \tag{10}$$

is continuous on both variables and Lipschitz in time with constant $L > 0$: for any $z \in E$ and $t_1, t_2 \in \mathbb{T}$

$$|f_{t_1}(z) - f_{t_2}(z)| \leq L|t_1 - t_2|. \tag{11}$$

Assumption 3. The homogeneous partition $\{B_\ell^m\}_{\ell=1}^{q_m}$ of $\mathbb{T} = [0; t_f]$, where the bins have size $b_m := t_f/q_m$, is such that

$$\lim_{m\to\infty} b_m = 0, \tag{12}$$

$$\lim_{m\to\infty} m b_m = \infty. \tag{13}$$

Let $\mathcal{S} = \{(z_k)_{k=1}^N \in E^N : N \in \mathbb{N}^*\}$ be the set of finite sequences with values in the compact set $E \subset \mathbb{R}^d$. We also need to assume that the statistic $\Theta : \mathcal{S} \to L^1(E, \mathbb{R}_+)$ used to build the density estimators leads to uniformly consistent density estimations in a standard i.i.d setting, which is summarized in the following assumption:

Assumption 4. Let $\mathcal{G}$ be an arbitrary family of probability density functions on E.Given a density $\rho \in \mathcal{G}$, let S_ρ^N be an *i.i.d* sample of size N valued in $\mathcal{S}$. The estimator obtained by applying Θ to S_ρ^N, denoted by

$$\hat{\rho}^N := \Theta[S_\rho^N] \in L^1(E, \mathbb{R}_+), \tag{14}$$

is a (pointwise) consistent density estimator, uniformly in ρ:

$$\begin{gathered}\text{For all } z \in E, \varepsilon > 0, \alpha_1 > 0, \text{ there is } N_{\varepsilon,\alpha_1} > 0 \text{ such that, for any } \rho \in \mathcal{G}, \\ N \geq N_{\varepsilon,\alpha_1} \Rightarrow \mathbb{P}\left(\left|\hat{\rho}^N(z) - \rho(z)\right| < \varepsilon\right) > 1 - \alpha_1.\end{gathered} \tag{15}$$

For $m \in \mathbb{N}^*$, let $\ell^m : \mathbb{T} \to \mathbb{N}^*$ be the function mapping any point $t \in \mathbb{T} = [0; t_f]$ to the index of the bin containing it:

$$\ell^m(t) := \left\lceil \frac{t}{b_m} \right\rceil. \tag{16}$$

We denote by $\hat{f}^m_{\ell^m(t)}$ the estimator obtained by applying Θ to the subset of data points $\mathcal{T}^m_{\ell^m(t)}$ whose sampling times fall in the bin containing t.

Theorem 1. *Under assumptions 1 to 4, for any $z \in E$ and $t \in \mathbb{T}$, $\hat{f}^m_{\ell^m(t)}(z)$ consistently approximates the marginal density $f_t(z)$ as the number of curves m grows:*

$$\forall \varepsilon > 0, \quad \lim_{m\to\infty} \mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f_t(z)| < \varepsilon\right) = 1. \tag{17}$$

Remark 1. Note that, unlike assumption 4, the convergence in theorem 1 is written in terms of m. This is a big difference since the number of observation points used is supposed to be controlled in the general setting of assumption 4, while it is a random variable in theorem 1 (number of observations falling in the bin $\mathcal{T}^m_{\ell^m(t)}$).

Before explaining the proof of such a theorem, notice that the observations falling into a certain bin for a given number of curves m follow some distribution whose density

function can be explicitly derived. Indeed, for $\mathcal{V} \subset E$ and $B \subset \mathbb{T}$ two compact sets, we have

$$\mathbb{P}\left(Z_T \in \mathcal{V} | T \in B\right) = \frac{\mathbb{P}\left(\{Z_T \in \mathcal{V}\} \bigcap \{T \in B\}\right)}{\mathbb{P}\left(T \in B\right)} = \frac{\int_{\mathcal{V}} \int_B f_t(z)\nu(t)dt \; dz}{\int_B \nu(t)dt}. \tag{18}$$

This shows that $(Z_T | Z_T \in \mathcal{T}^m_{\ell^m(t)}) = (Z_T | T \in B^m_{\ell^m(t)})$ follows a distribution of density

$$f^m_{\ell^m(t)}(z) := \frac{\int_{\tau^m_{\ell^m(t)-1}}^{\tau^m_{\ell^m(t)}} f_t(z)\nu(t)dt}{\int_{\tau^m_{\ell^m(t)-1}}^{\tau^m_{\ell^m(t)}} \nu(t)dt}. \tag{19}$$

The proof of theorem 1 relies on the fact that $f^m_{\ell^m(t)}$ converges pointwise to the marginal density f_t as m tends to infinity. It is indeed quite straightforward to show this by using the assumptions that f_t is Lipschitz in time (11) and that the bin sizes b_m tend to 0 (12). From there, the idea is to try to apply the consistency result from assumption 4 to show that $\hat{f}^m_{\ell^m(t)}$ converges pointwise in probability to $f^m_{\ell^m(t)}$. However, two main difficulties arise here:

1. $\hat{f}^m_{\ell^m(t)}$ is trained using the observations from $\mathcal{T}^m_{\ell^m(t)}$ and the number of elements contained in this subset, denoted by $N^m_{\ell^m(t)}$, is random;
2. we need to train $\hat{f}$ on i.i.d observations whose number tend to infinity in order to apply (15).

The first difficulty can be tackled by conditioning on $N^m_{\ell^m(t)}$. For the second one, we use the fact that, as the bin size tend to 0 and as the number n of observations per curve is fixed with respect to m, than each training subset has, with high probability, at most one observation per curve asymptotically. Hence, because the curves are independent observations of Z, the observations contained in $\mathcal{T}^m_{\ell^m(t)}$ for m large enough will be independently drawn from $f^m_{\ell^m(t)}$ with probability 1. Furthermore, we can show that if the bin size does not decrease too fast, as required by (13), than $N^m_{\ell^m(t)}$ diverges to $+\infty$ in probability. The detailed proof of theorem 1 can be found in the supplementary material.

Example of a kernel estimator with deterministic kernel. In this paragraph we state a stronger consistency result for this particular setting.

$$\hat{f}^m_{\ell^m(t)}(z) = \frac{1}{\sigma N^m_{\ell^m(t)}} \sum_{z_k \in \mathcal{T}^m_{\ell^m(t)}} K\left(\frac{z_k - z}{\sigma}\right) = \frac{1}{N^m_{\ell^m(t)}} \sum_{z_k \in \mathcal{T}^m_{\ell^m(t)}} K_\sigma(z_k - z). \tag{20}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is symmetric kernel summing to 1 and $\sigma > 0$ is the bandwidth, chosen to be scalar here for simplicity. More details on this type of estimators are given in section 3.

We denote the second moments of the random variables of density K_σ and K^2_σ respectively by

$$\sigma^2_{K_\sigma} = \int w^2 K_\sigma(w)dw = \sigma^2 \int w^2 K(w)dw = \sigma^2 \sigma^2_K, \tag{21}$$

$$\sigma^2_{K^2_\sigma} = \int w^2 K_\sigma(w)^2 dw = \sigma \int w^2 K(w)^2 dw = \sigma \sigma^2_{K^2}, \tag{22}$$

and we denote the kernel risk by

$$R(K_\sigma) = \int K_\sigma(w)^2 dw = \frac{1}{\sigma} \int K(w)^2 dw = \frac{1}{\sigma} R(K). \tag{23}$$

For this particular setting, we state in theorem 2 that, under certain conditions, $\hat{f}^m_{\ell^m(t)}(z)$ approximates $f_t(z)$ consistently in expected squared-error, which is stronger than the convergence in probability stated in theorem 1:

Assumption 5. The function $(t, z) \in \mathbb{T} \times E \mapsto f_t(z)$ is $\mathcal{C}^4(E)$ in z and $\mathcal{C}^1(\mathbb{T})$ in t ; the Lipschitz constant of the function

$$t \mapsto \frac{d^2 f_t}{dz^2}(z) := f_t''(z) \tag{24}$$

is denoted by $L'' > 0$: for any $z \in E$ and $t_1, t_2 \in \mathbb{T}$,

$$|f''_{t_1}(z) - f''_{t_2}(z)| \leq L''|t_1 - t_2|. \tag{25}$$

Theorem 2. *Under assumptions 1, 3 and 5, if $\hat{f}^m_{\ell^m(t)}$ is a kernel estimator of the form (20) where the kernel K and the bandwidth $\sigma := \sigma_m$ are deterministic (i.e. do not depend on the data), such that $\sigma_K < \infty$, $\sigma_{K^2} < \infty$, $R(K) < \infty$ and if*

$$\lim_{m\to\infty} \sigma_m = 0, \qquad \lim_{m\to\infty} m b_m \sigma_m = +\infty, \tag{26}$$

then

$$\lim_{m\to\infty} \mathbb{E}\left[(\hat{f}^m_{\ell^m(t)}(z) - f_t(z))^2 \right] = 0. \tag{27}$$

As an example, according to (13) from assumption 3, theorem 2 applies to the case of a Gaussian kernel and a bandwidth $\sigma_m = 1/\sqrt{m b_m}$. Unfortunately, it does not apply to the marginal density estimator presented in the next section, whose kernel is random.

3. Possible Choice of Density Estimator: the Self-Consistent Estimator

In the previous section we presented a general estimator of some discrepancy used to quantify how close a certain curve is to a set of other curves, called the Mean Marginal Likelihood. As explained, such plug-in estimator is based on the aggregation of other consistent density estimators trained on uniform bins. One may wonder what local density estimator to use in this situation.

As most statistical learning problems, density estimation can be tackled in a parametric or a nonparametric setting. In the first case, a specific class of density functions has to be fixed *a priori*, a finite set of unknown parameters needing to be tuned using information contained in the data. Such approaches, as for example Maximum Likelihood estimation, are known to be fast to train and evaluate, they have the best learning rate attainable and are usually very scalable and accurate if the model assumptions are correct. However, the nonparametric density learning techniques make little to no assumptions on the shape of the density to be estimated. As explained in section 2.1, the marginal densities at different times may greatly vary in shape, which is why we preferred to consider nonparametric estimators in this article and more precisely the popular kernel density estimators.

For d-dimensional data, kernel density estimators (KDE) have the following general form when trained on N i.i.d observations $\{x_k\}_{k=1}^N$ of a random variable X of density f:

$$\hat{f}^N(x) = \frac{1}{N \det H} \sum_{k=1}^{N} K\left(H^{-1}(x - x_k)\right). \tag{28}$$

In (28), the function $K : \mathbb{R}^d \to \mathbb{R}_+$ is called a *smoothing kernel* and $H \in GL_d(\mathbb{R})$ is the kernel's *bandwidth* matrix. In general, the main drawback of kernel density estimators (28) lies on the subjective choice of the kernel K and bandwidth H. However, it is well-known folklore in the density estimation literature (see e.g. Wasserman, 2004, chapter 20.3) that KDE's accuracy is not really sensitive to the choice of K and depends mainly on the bandwidth H used. Several rules and heuristics have been suggested since then to choose such a parameter, but they are usually based on quite strict assumptions, such as Silverman's rule of thumb for the estimation of a 1-dimensional Gaussian density (Silverman, 1986). Another possibility is to select H through cross-validation (Stone, 1984) but this approach is computationally intensive, specially if d is larger than 1. For these reasons, we decided to consider a similar method proposed by Bernacchia and Pigolotti (2011), called the *self-consistent density estimator*. It consists indeed in a KDE whose kernel incorporates the bandwidth and is learned directly from the data, hence not requiring any parameter tuning. Its derivation is based on the use of a fixed-point equation to approximate the optimal kernel estimator in the sense of the Mean Integrated Squares-Error (MISE). The obtained estimator takes the form of the Fourier transform of the following characteristic function estimator:

$$\hat{\Phi}_{sc}(s) := \frac{N\Delta(s)}{2(N-1)} \left(1 + \sqrt{1 - \frac{(\Delta_N^{min})^2}{|\Delta(s)|^2}}\right) \mathbf{1}_{A_N}(s),$$

where s is the Fourier variable, N is the number of training observations $\{x_k\}_{k=1}^N$, Δ is the empirical characteristic function

$$\Delta(s) := \frac{1}{N} \sum_{k=1}^{N} e^{i x_k \cdot s}, \qquad \Delta_N^{min} := \sqrt{\frac{4(N-1)}{N^2}}, \tag{29}$$

$i = \sqrt{-1}$ and $\mathbf{1}_{A_N}$ denotes the indicator function over an arbitrary subset $A_N \subset \mathbb{S}_N$ of the frequencies in

$$\mathbb{S}_N := \left\{s : |\Delta(s)|^2 \geq (\Delta_N^{min})^2\right\}. \tag{30}$$

Bernacchia and Pigolotti (2011) proved for 1D data that, under mild assumptions on A_N, the self-consistent estimator converges almost-surely to the true density f as the number of observations N grows and is hence (strongly) consistent. This result is summarized in the following theorem:

Theorem 3 ((Bernacchia and Pigolotti, 2011)). *Let $[-t^*; t^*] = A_N \subset \mathbb{S}_N$ be an interval of frequencies in $\mathbb{R}$. Assuming that f is $L^2(E, \mathbb{R}_+)$ and its Fourier transform $\Phi = \mathcal{F}[f]$ is $L^1(\mathbb{R}, \mathbb{R}^+)$, if the bounds of A_N are such that*

$$\lim_{N\to\infty} t^* = \infty, \qquad \lim_{N\to\infty} \frac{t^*}{\sqrt{N}} = 0, \tag{31}$$

then the density estimator defined by

$$\hat{f}_{sc}^{N}(x) := \mathcal{F}^{-1}[\hat{\Phi}_{sc}](x), \quad \forall x \in E \tag{32}$$

converges almost surely to f as N tends to infinity:

$$\mathbb{P}\left(\lim_{N\to\infty} \hat{f}_{sc}^{N}(x) = f(x)\right) = 1, \quad \forall x \in E. \tag{33}$$

It has been demonstrated through extensive numerical experiments in Bernacchia and Pigolotti (2011); O'Brien et al. (2014); O'Brien et al. (2016) that the self-consistent estimator achieves state-of-the-art MISE accuracy for many types of underlying densities. Furthermore, modern implementations of this estimator proposed by O'Brien et al. (2014); O'Brien et al. (2016)are shown to be several times faster to compute than a regular KDE. This is achieved thanks to the smart use of the Non Uniform Fast Fourier Transform (Greengard and Lee, 2004) to compute the empirical characteristic function Δ. This property is particularly important in our case because of the potentially large number of trainings needed to compute the EMML, which is equal to the number of bins of $\mathbb{T}$'s partition.

For all these reasons, all EMML numerical experiments presented in section 4 make use of this density estimator. More details concerning its derivation and implementation can be found in the supplementary material.

4. Application to the Assessment of Optimized Aircraft Trajectories

4.1. Experiments Motivation

In this section we illustrate our approach on real data recorded from $m = 424$ flights of the same medium haul aircraft $\mathcal{T} = \{\boldsymbol{z}^1, \ldots, \boldsymbol{z}^m\}$, which corresponds to 334 531 observation points. These trajectories are used to estimate the differential system describing the aircraft dynamics. Then, by numerically solving an optimal control problem defined using the estimated aircraft dynamics, a new trajectory $\boldsymbol{y}$ is obtained, supposed to minimize the overall fuel consumption for some future flight (see e.g. Rommel et al., 2017).

Note that:

1. The dynamics model is not guaranteed to be valid outside of the region occupied by the data used to estimate it. Hence, it is natural to want the optimized trajectory to avoid going too far from its validity region.
2. Furthermore, it is desirable for the proposed trajectory not to be too unusual compared to standard climb profiles for better acceptance by the pilots and Air Traffic Control.

The two previous points motivate the need for an indicator of closeness between the optimized trajectory and the set of recorded flights.

4.2. Experiments Design

Training set. The training data used for our experiments were extracted from the *Quick Access Recorder* (QAR) of the same aircraft, whose sampling rate is of one measurement per second. We only used the recordings of 5 variables, which are the altitude h, the true

airspeed V, the path angle γ, the angle of attack α and the throttling position N_1. These variables are not all directly accessible and some were computed from other measurements using standard formulas from flight mechanics (see e.g. Rommel et al., 2017). Only the portion corresponding to the climb phase of these signals was kept for our experiments, i.e. data corresponding to altitudes between 1524 m = 5000 ft and the *top of climb* (cruise altitude, specific to each flight). The training set of curves obtained by the described procedure is displayed on figure 2a.

Test set. In order to evaluate the estimated mean marginal likelihood on relevant examples, the following test flights were considered:

1. 50 real flights, extracted from the training set before training;
2. 50 simulated trajectories which were optimized as described in section 4.1, with constraints keeping the resulting speed V and N_1 between reasonable operational bounds;
3. and another 50 simulated trajectories optimized without the operational constraints.

We evaluated the likelihood of these 150 test trajectories using the EMML and the competing methods described in the following section in order to assess and compare their discriminative power and computation efficiency.

4.3. Alternate Approaches Based on Standard Methods

The problem of quantifying how close a newly observed random curve is with respect to a set of observations from the same stochastic process has not been treated by the statistical learning literature to our knowledge. However, this problem is related to other more standard approaches from Functional Data Analysis and Conditional Density Estimation, which could be adapted quite straightforwardly for this purpose. For this reason, we discuss in the following paragraphs the characteristics of two of these other existing methods, before comparing them to our approach in the numerical results of section 4.5.

4.3.1. Functional Principal Components Analysis

Functional Principal Components Analysis (FPCA) is a standard tool in FDA capable of building a small number of descriptors which summarize the structure of a set of random functions. As explained for example in Nicol (2013), this dimension reduction method can be used to project the train set of infinite-dimensional random trajectories into a finite (low) dimensional space.

Following the same reasoning used to derive the MML, our idea here consists in estimating the density function of these low dimensional representations of the training set. Then, after projecting the new trajectory $\boldsymbol{y}$ into the same descriptors, we can evaluate the density estimate at it and obtain an approximation of its likelihood.

4.3.2. Least-Squares Conditional Density Estimation

From a completely different point of view, we could forget for a moment that we are considering a random process Z and look at (T, Z_T) as a pair of standard random variables valued on the finite dimensional space $\mathbb{T} \times E$. In this case, we could see the marginal densities f_t as the conditional probability density functions

$$f_{Z_T|T}(t, z) = \frac{f_{(T,Z_T)}(t, z)}{f_T(t)}, \qquad (t, z) \in \mathbb{T} \times E. \tag{34}$$

We could hence estimate (34) at the observed points of the new trajectory $\boldsymbol{y}$ and use them to compute the EMML indicator (8). It is however well-known in density ratio estimation that approximating $f_{(T,Z_T)}(t,z)$ and $f_T(t)$ separately before building the ratio in (34) is not a good idea because it magnifies the errors. For this reason, Sugiyama et al. (2010) proposed to use a linear model for this purpose

$$f_{Z_T|T}(t,z) = \theta^\top \phi(t,z), \tag{35}$$

where $\theta = (\theta_1, \ldots, \theta_p)$ is a vector of scalar parameters and $\phi(t,z) = (\phi_1(t,z), \ldots, \phi_p(t,z))$ is a family of nonnegative basis functions. The parameters θ are then chosen so as to minimize a L^2-penalized least-squares criterion, which is shown to have a closed-form solution. This method was coined *Least-Squares Conditional Density Estimation* (LSCDE) by the authors, and is also known as *Unconstrained Least-Squares Importance Fitting* (uLSIF) in the density ratio estimation literature (Kanamori et al., 2009). The extensive numerical results presented in Sugiyama et al. (2010) indicate that this approach have state-of-the-art accuracy in conditional density estimation.

4.4. Algorithms Settings

For all the methods tested, the altitude h played the role of "time". This is a natural assumption made when optimizing the climb profile of a civil airliner, since the altitude is an increasing function of the time and every other variable depends on it. This allowed us to reduce the dimension of our problem from 5 to 4.

MML with Self-Consistent Kernel Estimator settings. The python library FASTKDE (O'Brien et al., 2014; O'Brien et al., 2016) was used to compute the marginal densities from the bins data. It contains the implementation of the Self-Consistent kernel estimator described in section 3. The precision of the density estimations were set to single. The *confidence levels* were approximated by numerical integration using the trapezoidal rule over a fine grid of approximately 300 points per bin.

Concerning bin sizes, we chose to use an uneven partition in our experiments. The reason for this are the *climb-steps* visible in the trajectories between 3000 and 4000 m, which correspond to phases during which the aircraft decreases considerably its ascent speed, leading to slowly increasing altitudes. Such behaviors translate into rapidly increasing speeds V with respect to the altitude, as well as into plummeting values of γ and N_1 (see figure 2a). This brought us to consider tighter bins around these climb-step altitudes:

- between 1524 and 3000m and between 4000 and 12000m, we partitioned the altitudes homogeneously into bins of size $b_m^{(1)} = 21\text{m} \simeq 1/\sqrt{m}$ (which satisfies assumption 3);
- between 3000 and 4000m, we used a bin size twice smaller $b_m^{(2)} = 10\text{m} \simeq b_m^{(1)}/2$;

FPCA settings. Concerning the Functional Principal Components Analysis method, all training and testing flights were resampled on an equispaced grid of altitudes, using a step size of 5m. The trajectories were then centered and decomposed into a basis of 128 cubic B-splines. The SVD decomposition was carried using the `PCA` class from `scikit-learn` python library (Pedregosa et al., 2011). We kept 4 components for each variable (V, γ, α and N_1), which was enough to explain more than $90\%, 65\%, 60\%$ and 75% of their respective variance. A Gaussian mixture model was used to estimate the density

of the training trajectory scores obtained by the projection into the principal functions. The model was trained using a standard EM algorithm, implemented in `scikit-learn` as well. The number of components was selected between 1 and 5 using the Bayesian information criterion (BIC).

LS-CDE settings. For the Least-Squares Conditional Density Estimation, the python package `densratio` (Makiyama, 2016), implementing the uLSIF method from Kanamori et al. (2009) was used and adapted. The basis functions chosen were $p = 100$ Gaussian kernels with the same variance σ and different centers. These centers were randomly drawn from the training data points using a uniform distribution, as suggested in Sugiyama et al. (2010). The variance σ, as well as the L^2 penalty weight λ needed for minimizing the least-squares criterion were selected by cross-validation.

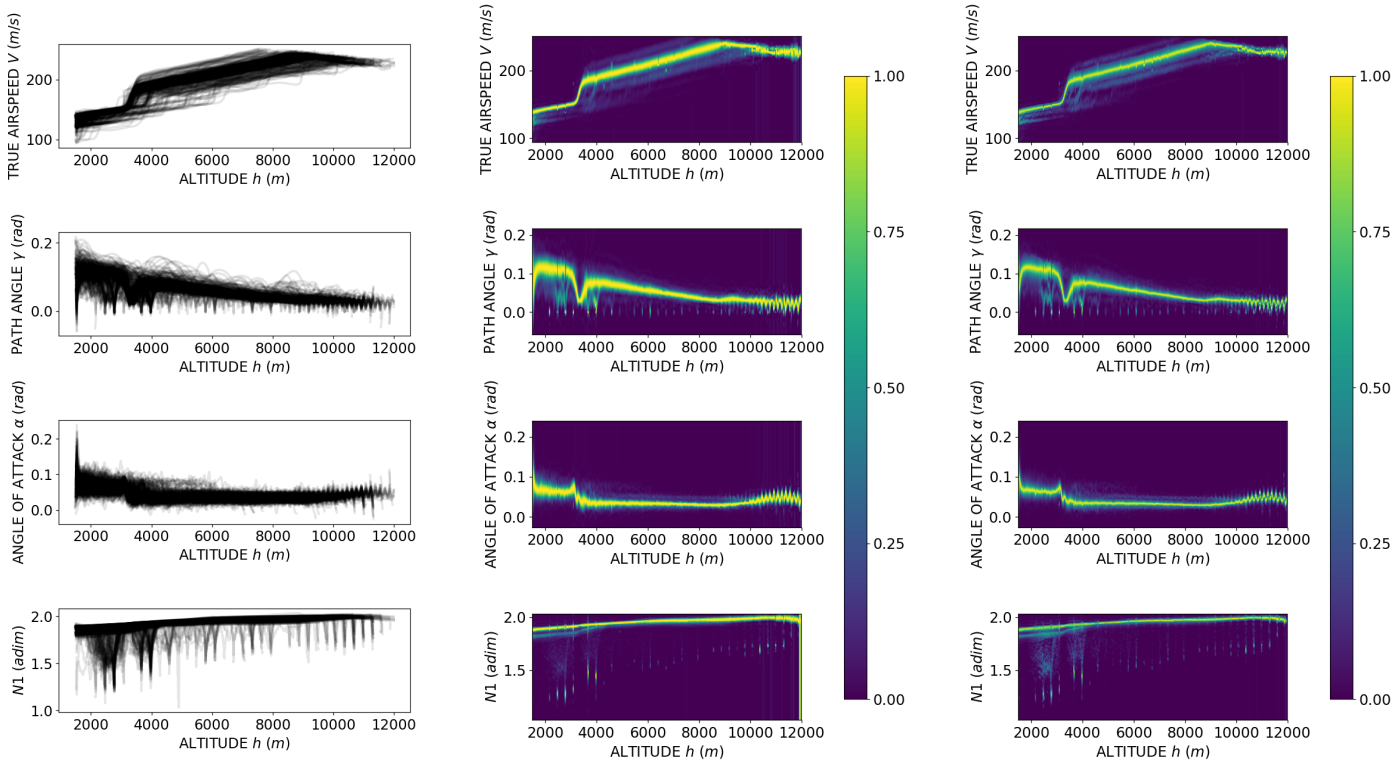
4.5. Results and Comments

Figures 2b and 2c show heatmaps encoding the estimated marginal likelihoods using the normalized density (3) and the confidence level (4). We notice that both figures are similar and seem to catch the shape of the plot from figure 2a, including the multi-modalities visible for N_1 below $h = 4000$m for example.

Table 1a contains the estimated Mean Marginal Likelihood scores averaged over each test flight category. The training was carried on each dimension separately and the average total training time was of 5 seconds on a laptop (2.30 GHz, 7.7 GB). First of all, we notice for both types of scaling functions that the test flight categories are nicely separated by three really distinct ranges of scores. Furthermore, the higher differences between the two types of optimized flights can be seen for the variables V and N_1, which makes sense since those are the variables which are left free for *Opt2* flights and constrained for *Opt1* flights. As expected from figures 2b and 2c, the performances of both confidence level and normalized density based MML are comparable in terms of discrimination power and seem adequate for the task of assessing optimized aircraft climb trajectories.

Table 1b contains the estimated Mean Marginal Likelihood scores in a 2-dimensional setting, where the pairs (V, γ) and (α, N_1) have been treated together. The average training time needed here was 16 times larger than in the 1D case, i.e. 1 minute 20 seconds. The scores observed are globally really low and the test flight categories are not well separated. Moreover, we expected to obtain large scores for the real flights, since we used the marginal densities of their category to build the criterion, but this is not the case here. We conclude that the MML criteria based on the self-consistent estimator does not work so well in higher dimension and we suspect this to be related to the curse of dimensionality. Indeed, it is well-known (see e.g. Wasserman, 2004, chapter 21.3) that as the dimension grows, the amount of data needed to attain a given accuracy with kernel density estimators skyrockets, which may explain this poor performance.

From a practical point of view, a reference value or threshold is needed if one wanted to use our method to determine automatically whether a given optimized flight should be accepted or not. In such a context, a quite straightforward solution would be to use a leave-one-out cross-validation approach: compute the MML score of each real flight leaving it out of the training data and then averaging over the obtained scores. These reference values have been computed for our dataset and are summarized in table 2. We note that the values obtained are very close to the average scores of the real flights showed in table 1a.



(a) Training set of 424 trajectories

(b) Normalized density heatmap

(c) Confidence level heatmap

Figure 2: Estimated marginal densities using two types of scaling functions

Table 1: Average and standard deviation of the Mean Marginal Likelihood scores using confidence level and normalized density for 50 real flights (*Real*), 50 optimized flights with operational constraints (*Opt1*) and 50 optimized flights without constraints (*Opt2*).

(a) 1-dimensional case

Var.	Confidence level			Normalized Density		
	Real	Opt1	Opt2	Real	Opt1	Opt2
V	0.52 ± 0.16	0.38 ± 0.14	0.15 ± 0.09	0.63 ± 0.16	0.45 ± 0.16	0.17 ± 0.10
γ	0.54 ± 0.09	0.24 ± 0.12	0.22 ± 0.09	0.67 ± 0.09	0.33 ± 0.17	0.29 ± 0.11
α	0.53 ± 0.06	0.08 ± 0.05	0.02 ± 0.01	0.65 ± 0.07	0.10 ± 0.06	0.02 ± 0.01
N_1	0.47 ± 0.24	0.71 ± 0.00	0.03 ± 0.01	0.57 ± 0.27	0.83 ± 0.01	0.04 ± 0.02
Mean	0.52 ± 0.07	0.35 ± 0.06	0.10 ± 0.02	0.63 ± 0.07	0.43 ± 0.08	0.13 ± 0.02

(b) 2-dimensional case

Var.	Confidence level			Normalized Density		
	Real	Opt1	Opt2	Real	Opt1	Opt2
(V, γ)	0.09 ± 0.05	0.11 ± 0.05	0.03 ± 0.03	0.05 ± 0.03	0.06 ± 0.03	0.01 ± 0.01
(α, N_1)	0.03 ± 0.02	0.01 ± 3e-3	0.01 ± 2e-3	0.02 ± 0.01	4e-3 ± 2e-3	3e-3 ± 1e-3
Mean	0.06 ± 0.03	0.06 ± 0.02	0.02 ± 0.01	0.03 ± 0.02	0.03 ± 0.02	0.01 ± 0.01

Table 2: Leave-one-out cross-validated MML scores of the training trajectories.

Var.	Confidence level	Normalized Density
V	0.50 ± 0.19	0.60 ± 0.22
γ	0.51 ± 0.13	0.63 ± 0.15
α	0.51 ± 0.15	0.62 ± 0.17
N_1	0.51 ± 0.22	0.61 ± 0.24
Mean	0.51 ± 0.21	0.62 ± 0.22

Table 3a contains the scores obtained using the Functional PCA based method presented in section 4.3.1. The training time needed here was of 20 seconds in average. As for the MML in 2D, the real flights' scores are surprisingly low and the two types of simulated trajectories are not well discriminated by the criterion. This might be caused by the fact that this method encodes each training trajectory by a single point in the 4-dimensional space spanned by the principal functions. The training set obtained is hence of $m = 424$ points, which might be too small too attain sufficient accuracy from the Gaussian mixture density estimator in such a high dimension. The principal functions used and scatter plots of the projected trajectories can be found in the online version.

Concerning the LS-CDE approach, because the algorithm needs large Gram matrices (of size $O(nm^2)$) to be stored, we encountered several memory problems when trying to run it on our dataset of 334 531 observation points. For this reason, our results were obtained by applying it to 100 uniform batches. These batches were obtained by partitioning the data according to the altitude. Although the three categories are well-separated by this

method, as shown on table 3b, the total time needed to train the estimators on every batch was approximately 14 hours.

We didn't test both alternate methods in the 2D setting since the problems observed in 1D (curse of dimensionality for FPCA and memory/time for LS-CDE) would be aggravated.

Table 3: Average and standard deviation of the normalized density scores using Functional PCA and Least-Squares Conditional Density Estimation of 50 real flights (*Real*), 50 optimized flights with operational constraints (*Opt1*) and 50 optimized flights without constraints (*Opt2*).

(a) FPCA

Var.	Real	Opt1	Opt2
V	0.15 ± 0.22	4.9e-04 ± 9.0e-04	2.1e-05 ± 8.1e-05
γ	0.20 ± 0.22	9.3e-03 ± 1.5e-02	1.4e-02 ± 2.2e-02
α	0.28 ± 0.28	1.2e-05 ± 1.8e-05	7.0e-08 ± 1.7e-07
N_1	7.6e-03 ± 6.1e-03	1.6e-02 ± 2.3e-04	1.3e-06 ± 6.7e-07
Mean	0.16 ± 0.12	6.4e-03 ± 3.8e-03	3.6e-03 ± 5.4e-03

(b) LS-CDE

Var.	Real	Opt1	Opt2
V	0.81 ± 0.13	0.63 ± 0.11	0.40 ± 0.23
γ	0.65 ± 0.05	0.55 ± 0.10	0.53 ± 0.08
α	0.91 ± 0.02	0.74 ± 0.03	0.68 ± 0.01
N_1	0.72 ± 0.10	0.79 ± 0.01	0.35 ± 0.05
Mean	0.77 ± 0.05	0.68 ± 0.04	0.49 ± 0.06

In conclusion, our numerical results indicate that the MML criterion has better discriminative power than FPCA and LS-CDE for the task of assessing curves relatively to a set of "good" examples. Furthermore, the training time and memory needed for using LS-CDE in datasets of this size seems crippling. Concerning the FPCA method, it does not seem to be applicable to datasets with so few curves and present a higher training time than MML.

5. Conclusions

In this paper we proposed a new approach for a problem which seemed unaddressed by the statistical learning community: quantifying the closeness from a curve to a set of random functions. We introduced a class of probabilistic criteria for this context called the Mean Marginal Likelihood (MML), and analyzed two possible scaling functions used to build them. We also derived a class of estimators of our criteria, which make use of local density estimators proved to consistently approximate the marginal densities of a random process. For practical applications, we suggested a particular flexible density estimator believed to have the right properties needed in this setting, called the self-consistent kernel estimator.

Numerical experiments using real aircraft data were carried to compare the MML with other well-established approaches from Functional Data Analysis and Conditional Density Estimation. The results show that, although the MML does not take into account the temporal structure of the data as other standard functional data analysis methods, it is a good candidate for the type of applications suggested. This seems to be especially the case if the number of training trajectories is too small for using FPCA or if the total number of observation points is too large for conditional density estimation. Moreover, the training times obtained for MML are by far the shortest among the compared methods, which confirms the relevance of the self-consistent kernel estimator. Furthermore, the ease to visualize, localize and interpret discrepancy zones allowed by MML make it a good exploratory analysis tool for functional data (see e.g. figure 2). We also note that our method does not perform as well in a multidimensional setting, but that the training time should not be an obstacle. In future work we intend to test the MML with parametric density estimators, which should be less affected by the curse of dimensionality.

References

Bernacchia, A., Pigolotti, S., 2011. Self-consistent method for density estimation. Journal of the Royal Statistical Society 73 (3), 407–422.

Bosq, D., 2012. Nonparametric statistics for stochastic processes: estimation and prediction. Vol. 110. Springer Science & Business Media.

Cooley, J. W., Tukey, J. W., 1965. An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation 19, 297–301.

Cribari-Neto, F., L. P. Vasconcellos, K., L. Garcia, N., 2000. A note on inverse moments of binomial variates. Brazilian Review of Econometrics 20, 269–277.

Dabo-Niang, S., 2004. Kernel density estimator in an infinite-dimensional space with a rate of convergence in the case of diffusion process. Applied mathematics letters 17 (4), 381–386.

Dabo-Niang, S., Ferraty, F., Vieu, P., 2007. On the using of modal curves for radar waveforms classification. Computational Statistics & Data Analysis 51 (10), 4878–4890.

Dutt, A., Rokhlin, V., 1993. Fast Fourier Transforms for Nonequispaced Data. SIAM Journal on Scientific Computing 14 (6), 1368–1393.

Ferraty, F., Vieu, P., 2003. Curves discrimination: a nonparametric functional approach. Computational Statistics & Data Analysis 44 (1-2), 161–173.

Glad, I. K., Hjort, N. L., Ushakov, N. G., 2003. Correction of density estimators that are not densities. Scandinavian Journal of Statistics 30 (2), 415–427.

Greengard, L., Lee, J.-Y., 2004. Accelerating the Nonuniform Fast Fourier Transform. SIAM Review 46 (3), 443–454.

Gregorutti, B., Michel, B., Saint-Pierre, P., 2015. Grouped variable importance with random forests and application to multiple functional data analysis. Computational Statistics & Data Analysis 90, 15–35.

Hall, P., Heckman, N. E., 2002. Estimating and depicting the structure of a distribution of random functions. Biometrika 89 (1), 145–158.

Jacod, J., 2007. Lecture notes on "Mouvement brownien et calcul stochastique".

Kanamori, T., Hido, S., Sugiyama, M., 2009. A least-squares approach to direct importance estimation. Journal of Machine Learning Research 10 (Jul), 1391–1445.

Makiyama, K., Dec. 2016. densratio, A Python Package for Density Ratio Estimation. Dowloaded on may 4th 2018.
URL `https://github.com/hoxo-m/densratio_py`

Nicol, F., 2013. Functional principal component analysis of aircraft trajectories. In: Proceedings of the 2nd International Conference on Interdisciplinary Science for Innovative Air Traffic Management (ISIATM).

O'Brien, T. A., Collins, W. D., Rauscher, S. A., Ringler, T. D., 2014. Reducing the Computational Cost of the ECF Using a nuFFT: A Fast and Objective Probability Density Estimation Method. Computational Statistics & Data Analysis 79, 222–234.

O'Brien, T. A., Kashinath, K., Cavanaugh, N. R., Collins, W. D., O'Brien, J. P., 2016. A fast and objective multidimensional kernel density estimation method: fastkde. Computational Statistics & Data Analysis 101, 148–160.

Pedregosa, F., et al., 2011. Scikit-learn: Machine learning in Python. JMLR 12, 2825–2830.

Prakasa Rao, B. L. S., 2010a. Nonparametric density estimation for functional data by delta sequences. Brazilian Journal of Probability and Statistics 24 (3), 468–478.

Prakasa Rao, B. L. S., 2010b. Nonparametric density estimation for functional data via wavelets. Communications in Statistics—Theory and Methods 39 (8-9), 1608–1618.

Ramsay, J. O., Silverman, B. W., 2007. Applied functional data analysis: methods and case studies. Springer.

Rommel, C., Bonnans, J. F., Gregorutti, B., Martinon, P., 2017. Aircraft dynamics identification for optimal control. In: Proceedings of the 7th European Conference for Aeronautics and Aerospace Sciences.

Scott, D. W., 2015. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons.

Silverman, B. W., 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall.

Stone, C. J., 1984. An asymptotically optimal window selection rule for kernel density estimates. The Annals of Statistics 12 (4), 1285–1297.

Sugiyama, M., Takeuchi, I., Suzuki, T., Kanamori, T., Hachiya, H., Okanohara, D., 2010. Conditional density estimation via least-squares density ratio estimation. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 781–788.

Tsybakov, A. B., 2008. Introduction to Nonparametric Estimation, 1st Edition. Springer.
Wasserman, L., 2004. All of statistics: a concise course in statistical Inference. Springer texts in statistics. Springer.
Watson, G. S., Leadbetter, M. R., 1963. On the estimation of the probability density. The Annals of Mathematical Statistics 34 (2), 480–491.

Appendices

A. Marginal densities consistency proof

In this section we prove theorem 1 from section 2.3. It relies on 5 lemmas stated and proved hereafter.

Lemma 1. *Let $t \in \mathbb{T}$ and $z \in E$. Under assumptions 2 and 3, $f^m_{\ell^m(t)}(z)$ converges to $f_t(z)$:*

$$\lim_{m\to\infty} |f_t(z) - f^m_{\ell^m(t)}(z)| = 0. \tag{36}$$

Proof.

$$\begin{aligned} |f_t(z) - f^m_{\ell^m(t)}(z)| &= \left| f_t(z) - \frac{\int_{\mathcal{B}^m_{\ell^m(t)}} f_s(z)\nu(s)ds}{\int_{B^m_{\ell^m(t)}} \nu(s)ds} \right|, \\ &= \frac{\left| \int_{\mathcal{B}^m_{\ell^m(t)}} (f_t(z) - f_s(z))\nu(s)ds \right|}{\int_{B^m_{\ell^m(t)}} \nu(s)ds}, \\ &\leq \frac{\int_{\mathcal{B}^m_{\ell^m(t)}} |f_t(z) - f_s(z)|\nu(s)ds}{\int_{B^m_{\ell^m(t)}} \nu(s)ds}. \end{aligned} \tag{37}$$

According to assumption 2,

$$|f_t(z) - f_s(z)| \leq L|t-s| \leq L|\tau^m_{\ell^m(t)} - \tau^m_{\ell^m(t)-1}| = Lb_m. \tag{38}$$

Hence,

$$|f_t(z) - f^m_{\ell^m(t)}(z)| \leq Lb_m \frac{\int_{\mathcal{B}^m_{\ell^m(t)}} \nu(s)ds}{\int_{B^m_{\ell^m(t)}} \nu(s)ds} = Lb_m. \tag{39}$$

Since $b_m \to 0$ by assumption 3, the conclusion follows. □

Lemma 2. *Let $m \in N^*$ and $t \in \mathbb{T}$. Under assumption 3, the probability that $\mathcal{T}^m_{\ell^m(t)}$ (the subset of training points whose sampling time fall in the bin containing t) contains at most one observation point per curve is asymptotically equal to 1, meaning that for large enough m, the observations in $\mathcal{T}^m_{\ell^m(t)}$ will be independent with high probability:*

$$\lim_{m\to\infty} \mathbb{P}(N^m_{r,\ell^m(t)} \leq 1) = 1, \quad r = 1, \ldots, m, \tag{40}$$

where $N^m_{r,\ell^m(t)}$ denotes the number of observations of $\boldsymbol{z}^r$ in $\mathcal{T}^m_{\ell^m(t)}$.

Proof. Let $1 \leq r \leq m$ and

$$\mathcal{T}^m_{r,\ell^m(t)} := \{(t^r_k, z^r_k) : t^r_k \in B^m_{\ell^m(t)}; 1 \leq k \leq n\} \tag{41}$$

be the set of observations of the r^{th} curve with times lying in the $\ell^m(t)^{th}$ bin $B^m_{\ell^m(t)}$. Let $N^m_{r,\ell^m(t)}$ be the number of elements in $\mathcal{T}^m_{\ell^m(t)}$. The random variable $N^m_{r,\ell^m(t)}$ follows a binomial law $\mathcal{B}(n, P^m_{\ell^m(t)})$, where

$$P^m_{\ell^m(t)} = \int_{B^m_{\ell^m(t)}} \nu(t)dt \tag{42}$$

is the probability of a new observation of the r^{th} curve falling in $\mathcal{T}^m_{r,\ell^m(t)}$. The probability of $N^m_{r,\ell^m(t)}$ being at most equal to 1 writes

$$\begin{aligned}\mathbb{P}(N^m_{r,\ell^m(t)} \leq 1) &= \sum_{k=0}^{1} \left(\begin{array}{c} n \\ k \end{array} \right) (P^m_{\ell^m(t)})^k (1 - P^m_{\ell^m(t)})^{n-k}, \\ &= (1 - P^m_{\ell^m(t)})^n + nP^m_{\ell^m(t)}(1 - P^m_{\ell^m(t)})^{n-1}, \\ &= (1 - P^m_{\ell^m(t)})^{n-1}(1 + (n-1)P^m_{\ell^m(t)}).\end{aligned} \tag{43}$$

As

$$P^m_{\ell^m(t)} = \int_{B^m_{\ell^m(t)}} \nu(t)dt \leq b_m \nu_+, \tag{44}$$

we have

$$\mathbb{P}(N^m_{r,\ell^m(t)} \leq 1) \geq (1 - P^m_{\ell^m(t)})^{n-1} \geq (1 - b_m\nu_+)^{n-1}. \tag{45}$$

Since $b_m \to 0$ according to assumption 3, we obtain

$$\lim_{m\to\infty} \mathbb{P}(N^m_{r,\ell^m(t)} \leq 1) = 1, \quad r = 1, \ldots, m. \tag{46}$$

□

Lemma 3. *Under assumption 3 and for any $t \in \mathbb{T}$, the number $N^m_{\ell^m(t)}$ of observations falling in $\mathcal{T}^m_{\ell^m(t)}$ diverges in probability to $+\infty$:*

$$\forall M > 0, \quad \lim_{m\to\infty} \mathbb{P}\left(N^m_{\ell^m(t)} > M\right) = 1. \tag{47}$$

Proof. Let $M > 0$. As in the proof of lemma 2, we have $N^m_{\ell^m(t)} \sim B(mn, P^m_{\ell^m(t)})$ and hence,

$$\mathbb{P}\left(N^m_{\ell^m(t)} \leq M\right) = \sum_{k=1}^{\lfloor M \rfloor} \binom{nm}{k} (P^m_{\ell^m(t)})^k (1 - P^m_{\ell^m(t)})^{nm-k}. \tag{48}$$

As the sum has is finite when we fix M, the limit of expression (48) when m tends to infinity is the sum of the limits of its terms.

$$\begin{aligned}&\binom{nm}{k} (P^m_{\ell^m(t)})^k (1 - P^m_{\ell^m(t)})^{nm-k} \\ &\qquad \leq \binom{nm}{k} (\nu_+ b_m)^k (1 - \nu_- b_m)^{nm-k}, \\ &\qquad \underset{m\to\infty}{\sim} (\nu_+ nm b_m)^k (1 - \nu_- b_m)^{nm}, \\ &\qquad = (nm b_m)^k \exp\left[nm \log\left(1 - \nu_- b_m\right)\right], \\ &\qquad \underset{m\to\infty}{\sim} (nm b_m)^k \exp\left[-\nu_- nm b_m + o(m b_m)\right].\end{aligned} \tag{49}$$

Using (13) from assumption 3, we obtain that (49) tends to 0 which proves that

$$\lim_{m\to\infty} \mathbb{P}\left(N^m_{\ell^m(t)} > M\right) = 1 - \lim_{m\to\infty} \mathbb{P}\left(N^m_{\ell^m(t)} \leq M\right) = 1.$$

□

In what follows, for any $M > 0$ we denote C_M the following event:

$$C_M := \{N^m_{\ell^m(t)} > M\} \bigcap_{r=1}^{m} \{N^m_{r,\ell^m(t)} \leq 1\}. \tag{50}$$

Lemma 4. *For any $t \in \mathbb{T}$, if assumption 3 holds,*

$$\forall M > 0, \quad \lim_{m\to\infty} \mathbb{P}(C_M) = 1. \tag{51}$$

Proof. Let $M > 0$. We have by definition of the conditional probability

$$\mathbb{P}(C_M) = \mathbb{P}\left(N^m_{\ell^m(t)} \middle| N^m_{r,\ell^m(t)} \leq 1; r = 1, \ldots, m\right) \times \mathbb{P}\left(N^m_{r,\ell^m(t)} \leq 1; r = 1, \ldots, m\right). \tag{52}$$

As the variables $N^m_{r,\ell^m(t)}$ are independent

$$\mathbb{P}\left(N^m_{r,\ell^m(t)} \leq 1; r = 1, \ldots, m\right) = \prod_{r=1}^{m} \mathbb{P}\left(N^m_{r,\ell^m(t)} \leq 1\right) \tag{53}$$

which tends to 1 when m grows according to lemma 2.

Furthermore, as $N^m_{\ell^m(t)} = \sum_{r=1}^{m} N^m_{r,\ell^m(t)} \sim \mathcal{B}(nm, P^m_{\ell^m(t)})$, we know that the random variable $(N^m_{\ell^m(t)} | N^m_{r,\ell^m(t)} \leq 1; r = 1, \ldots, m)$ is also a binomial random variable where at least $m(n-1)$ Bernouilli have failed, leaving m trials. Hence, we know that $(N^m_{\ell^m(t)} | N^m_{r,\ell^m(t)} \leq 1; r = 1, \ldots, m) \sim \mathcal{B}(r$ As in the proof of lemma 3,

$$\mathbb{P}\left(N^m_{\ell^m(t)} \leq M \middle| N^m_{r,\ell^m(t)} \leq 1; r = 1, \ldots, m\right) = \sum_{k=1}^{\lfloor M \rfloor} \binom{m}{k} (P^m_{\ell^m(t)})^k (1 - P^m_{\ell^m(t)})^{m-k}, \tag{54}$$

whose limit is the sum of the limits of the sum terms. As in the proof of lemma 3, we have

$$\binom{m}{k} (P^m_{\ell^m(t)})^k (1 - P^m_{\ell^m(t)})^{nm-k} \simeq (mb_m)^k \exp\left[-\nu_- m b_m + o(mb_m)\right], \tag{55}$$

and hence, by using (13) from assumption 3,

$$\lim_{m\to\infty} \mathbb{P}\left(N^m_{\ell^m(t)} \leq M \middle| N^m_{r,\ell^m(t)} \leq 1; r = 1, \ldots, m\right) = 0. \tag{56}$$

Combining (52), (53) and (56) we obtain (51). □

Lemma 5. *For any $z \in E$ and $t \in \mathbb{T}$, if assumptions 3 and 4 hold for some function Θ used to compute $\hat{f}^m_{\ell^m(t)}$, then*

$$\forall \varepsilon > 0, \qquad \lim_{m\to\infty} \mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| < \varepsilon\right) = 1. \tag{57}$$

PROOF. The randomness of $\hat{f}^m_{\ell^m(t)}(z)$ comes from both the sample of observations drawn from the bin's density $f^m_{\ell^m(t)}$ and the random number of observations falling in the bin $N^m_{\ell^m(t)}$. Hence, the idea here is to separate these two sources of randomness by conditioning on one of them. As we would like to use the result from assumption 4, it makes sense to condition on $N^m_{\ell^m(t)}$ here.

Indeed, for any $m > 0$, $\hat{f}^m_{\ell^m(t)}(z) = \Theta[\mathcal{T}^m_{\ell^m(t)}](z)$, where $\mathcal{T}^m_{\ell^m(t)}$ is a sample of $N^m_{\ell^m(t)}$ observations drawn from $f^m_{\ell^m(t)}$. Hence, according to assumption 4, for any $\varepsilon > 0$, for any $\alpha_1 > 0$, there is some $N_{\varepsilon,\alpha_1,m}$ such that

$$\mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| < \varepsilon \middle| C_{N_{\varepsilon,\alpha_1,m}}\right) > 1 - \alpha_1. \tag{58}$$

As seen in lemma 2, the conditioning on $N^m_{r,\ell^m(t)} \leq 1$ comes from the fact that when $\mathcal{T}^m_{\ell^m(t)}$ contains at most one observation per curve, those observations are independent. Furthermore, for ε, α_1 fixed, let $M := \sup_{m'}\{N_{\varepsilon,\alpha_1,m'}\}$ (which exists according to assumption 4). In this case, when assumption 3 holds, we have from lemma 4 that for any $\alpha_2 > 0$,there is some $m_{\varepsilon,\alpha_1,\alpha_2}$ such that,

$$m \geq m_{\varepsilon,\alpha_1,\alpha_2} \quad \Rightarrow \quad \mathbb{P}\left(C_M\right) > 1 - \alpha_2. \tag{59}$$

Hence, as $N^m_{\ell^m(t)} > M \Rightarrow N^m_{\ell^m(t)} > N_{\varepsilon,\alpha_1,m}$, we have

$$m \geq m_{\varepsilon,\alpha_1,\alpha_2} \quad \Rightarrow \quad \mathbb{P}\left(C_{N_{\varepsilon,\alpha_1,m}}\right) > 1 - \alpha_2. \tag{60}$$

Moreover, we can also write the following useful inequality:

$$\begin{aligned}
&\mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| < \varepsilon\right) \\
&\quad > \mathbb{P}\left(\{|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| < \varepsilon\} \bigcap C_{N_{\varepsilon,\alpha_1,m}}\right) \\
&\quad = \mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| < \varepsilon \middle| C_{N_{\varepsilon,\alpha_1,m}}\right) \times \mathbb{P}\left(C_{N_{\varepsilon,\alpha_1,m}}\right).
\end{aligned} \tag{61}$$

By combining (58), (60) and (61) we get that for any $\varepsilon, \alpha_1, \alpha_2 > 0$, there is $m_{\varepsilon,\alpha_1,\alpha_2}$ such that

$$m \geq m_{\varepsilon,\alpha_1,\alpha_2} \Rightarrow \mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| < \varepsilon\right) > 1 - \alpha_3, \tag{62}$$

with $\alpha_3 = \alpha_1 + \alpha_2 - \alpha_1\alpha_2$. □

Using these lemmas we can finally prove theorem 1:

PROOF OF THEOREM 1. Let $(t, z) \in \mathbb{T} \times E$ and let $\varepsilon > 0$. According to lemma 1, under assumptions 2 and 3, there is $m_\varepsilon \in \mathbb{N}^*$ such that for any $m \geq m_\varepsilon$,

$$|f^m_{\ell^m(t)}(z) - f_t(z)| < \varepsilon \tag{63}$$

As

$$\begin{aligned}
|\hat{f}^m_{\ell^m(t)}(z) - f_t(z)| &\leq |\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| + |f^m_{\ell^m(t)}(z) - f_t(z)|, \\
&\leq |\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| + \varepsilon,
\end{aligned} \tag{64}$$

we have

$$\{|\hat{f}^m_{\ell^m(t)}(z) - f_t(z)| > 2\varepsilon\} \subset \{|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| > \varepsilon\},$$

and

$$\mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f_t(z)| > 2\varepsilon\right) \leq \mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| > \varepsilon\right). \tag{65}$$

According to lemma 5 we have that

$$\lim_{m\to\infty} \mathbb{P}\left(|\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z)| > \varepsilon\right) = 0, \tag{66}$$

which allow us to obtain (17) from (65). □

In the following section we prove an even stronger convergence (in the L^2 sense) for standard kernel density estimators.

B. L^2 consistency for kernel density estimators

We assume here that $\hat{f}^m_{\ell^m(t)}$ is a standard kernel density estimator of the form (20).

Remark 2. Such an estimator is only defined for $N^m_{\ell^m(t)} > 0$. This is why, in the remaining section, we consider the conditioned random variables $(N^m_{\ell^m(t)}|N^m_{\ell^m(t)} > 0)$ and $(\hat{f}^m_{\ell^m(t)}|N^m_{\ell^m(t)} > 0)$, which are still denoted $N^m_{\ell^m(t)}$ and $\hat{f}^m_{\ell^m(t)}$ for simplicity. Furthermore,

$$\mathbb{P}\left(\cdot|N^m_{\ell^m(t)} > 0\right) = \frac{\mathbb{P}\left(\cdot \bigcap \{N^m_{\ell^m(t)} > 0\}\right)}{\mathbb{P}\left(N^m_{\ell^m(t)} > 0\right)}, \tag{67}$$

and, according to lemma 3, $\lim_{m\to\infty} \mathbb{P}\left(N^m_{\ell^m(t)} > 0\right) = 1$.

In the following, we use the following notations for the conditional expectation and variance:

$$\tilde{\mathbb{E}}[\cdot] := \mathbb{E}\left[\cdot|N^m_{\ell^m(t)} > 0\right], \qquad \tilde{\mathrm{Var}}[\cdot] := \mathrm{Var}\left[\cdot|N^m_{\ell^m(t)} > 0\right]. \tag{68}$$

In the remaining of this section we derive the conditions under which $\hat{f}^m_{\ell^m(t)}$ converges in expected squared-error to f_t. We recall that a sufficient condition for this is having its bias and variance tending to 0. The proof was greatly inspired by the derivations presented in Scott (2015) for the standard multivariate case.

Similarly to lemma 1, we can prove the following convergence result:

Lemma 6. *Under assumption 5, for any* $(t, z) \in \mathbb{T} \times E$,

$$\frac{d^2 f^m_{\ell^m(t)}}{dz^2}(z) := \left(f^m_{\ell^m(t)}\right)''(z) = \frac{\int_{\tau^m_{\ell-1}}^{\tau^m_\ell} f''_t(z)\nu(t)dt}{\int_{\tau^m_{\ell-1}}^{\tau^m_\ell} \nu(t)dt}, \tag{69}$$

and

$$\lim_{m\to\infty} \left|\left(f^m_{\ell^m(t)}\right)''(z) - f''_t(z)\right| = 0. \tag{70}$$

PROOF. Similar argument to lemma 1. □

Furthermore, the following bounds of the estimator's bias and variance hold under the same assumption:

Lemma 7. *If assumption 5 holds,*

$$\left|\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] - f^m_{\ell^m(t)}(z)\right| \leq \frac{1}{2}\left\|\left(f^m_{\ell^m(t)}\right)''\right\|_\infty \tilde{\mathbb{E}}\left[\sigma^2_{K_\sigma}\right], \tag{71}$$

$$\begin{aligned}\tilde{\mathrm{Var}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] \leq\ & f^m_{\ell^m(t)}(z)\tilde{\mathbb{E}}\left[\frac{R(K_\sigma)}{N^m_{\ell^m(t)}}\right] + \\ & \frac{1}{2}\left\|\left(f^m_{\ell^m(t)}\right)''\right\|_\infty \tilde{\mathbb{E}}\left[\frac{\sigma^2_{K^2_\sigma}}{N^m_{\ell^m(t)}}\right] + \frac{1}{4}\left\|\left(f^m_{\ell^m(t)}\right)''\right\|^2_\infty \tilde{\mathbb{E}}\left[(\sigma^2_{K_\sigma})^2\right].\end{aligned} \tag{72}$$

PROOF. From the law of total expectation (e.g. Wasserman, 2004, p.55)

$$\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] = \tilde{\mathbb{E}}\left[\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right]. \tag{73}$$

For large enough m, $\hat{f}^m_{\ell^m(t)}$ writes as an empirical average of independent identically distributed random variables $Z_{m,t}$ of density $f^m_{\ell^m(t)}$ and hence

$$\begin{aligned}\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right] &= \tilde{\mathbb{E}}\left[K_\sigma(z - Z_{m,t})|N^m_{\ell^m(t)}\right] = \tilde{\mathbb{E}}\left[K_\sigma(z - Z_{m,t})\right], \\ &= \int K_\sigma(z-x)f^m_{\ell^m(t)}(x)dx = \int K_\sigma(w)f^m_{\ell^m(t)}(z-w)dw.\end{aligned} \tag{74}$$

The second order integral Taylor expansion of $f^m_{\ell^m(t)}$ around z gives

$$f^m_{\ell^m(t)}(z-w) = f^m_{\ell^m(t)}(z) - \left(f^m_{\ell^m(t)}\right)'(z)w + \int_0^1 (1-x)\left(f^m_{\ell^m(t)}\right)''(z-xw)w^2dx, \tag{75}$$

which leads to

$$\begin{aligned}\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right] = f^m_{\ell^m(t)}(z)\int K_\sigma(w)dw - \left(f^m_{\ell^m(t)}\right)'(z)\int wK_\sigma(w)dw \\ + \int\int_0^1 (1-x)\left(f^m_{\ell^m(t)}\right)''(z-xw)w^2K_\sigma(w)dx\ dw.\end{aligned} \tag{76}$$

As K_σ is a symmetric probability density function, we have

$$\int K_\sigma(w)dw = 1, \qquad \int wK_\sigma(w)dw = 0, \tag{77}$$

which leads to

$$\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right] = f^m_{\ell^m(t)}(z) + \int\int_0^1 (1-x)\left(f^m_{\ell^m(t)}\right)''(z-xw)w^2K_\sigma(w)dx\ dw. \tag{78}$$

By combining (73) and (78) we obtain the following bias expression

$$\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] - f^m_{\ell^m(t)}(z) = \tilde{\mathbb{E}}\left[\int\int_0^1 (1-x)\left(f^m_{\ell^m(t)}\right)''(z-xw)w^2K_\sigma(w)dx\ dw\right], \quad (79)$$

proving bound (71).

Concerning the variance, we apply the law of total variance (e.g. Wasserman, 2004, p.55):

$$\tilde{\mathrm{Var}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] = \tilde{\mathbb{E}}\left[\tilde{\mathrm{Var}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right] + \tilde{\mathrm{Var}}\left[\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right]. \quad (80)$$

As in (74), we may express the conditional variance of $\hat{f}^m_{\ell^m(t)}(z)$ using the conditional variance of the kernel, which brings the first term in (80) to

$$\tilde{\mathbb{E}}\left[\tilde{\mathrm{Var}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right] = \tilde{\mathbb{E}}\left[\frac{1}{N^m_{\ell^m(t)}}\tilde{\mathrm{Var}}\left[K_\sigma(z-Z_{m,t})|N^m_{\ell^m(t)}\right]\right]. \quad (81)$$

Furthermore, the kernel variance can be developped as follows

$$\tilde{\mathrm{Var}}\left[K_\sigma(z-Z_{m,t})|N^m_{\ell^m(t)}\right] = \tilde{\mathbb{E}}\left[K_\sigma(z-Z_{m,t})^2|N^m_{\ell^m(t)}\right] - \tilde{\mathbb{E}}\left[K_\sigma(z-Z_{m,t})|N^m_{\ell^m(t)}\right]^2, \quad (82)$$

which is hence smaller than the first term, i.e. the kernels second moment. Using integral Taylor expansion of $f^m_{\ell^m(t)}$ around z (75) truncated to the first order, such a quantity can be written as follows:

$$\begin{aligned}
\tilde{\mathbb{E}}\left[K_\sigma(z-Z_{m,t})^2|N^m_{\ell^m(t)}\right] &= \int K_\sigma(z-\lambda)^2 f^m_{\ell^m(t)}(\lambda)d\lambda \\
&= \int K_\sigma(w)^2 f^m_{\ell^m(t)}(z-w)dw \\
&= f^m_{\ell^m(t)}(z)R(K_\sigma) - \left(f^m_{\ell^m(t)}\right)'(z)\underbrace{\int wK_\sigma(w)^2dw}_{=0} \\
&\quad + \int\int_0^1 (1-x)w^2\left(f^m_{\ell^m(t)}\right)''(z-xw)K_\sigma(w)^2dx\ dw,
\end{aligned} \quad (83)$$

where we used the fact that K_σ^2 is an even function. By using (81)-(82) and (83) we get

$$\tilde{\mathbb{E}}\left[\tilde{\mathrm{Var}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right] \leq f^m_{\ell^m(t)}(z)\tilde{\mathbb{E}}\left[\frac{R(K_\sigma)}{N^m_{\ell^m(t)}}\right] + \frac{1}{2}\left\|\left(f^m_{\ell^m(t)}\right)''\right\|_\infty \tilde{\mathbb{E}}\left[\frac{\sigma^2_{K^2_\sigma}}{N^m_{\ell^m(t)}}\right]. \quad (84)$$

We still need to bound the second term in (80). We have

$$\begin{aligned}
&\tilde{\mathrm{Var}}\left[\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right] \\
&\quad = \mathbb{E}\left[\left(\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right] - \underbrace{\tilde{\mathbb{E}}\left[\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right]}_{\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right]}\right)^2\right].
\end{aligned} \quad (85)$$

By denoting

$$G_m = \int \int_0^1 (1-x) \left(f^m_{\ell^m(t)}\right)'' (z - xw) w^2 K_\sigma(w) dx \; dw \tag{86}$$

and plugging (78)-(79) in (85), we obtain

$$\tilde{\text{Var}}\left[\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right] = \tilde{\mathbb{E}}\left[\left(G_m - \tilde{\mathbb{E}}\left[G_m\right]\right)^2\right] \leq \tilde{\mathbb{E}}\left[G_m^2\right]. \tag{87}$$

Furthermore,

$$\begin{aligned} \tilde{\mathbb{E}}\left[G_m^2\right] &= \tilde{\mathbb{E}}\left[\left(\int \int_0^1 (1-x) \left(f^m_{\ell^m(t)}\right)'' (z - xw) w^2 K_\sigma(w) dx \; dw\right)^2\right], \\ &\leq \frac{1}{4} \left\| \left(f^m_{\ell^m(t)}\right)'' \right\|^2_\infty \tilde{\mathbb{E}}\left[\left(\int w^2 K_\sigma(w) dw\right)^2\right], \end{aligned} \tag{88}$$

which leads to the following bound of the second term in (80)

$$\tilde{\text{Var}}\left[\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)|N^m_{\ell^m(t)}\right]\right] \leq \frac{1}{4} \left\| \left(f^m_{\ell^m(t)}\right)'' \right\|^2_\infty \tilde{\mathbb{E}}\left[(\sigma^2_{K_\sigma})^2\right], \tag{89}$$

and proves inequality (72). □

As a direct consequence, we get the following bounds for the simpler case where kernel and bandwidth are deterministic:

Lemma 8. *If $\sigma = \sigma_m$ depends on m and K is fixed, both deterministic (do not depend on the sample), then under assumption 5:*

$$\left|\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] - f^m_{\ell^m(t)}(z)\right| \leq \frac{1}{2} \left\| \left(f^m_{\ell^m(t)}\right)'' \right\|_\infty \sigma^2_K \sigma^2_m, \tag{90}$$

$$\begin{aligned} \tilde{\text{Var}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] \leq \tilde{\mathbb{E}}\left[\frac{1}{N^m_{\ell^m(t)}}\right] &\left(\frac{f^m_{\ell^m(t)}(z) R(K)}{\sigma_m} + \left\| \left(f^m_{\ell^m(t)}\right)'' \right\|_\infty \frac{\sigma^2_{K^2} \sigma_m}{2}\right) \\ &+ \frac{1}{4} \left\| \left(f^m_{\ell^m(t)}\right)'' \right\|^2_\infty (\sigma^2_K \sigma^2_m)^2. \end{aligned} \tag{91}$$

Proof. Obtained directly by using (21)-(23) together with lemma 7. □

We know from lemmas 1 and 6 that, under assumptions 3 and 5:

$$\lim_{m \to \infty} f^m_{\ell^m(t)}(z) = f_t(z) < \infty, \tag{92}$$

$$\lim_{m \to \infty} \left(f^m_{\ell^m(t)}\right)'' (z) = f''_t(z) < \infty. \tag{93}$$

Hence, by looking at expressions (90) and (91), it seems clear that the convergence of the bias and the variance to zero will strongly depend on the asymptotics of $\tilde{\mathbb{E}}\left[1/N^m_{\ell^m(t)}\right]$. This motivates the following lemma:

Lemma 9. *If assumption 1 to 3 hold, as* $m \to \infty$

$$\tilde{\mathbb{E}}\left[\frac{1}{N^m_{\ell^m(t)}}\right] = O\left(\frac{1}{nmP^m_{\ell^m(t)}}\right) = O\left(\frac{1}{mb_m}\right). \tag{94}$$

Proof. We recall that $N^m_{\ell^m(t)} \sim \mathcal{B}(nm, P^m_{\ell^m(t)})$, with $P^m_{\ell^m(t)} = \int_{B^m_{\ell^m(t)}} \nu(t)dt$. According to the main theorem proved in Cribari-Neto et al. (2000),

$$\begin{aligned} S_1 &:= \mathbb{E}\left[\frac{1}{1+N^m_{\ell^m(t)}}\right] = \sum_{k=0}^{nm} \binom{nm}{k} \frac{(P^m_{\ell^m(t)})^k (1-P^m_{\ell^m(t)})^{nm-k}}{1+k} \\ &= O\left(\frac{1}{nmP^m_{\ell^m(t)}}\right). \end{aligned} \tag{95}$$

As noted in remark 2, $\hat{f}^m_{\ell^m(t)}$ is not defined for $N^m_{\ell^m(t)} = 0$. This motivates the conditioning of the variable $\frac{1}{N^m_{\ell^m(t)}}$ by the event $N^m_{\ell^m(t)} > 0$ and we have

$$S_0 := \mathbb{E}\left[\frac{1}{N^m_{\ell^m(t)}} \middle| N^m_{\ell^m(t)} > 0\right] = \sum_{k=1}^{nm} \binom{nm}{k} \frac{(P^m_{\ell^m(t)})^k (1-P^m_{\ell^m(t)})^{nm-k}}{k}. \tag{96}$$

By computing the difference between the S_1 and S_0 we obtain the following bound:

$$\begin{aligned} S_0 - S_1 &= -\binom{nm}{0}(1-P^m_{\ell^m(t)})^{nm} + \sum_{k=1}^{nm} \binom{nm}{k} \frac{(P^m_{\ell^m(t)})^k (1-P^m_{\ell^m(t)})^{nm-k}}{k(k+1)} \\ &\leq \sum_{k=1}^{nm} \binom{nm}{k} \frac{(P^m_{\ell^m(t)})^k (1-P^m_{\ell^m(t)})^{nm-k}}{k+1} \leq S_1. \end{aligned} \tag{97}$$

We conclude that $S_0 \leq 2S_1$ and hence

$$\mathbb{E}\left[\frac{1}{N^m_{\ell^m(t)}} \middle| N^m_{\ell^m(t)} > 0\right] = O\left(\frac{1}{nmP^m_{\ell^m(t)}}\right). \tag{98}$$

Finally, the last equality in (94) comes from the fact that $P^m_{\ell^m(t)} \geq \nu_- b_m$. □

In conclusion, we can now prove theorem 2 stating that the kernel marginal density estimator will be consistent in expected squared-error (which implies convergence in probability stated in theorem (1)).

Proof of Theorem 2. By (90) and (26), the bias of $\hat{f}^m_{\ell^m(t)}(z)$ converges to 0:

$$\lim_{m\to\infty} \left|\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] - f^m_{\ell^m(t)}(z)\right| = 0. \tag{99}$$

Similarly, as σ_m converges to 0, the two last terms in (91) shrink. Concerning the first term, we can conclude from lemma 9 and from condition (26) that

$$\lim_{m\to\infty} \tilde{\mathbb{E}}\left[\frac{1}{N^m_{\ell^m(t)}}\right] \frac{1}{\sigma_m} = 0, \tag{100}$$

which means that the variance of $\hat{f}^m_{\ell^m(t)}(z)$ also converges to 0:

$$\lim_{m\to\infty} \tilde{\mathrm{Var}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] = 0. \tag{101}$$

From the bias-variance decomposition of the expected squared-error between $\hat{f}^m_{\ell^m(t)}(z)$ and $f^m_{\ell^m(t)}(z)$ we have

$$\begin{aligned}\tilde{\mathbb{E}}\left[(\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z))^2\right] &= \tilde{\mathbb{E}}\left[\left(\hat{f}^m_{\ell^m(t)}(z) - \tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right]\right.\right.\\ &\qquad \left.\left.+\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] - f^m_{\ell^m(t)}(z)\right)^2\right],\\ &= \tilde{\mathbb{E}}\left[\left(\hat{f}^m_{\ell^m(t)}(z) - \tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right]\right)^2\right]\\ &\qquad + \tilde{\mathbb{E}}\left[\left(\tilde{\mathbb{E}}\left[\hat{f}^m_{\ell^m(t)}(z)\right] - f^m_{\ell^m(t)}(z)\right)^2\right].\end{aligned}$$

Furthermore, by Jensen inequality we have that,

$$\frac{1}{2}\tilde{\mathbb{E}}\left[(\hat{f}^m_{\ell^m(t)}(z) - f_t(z))^2\right] \leq \tilde{\mathbb{E}}\left[(\hat{f}^m_{\ell^m(t)}(z) - f^m_{\ell^m(t)}(z))^2\right] + (f^m_{\ell^m(t)}(z) - f_t(z))^2. \tag{102}$$

Finally, by using lemma 1 in conjunction with (99), (101) and (102), we obtain that both terms in (102) tend to 0, leading to result (27). □

C. Derivation of the Self-Consistent Density Estimator

In this section we present in more details the self-consistent estimator proposed by Bernacchia and Pigolotti (2011) and extended by O'Brien et al. (2014); O'Brien et al. (2016).

C.0.1. Optimal Kernel Density Estimator

Let $S^N = \{z_k\}_{k=1}^N \subset E$ be a sample of N observations drawn from a common density function f. We suppose that $f \in L^2(E, \mathbb{R}_+)$ and that $N > 0$ is deterministic. We consider a kernel estimator of f

$$\hat{f}(z) := \frac{1}{N}\sum_{k=1}^{N} K(z - z_k), \qquad \forall z \in E, \tag{103}$$

where $K : \mathbb{R}^d \to \mathbb{R}$ is a smoothing kernel in $L^2(\mathbb{R}^d, \mathbb{R})$. One can interpret (103) as a kernel density estimator with an implicit bandwidth $H \in GL_d(\mathbb{R}_+)$ hidden inside the expression of the kernel function $K(z) = \tilde{K}(H^{-1}z)/\det H$, $\forall z \in \mathbb{R}^d$.

Denote by $\mathbb{E}$ the expectation relative to the random sample S^N, defined for any deterministic function $\varphi : E^N \to \mathbb{R}$ by

$$\mathbb{E}\left[\psi(S^N)\right] := \int_E \cdots \int_E \varphi(z_1, \ldots, z_N) f(z_1) \ldots f(z_N) dz_1 \ldots dz_N. \tag{104}$$

In this context, a common quality measure of density estimators is the Mean Integrated Squared Error:

$$MISE := \mathbb{E}\left[\int_E (\hat{f}(z) - f(z))^2 dz\right]. \tag{105}$$

As we will show, it becomes relatively easy to minimize such a criterion with regard to the choice of the kernel K once we've shifted it to the Fourier domain. Hence, for any function $v \in L^2(\mathbb{R}^d, \mathbb{R})$, we define its Fourier transform hereafter with the following convention

$$\mathcal{F}[v](s) := \int_{\mathbb{R}^d} v(z) e^{iz\cdot s} dz, \qquad \forall s \in \mathbb{R}^d, \tag{106}$$

where $i = \sqrt{-1}$, its inverse being defined by

$$\mathcal{F}^{-1}[v](z) := \frac{1}{2\pi} \int_{\mathbb{R}^d} v(s) e^{-iz\cdot s} ds, \qquad \forall z \in \mathbb{R}^d. \tag{107}$$

As $f, \hat{f} \in L^2$, Plancherel's theorem gives that

$$MISE = \frac{1}{2\pi}\,\mathbb{E}\left[\int_{\mathbb{R}^d} |\hat{\Phi}(s) - \Phi(s)|^2 ds\right], \tag{108}$$

where $\Phi := \mathcal{F}[f]$, usually called the *characteristic function*, and $\hat{\Phi} := \mathcal{F}[\hat{f}]$. By noticing that $\hat{f}$ can be seen as the convolution between the kernel and a sum of Dirac functions centered on the data points

$$\hat{f}(z) = \left(K * \left(\frac{1}{N}\sum_{k=1}^{N} \delta_{z_k}\right)\right)(z), \tag{109}$$

it follows that

$$\hat{\Phi}(s) = \kappa(s)\Delta(s), \tag{110}$$

where

$$\kappa(s) := \mathcal{F}[K](s) \tag{111}$$

$$\Delta(s) := \mathcal{F}\left[\frac{1}{N}\sum_{k=1}^{N} \delta_{z_k}\right](s) = \frac{1}{N}\sum_{k=1}^{N} e^{iz_k\cdot s} \in \mathbb{C}. \tag{112}$$

The function Δ is commonly called the *empirical characteristic function* (ECF).

Plugging (110) in the MISE expression (108) and expanding the square gives:

$$\begin{aligned} MISE = \frac{1}{2\pi}\int_{\mathbb{R}^d} \Big[& |\kappa|^2\,\mathbb{E}\left[|\Delta|^2\right] + |\Phi|^2 \\ & -\kappa(\mathbb{E}\left[\Delta\right]\Phi^* + \mathbb{E}\left[\Delta^*\right]\Phi)\Big] ds, \end{aligned} \tag{113}$$

the arguments s being omitted for lighter notation and c^* denoting the complex conjugate of any $c \in \mathbb{C}$. Furthermore, as shown in (Tsybakov, 2008, section 1.3, lemma 1.2) for

example, the ECF is an unbiased estimator of the characteristic function

$$\begin{aligned}\mathbb{E}\left[\Delta(s)\right] &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}\left[e^{is\cdot z_k}\right] \\ &= \frac{1}{N}\sum_{k=1}^{N}\int_{-\infty}^{+\infty} e^{is\cdot z_k} f(z_k)dz_k \\ &= \Phi(s),\end{aligned} \tag{114}$$

and its second moment is

$$\mathbb{E}\left[|\Delta(s)|^2\right] = \mathbb{E}\left[\Delta(s)\Delta(s)^*\right] = \mathbb{E}\left[\Delta(s)\Delta(-s)\right] \tag{115}$$

$$= \mathbb{E}\left[\frac{1}{N^2}\sum_{j,k:j\neq k} e^{is\cdot(z_j - z_k)}\right] + \frac{1}{N} \tag{116}$$

$$= \frac{1}{N^2}\sum_{j,k:j\neq k}\mathbb{E}\left[e^{is\cdot z_j}e^{-is\cdot z_k}\right] + \frac{1}{N} \tag{117}$$

$$= \frac{1}{N^2}\sum_{j,k:j\neq k}\mathbb{E}\left[e^{is\cdot z_j}\right]\mathbb{E}\left[e^{-is\cdot z_k}\right] + \frac{1}{N} \tag{118}$$

$$= \frac{1}{N^2}\sum_{j,k:j\neq k}\Phi(s)\Phi(-s) + \frac{1}{N} \tag{119}$$

$$= \frac{N-1}{N}\Phi(s)\Phi(-s) + \frac{1}{N} \tag{120}$$

$$= \frac{N-1}{N}|\Phi(s)|^2 + \frac{1}{N}. \tag{121}$$

Passing from line (117) to (118) is based on the assumption that the random variables $\{z_k\}_{k=1}^N$ are independent.

Replacing (114) and (121) in (113), we obtain

$$MISE = \frac{1}{2\pi}\int_{\mathbb{R}^d}\frac{|\kappa|^2}{N}(1-|\Phi|^2) + |\Phi|^2(1-\kappa)^2 ds. \tag{122}$$

As initially shown in Watson and Leadbetter (1963), expression (122) can be minimized with respect to the transformed kernel κ. Indeed, for any $s \in \mathbb{R}^d$, we get the following first order optimality condition by differentiating the quadratic integrand in (122) relative to $\kappa(s)$:

$$\frac{1}{N}\kappa(1-|\Phi|^2) - |\Phi|^2(1-\kappa) = 0, \tag{123}$$

leading to the *optimal transformed kernel*

$$\kappa_{opt}(s) := \frac{N}{N-1+|\Phi(s)|^{-2}}, \qquad \forall s \in \mathbb{R}^d. \tag{124}$$

Hence the optimal density estimator relative to the MISE is given by

$$\boxed{\hat{f}_{opt}(z) = \mathcal{F}^{-1}[\hat{\Phi}_{opt}](z),} \tag{125}$$

where

$$\boxed{\hat{\Phi}_{opt}(s) = \kappa_{opt}(s)\Delta(s) = \frac{N\Delta(s)}{N - 1 + |\Phi(s)|^{-2}}.} \tag{126}$$

C.0.2. Self-Consistent Estimator

The practical problem with estimator (126) is that it depends on the true characteristic function Φ, which is unknown. Hence, the solutions to the fixed-point equation (127) was suggested by Bernacchia and Pigolotti (2011) to approximate $\hat{\Phi}_{opt}$:

$$\hat{\Phi}^N_{sc} = \frac{N\Delta}{N - 1 + |\hat{\Phi}^N_{sc}|^{-2}}. \tag{127}$$

This is justified by the fact that the optimal estimator $\hat{\Phi}_{opt}$ should be very close to the true characteristic function Φ, as illustrated by the MISE criterion (108).

Equation (127) can be transformed into a second order equation in $|\hat{\Phi}^N_{sc}|$,

$$(N-1)|\hat{\Phi}^N_{sc}|^2 - N|\Delta||\hat{\Phi}^N_{sc}| + 1 = 0, \tag{128}$$

which admits a solution in $\mathbb{R}_+$ provided that

$$|\Delta(s)|^2 \geq (\Delta^{min}_N)^2 := \frac{4(N-1)}{N^2}. \tag{129}$$

When inequality (129) holds, the two possible solutions of (128) are

$$|\hat{\Phi}^+| := \frac{N|\Delta|}{2(N-1)}\left(1 + \sqrt{1 - \frac{(\Delta^{min}_N)^2}{|\Delta|^2}}\right), \tag{130}$$

$$|\hat{\Phi}^-| := \frac{N|\Delta|}{2(N-1)}\left(1 - \sqrt{1 - \frac{(\Delta^{min}_N)^2}{|\Delta|^2}}\right). \tag{131}$$

After some analysis, we can show that $\hat{\Phi}^+$ is a stable fixed-point, while $\hat{\Phi}^-$ is unstable. Bernacchia and Pigolotti (2011) suggest to keep only the stable one, which brings us to the self-consistent estimator of the characteristic function:

$$\boxed{\hat{\Phi}^N_{sc}(s) := \frac{N\Delta(s)}{2(N-1)}\left(1 + \sqrt{1 - \frac{(\Delta^{min}_N)^2}{|\Delta(s)|^2}}\right)\mathbf{1}_{A_N}(s),} \tag{132}$$

where $\mathbf{1}_{A_N}$ denotes the indicator function over an arbitrary subset $A_N \subset \mathbb{S}_N$ of the frequencies in

$$\mathbb{S}_N := \left\{s : |\Delta(s)|^2 \geq (\Delta^{min}_N)^2\right\}. \tag{133}$$

Hence our new density estimator writes

$$\boxed{\hat{f}^N_{sc}(z) := \mathcal{F}^{-1}\left[\hat{\Phi}^N_{sc}\right](z), \qquad \forall z \in E.} \tag{134}$$

C.0.3. Practical Considerations

Heuristics for choosing A_N were proposed in Bernacchia and Pigolotti (2011); O'Brien et al. (2014) for the univariate case and in O'Brien et al. (2016) in a multivariate setting.

One practical problem with the self-consistent estimator is that $\hat{f}_{sc}^N = \mathcal{F}^{-1}[\hat{\Phi}_{sc}^N]$ is not lower-bounded by zero. This can be corrected by translating $\hat{f}_{sc}^N$ downwards until the positive part integrates to one and then setting the negative part to 0. Indeed, it was proven by Glad et al. (2003) that such a transformation induces no cost in terms of MISE accuracy.

Another practical drawback with estimator $\hat{\Phi}_{sc}^N$ is that the direct computation of the empirical characteristic function Δ can be expensive: $O(N \cdot M)$ exponential evaluations, where M is the number of frequency points $s \in \mathbb{R}^d$. Noting from definition (112) that the expression of Δ is equivalent to some Discrete Fourier Transform

$$\Delta(s) = \frac{1}{N} \sum_{k=1}^{N} a_k e^{is \cdot z_k}, \tag{135}$$

where the Fourier coefficients a_k are all equal to 1, the idea of using the Fast Fourier Transform algorithm (FFT) proposed by Cooley and Tukey (1965) seems natural. However, the latter only applies to the case of uniformly spaced data, which is not the case of $\{z_k\}_{k=1}^N$. For this reason, O'Brien et al. (2014) proposed to use an implementation of Nonuniform Fast Fourier Transform (NUFFT) developped by Greengard and Lee (2004).

It consists in interpolating the original data points $\{z_k\}_{k=1}^N$ on a new equispaced grid $\{\tilde{z}_j\}_{j=1}^{\tilde{N}_\ell}$ by using another Gaussian kernel density estimator:

$$\begin{aligned} \tilde{f}(\tilde{z}_j) &:= \frac{1}{N} \sum_{k=1}^{N} K_G(z_k - \tilde{z}_j), \\ &= K_G * \left(\frac{1}{N} \sum_{k=1}^{N} \delta_{z_k} \right) (\tilde{z}_j), \end{aligned} \tag{136}$$

with $K_G(z) := \exp\left(-\frac{z^2}{\sigma^2}\right)$, $\forall z \in \mathbb{R}^d$, and $\sigma \in \mathbb{R}_+^*$. The FFT can than be used to approximate $\tilde{\Phi}(s) := \mathcal{F}[\tilde{f}](s)$, and by dividing it by the transformed Gaussian kernel $\kappa_G(s) := \mathcal{F}[K_G](s)$, we obtain the ECF evaluation:

$$\Delta(s) = \tilde{\Phi}(s) \cdot [\kappa_G(s)]^{-1}. \tag{137}$$

As computing $\{\tilde{f}(\tilde{z}_j)\}_{j=1}^{\tilde{N}_\ell}$ still takes $O(\tilde{N}_\ell \cdot N)$ operations, Dutt and Rokhlin (1993) suggested to use only $N^c < N$ surrounding points from $\{z_k\}_{k=1}^N$ to evaluate each new grid point $\tilde{z}_j$. We obtain an overall complexity of $O(N^c \cdot \tilde{N}_\ell + M \log M)$ which, in the case where $N^c < M \leq N$, is better than the original DFT formulation $O(N \cdot M)$. The analysis conducted in Greengard and Lee (2004) indicates that simple precision can be achieved in (137) by normalizing the data $\{z_k\}_{k=1}^N$ to the range $[0, 2\pi]$ and setting $2N^c = 12$, $\tilde{N}_\ell = 2N_\ell$ and $\sigma^2 = 24/N^2$. Hence, for a desired precision, this step of the algorithm introduces no additional hyperparameter to be tuned (see Greengard and Lee (2004) for the double-precision settings).