Electronic Journal of Statistics
Vol. 10 (2016) 2922–2972
ISSN: 1935-7524
DOI: 10.1214/16-EJS1198

Nonparametric clustering of functional data using pseudo-densities

Mattia Ciollaro, Christopher R. Genovese* **and Daren Wang**

Department of Statistics
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
e-mail: ciollaro@cmu.edu; genovese@stat.cmu.edu; darenw@andrew.cmu.edu

Abstract: We study nonparametric clustering of smooth random curves on the basis of the L^2 gradient flow associated to a pseudo-density functional and we discuss the conditions under which the clustering is well-defined both at the population and at the sample level. We provide an algorithm to idenify significant local modes of the estimated pseudo-density, which are associated to informative sample clusters, and we prove its consistency and other statistical properties. Our theory is developed under weak assumptions, which essentially reduce to the integrability of the random curves. If the underlying probability distribution is supported on a finite-dimensional subspace, we show that the proposed pseudo-density functional and the expectation of a kernel density estimator induce the same gradient flow, hence the same population clustering. Although our theory is developed for smooth curves that belong to a potentially infinite-dimensional functional space, we provide consistent procedures that can be used with real functional data (discretized and noisy curves). We illustrate these procedures by means of applications both on simulated and real datasets.

MSC 2010 subject classifications: Primary 62G07, 62G86; Secondary 62G99.
Keywords and phrases: Modal clustering, pseudo-density, gradient flow, functional data analysis.

Received January 2016.

1. Introduction

In Functional Data Analysis (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Ferraty and Romain, 2011; Horváth and Kokoszka, 2012; Zhang, 2013; Hsing and Eubank, 2015), henceforth FDA, we think of curves (and other functions) as the fundamental unit of measurement. Clustering is an important problem in FDA because it is often of critical interest to identify subpopulations based on the shapes of the measured curves. In this paper, we study the problem of functional clustering in a fully infinite-dimensional setting. We are motivated by recent work on modal clustering in finite dimensions (Chacón, 2015, and references therein) that, in contrast to many commonly-used clustering methods,

*Research supported in part by National Science Foundation Grants NSF-DMS-1208354 and NFS-DMS-1513412.

has a population formulation, and by recent advances in clustering of functional data (Bongiorno and Goia, 2015). Specifically, we prove the existence of population clusters in the infinite-dimensional functional case, under mild conditions. We show that an analogue of the mean-shift algorithm (see, for example, Fukunaga and Hostetler, 1975; Cheng, 1995, and the more recent works of Comaniciu, Ramesh and Meer, 2001; Carreira-Perpiñán, 2006) can identify local modes of a "pseudo-density". We devise an algorithm to classify local modes as representatives of significant clusters, and under some regularity assumptions on the pseudo-density, we further show that the algorithm is consistent. We develop our theory assuming that the data are observed as continuous curves defined on some interval. Because in practice one does not observe continuous curves, we also show how to apply the procedures that we propose to real data (e.g. noisy measurements of random curves on a grid).

Modal clustering is typically a finite-dimensional problem, but motivated by the flourishing literature on FDA and by the increasing interest in developing sound frameworks and algorithms for clustering of random curves, we extend the idea of modal clustering to the case where X is a functional random variable valued in an infinite-dimensional space. In particular, we develop a theory of modal clustering for smooth random curves that are assumed to belong to the Hölder space $H^1([0,1])$ of curves defined on the standard unit interval whose first weak derivative is square integrable. We focus on $H^1([0,1])$ for concreteness, but our theory generalizes to any more general spaces. Furthermore, our theory is density-free and nonparametric, as no assumptions are made regarding the existence of a dominating measure for the law P of the functional data, nor it is assumed that P can be parametrized by a finite number of parameters.

In the finite-dimensional modal clustering problem, we have that $p : \mathcal{X} \to \mathbb{R}_+$ is the probability density function associated to the law P of a random variable X valued in $\mathcal{X} \subseteq \mathbb{R}^d$. If p is a Morse function (i.e. p is smooth and its Hessian is not singular at the critical points), then the local modes of p, $\mu_1, \ldots, \mu_k$, induce an partition of the sample space $\mathcal{X} = C_1 \cup C_2 \cup \cdots \cup C_k$ where the sets C_i satisfy

1. $P(C_i) > 0 \ \forall i = 1, \ldots, k$
2. $P(C_i \cap C_j) = 0$ if $i \neq j$
3. $P(\cup_{i=1}^k C_i) = 1$
4. $x \in C_i \iff$ the gradient ascent path on p that starts from x eventually converges to μ_i.

Note that this framework characterizes C_i as a high-density region surrounding the local mode μ_i of p and each set $C_i \in \mathcal{C}$ is thought of as a cluster at the population level. Unlike other approaches to clustering which define clusters exclusively at the sample level (consider k-means, for instance), modal clustering provides an inferential framework in which the essential partition $\mathcal{C}$ is a population parameter that one wants to infer from the data. In fact, as soon as an i.i.d. sample $X_1, \ldots, X_n \sim P$ and an estimator $\hat{p}$ of p are available, the goals of modal clustering are exactly

- estimating the local modes of p by means of the local modes of $\hat{p}$
- estimating the population clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ by means of the empirical partition $\hat{\mathcal{C}} = \{\hat{C}_1, \ldots, \hat{C}_{\hat{k}}\}$ induced by $\hat{p}$

Thus, the typical output of a modal clustering procedure consists of the estimated clustering structure $\hat{\mathcal{C}}$ and a set of cluster representatives $\hat{\mu}_1, \ldots, \hat{\mu}_{\hat{k}}$. At the sample level, each data point is then uniquely assigned to a cluster $\hat{C}_i \in \hat{\mathcal{C}}$ and represented by the corresponding local mode $\hat{\mu}_i$.

Because it is generally not possible to define a probability density function in infinite-dimensional Hilbert spaces, we instead focus on a surrogate notion of density which we call "pseudo-density". Generally, by pseudo-density we mean any suitably smooth functional which maps the sample space $\mathcal{X}$ into the positive reals $\mathbb{R}_+ = [0, \infty)$. In particular, we focus on a family of pseudo-densities $\mathcal{P} = \{p_h : \mathcal{X} \to \mathbb{R}_+; h > 0\}$ which is parametrized by a bandwidth parameter h and, more specifically, p_h is the expected value of a kernel density estimator,

$$p_h(x) = E_P\, K\left(\frac{\|X - x\|_{L^2}^2}{h}\right), \tag{1.1}$$

where K is an appropriately chosen kernel function and h is the bandwidth parameter. Clusters of curves are then defined in terms of the L^2 gradient flow associated to p_h.

The gradient flow associated to $p_h \in \mathcal{P}$ is the collection of the gradient ascent paths $\pi_x : \mathbb{R}_+ \to \mathcal{X}$ corresponding to the solution of the initial value problem

$$\begin{cases} \frac{d}{dt}\pi_x(t) = \nabla p_h(\pi_x(t)) \\ \pi_x(0) = x, \end{cases} \tag{1.2}$$

where $\nabla p_h(x)$ is the L^2 functional gradient of p_h at $x \in \mathcal{X}$. In complete analogy with the finite-dimensional case, the gradient of p_h induces a vector field and a gradient ascent path is a path in $H^1([0,1])$, $\pi_x \subset H^1([0,1])$, that solves the initial value problem and flows along the direction of the vector field (at any time $t \geq 0$, $\pi_x(t)$ is an element of $H^1([0,1])$). The path π_x has the property that, at any time $t \geq 0$, the derivative of $\pi_x(t)$ corresponds to the gradient of p_h evaluated at $\pi_x(t)$. If the trajectory π_x converges to a local mode $\mu_i = \mu_i(h)$ of p_h as $t \to \infty$, then x is said to belong to the i-th cluster of p_h, $C_i = C_i(h)$. Thus, the cluster C_i is defined as the set

$$C_i = \left\{ x \in H^1([0,1]) : \lim_{t \to \infty} \|\pi_x(t) - \mu_i\|_{L^2([0,1])} \to 0 \right\}, \tag{1.3}$$

where π_x is a solution of the initial value problem of equation (1.2). According to the above definition, the i-th cluster of p_h corresponds to the basin of attraction of the i-th local mode μ_i of p_h, and the collection of the clusters C_i provides a summary of the subpopulations associated to the probability measure P.

The main contribution of our work is to identify conditions under which

1. there exist population clusters in functional data, i.e. the population clusters defined in equation (1.3) exist and are well-defined
2. these clusters are estimable.

Additionally, we describe a procedure to estimate the clusters and assess their statistical significance. This procedure can also be used as a means to perform bandwidth selection in practice.

As we further discuss later in the paper, the most remarkable challenge arising in the infinite-dimensional setting is the lack of compactness. As opposed to the finite-dimensional setting, in the functional case it is hard to show the existence, the uniqueness, and the convergence of the gradient ascent paths described by the initial value problem of equation (1.2), unless the sample space $\mathcal{X}$ can be compactly embedded in another space. We show that we can overcome this challenge by exploiting the compact embedding of $H^1([0,1])$ in $L^2([0,1])$, the space of square-integrable functions on the unit interval, and by studying equation (1.2) using these two non-equivalent topologies. For convenience, we focus on functional data belonging to $H^1([0,1])$ and on the gradient flow under the L^2 norm, but the exact same theory carries over to other function spaces, different norms and different pseudo-density functionals, as long as it is possible to compactly embed the sample space $\mathcal{X}$ in a larger space and the chosen pseudo-density functional is sufficiently smooth. In particular, we remark that the results of this paper can be straightforwardly generalized to arbitrary pairs of Sobolev spaces of integer order satisfying the compact embedding requirement.

The theory of clustering that we develop in this work is projection-free, since it does not involve projecting the random curves onto a finite-dimensional space. However, if the probability law P of the functional data is supported on a finite-dimensional space and admits a proper density with respect to the Lebesgue measure, we show that the gradient flow on the pseudo-density p_h and the gradient flow on the expectation of the kernel density estimator of the data coincide (and so coincide the corresponding population clusterings).

One of the most important practical tasks in modal clustering is to identify significant local modes, as these are associated to informative clusters. We provide an algorithm that

- identifies the local modes of the population pseudo-density p_h by analyzing its sample version $\hat{p}_h$; furthermore, all of the local modes of $\hat{p}_h$ identified by the algorithm converge asymptotically to their population correspondents of p_h
- is consistent (under additional regularity assumptions on p_h), in the sense that it establishes a one-to-one correspondence between the sample local modes that it identifies and their population equivalents.

While from a purely mathematical standpoint a sample of functional data $\{X_i\}_{i=1}^n$ is thought of as a collection of continuous curves defined on an interval, we never observe such objects in practice. Rather, we typically only observe noisy measurements of the X_i's at a set of design points $\{t_j\}_{j=1}^m$. As an intermediate

step, we therefore estimate the X_i's from these observations (which constitutes a typical regression problem), and then use the estimates as the input of our procedure.

The remainder of the paper is organized as follows. Section 2 provides a concise literature review. Section 3 is devoted to the development of our theory of population clustering for smooth random curves. In particular, there we study in detail the L^2 gradient flow on the pseudo-density p_h and establish that, in analogy to the finite-dimensional case, population clusters of smooth random curves can be defined in terms of the basins of attraction of the critical points of p_h. Section 4 describes the behavior of the L^2 gradient flow of p_h when the probability law P of the data is supported on a finite-dimensional subspace. Section 5 provides an algorithm to identify the significant local modes of $\hat{p}_h$ and shows that, under additional regularity assumptions, the algorithm is consistent. Section 6 extends the results of Section 5 to discretized and noisy functional data, and Section 7 provides examples of application of the clustering methodology described in this paper both to simulated and real functional data. Section 8 contains a discussion on the choice of the pseudo-density functional while Section 9 provides guidelines for the selection of the bandwidth parameter h. Section 10 summarizes the contributions of this paper. The proofs of the main results can be found in Appendix A, while auxiliary results (such as probability bounds for the estimation of the pseudo-density functional p_h and its derivatives) are deferred to Appendix B.

2. Related literature

The difficulties associated to the lack of proper density functions in infinite-dimensional spaces are well-known among statisticians. This has stimulated the introduction of various surrogate notions of density for functional spaces. The literature on pseudo-densities includes the work of Gasser, Hall and Presnell (1998), Hall and Heckman (2002), Delaigle and Hall (2010), and Ferraty, Kudraszow and Vieu (2012).

A population framework based on Morse theory for nonparametric modal clustering in the finite dimensional setting is presented in Chacón (2015). Whenever a proper density $p : \mathcal{X} \to \mathbb{R}^d$ exists and it is a Morse function, the problem of equation (1.2) induces an essential partition of the sample space $\mathcal{X} \subseteq R^d$ in the sense that each set C_i in the partition of $\mathcal{X}$ such that $P(C_i) > 0$ corresponds to the basin of attraction of a local mode μ_i of p, i.e. $C_i = \{x \in \mathcal{X} : \lim_{t\to\infty} \pi_x(t) = \mu_i\}$. Furthermore, if p has saddle points, the basin of attraction of each saddle is a null probability set (similarly, the basin of attraction of a local minimum is a singleton and hence negligible as well).

A number of gradient ascent algorithms have been developed to perform modal clustering in the finite-dimensional case. One of the most popular mode-finding and modal clustering algorithms is the *mean-shift algorithm* (Fukunaga and Hostetler, 1975; Cheng, 1995). A version of the mean-shift algorithm for functional data is discussed in Ciollaro et al. (2014). A gradient ascent algorithm for functional data is proposed in Hall and Heckman (2002).

The international FDA community is very vibrant. The most recent literature includes the work of Cuevas (2014) and Goia and Vieu (2015), who provide a general overview and an account of some key advances in high and infinite-dimensional statistics, as well as a book on functional ANOVA by Zhang (2013) and a book on the theoretical foundations of FDA by Hsing and Eubank (2015).

With particular focus on clustering, we would like to point out the recent work Bongiorno and Goia (2015) where the authors propose a clustering method for functional data based on the small ball probability function $\varphi_h(x) = P(\|X - x\|^2 \leq h)$ and on functional principal components. Finally, a recent overview of other clustering techniques for functional data can be found in Jacques and Preda (2013).

3. A population background for pseudo-density clustering of functional data

We denote by $X \sim P$ a functional random variable valued in $L^2([0,1])$, the space of square integrable functions on the unit interval with its canonical inner product $\langle x, y\rangle_{L^2} = \int_0^1 x(s)y(s)\,ds$ and induced norm $\|x\|_{L^2} = \sqrt{\langle x, x\rangle_{L^2}}$. As we previously mentioned, it is not possible to define a proper probability density function for P. Instead, we study the L^2 gradient flow of equation (1.2) associated to the functional

$$p_h(x) = E_P\, K\left(\frac{\|X - x\|_{L^2}^2}{h}\right) = \int_{\mathbb{R}} K(s)\, dP_{\|X-x\|_{L^2}^2/h}(s) \tag{3.1}$$

mapping $L^2([0,1])$ into $\mathbb{R}_+$, where $h > 0$ is a bandwidth parameter, $K : \mathbb{R}_+ \to \mathbb{R}_+$ is a kernel function, and $P_{\|X-x\|_{L^2}^2/h}$ denotes the probability measure induced by P through the map $X \mapsto \|X - x\|_{L^2}^2/h$. Note that p_h is closely related to the so-called *small-ball probability* function $\varphi_h(x) = P(\|X - x\|_{L^2}^2 \leq h)$. It is easy to see that $p_h(x) = \varphi_h(x)$ when $K(s) = \mathbb{1}_{[0,1]}(s)$, therefore p_h can be thought of as a smoother version of φ_h.

Unless otherwise noted, we make the following assumptions throughout the paper:

(H1) $K : \mathbb{R}_+ \to \mathbb{R}_+$ is twice continuously differentiable and the following bounds hold on the derivatives of $K_h(\cdot) = K(\cdot/h)$:
- $\sup_{t\in\mathbb{R}_+} \left|K_h(t^2)\right| \leq K_0 < \infty$
- $\sup_{t\in\mathbb{R}_+} \left|K_h'(t^2)t\right| \leq K_1 < \infty$
- $\sup_{t\in\mathbb{R}_+} \left\{\left|K_h^{(\ell-1)}(t^2)t^{\ell-2}\right| + \left|K_h^{(\ell)}(t^2)t^{\ell}\right|\right\} \leq K_\ell < \infty$, for $\ell = 2, 3$

where the constants K_0, K_1, K_ℓ may depend on h.

(H2) $K'(t^2) + K(t^2) \leq 0$ for all $t \in \mathbb{R}_+$.

(H3) X is P-almost surely absolutely continuous and its moments satisfy $E_P\|X\|_{L^2} \leq M_1 < \infty$ and $E_P\|X'\|_{L^2} \leq N_1 < \infty$ for some constants M_1 and N_1.

(H4) All the non-trivial critical points of p_h are isolated under the L^2 norm, i.e. there exists an open L^2 neighborhood around each critical point x^* of p_h with $p_h(x^*) > 0$ such that there are no other critical points of p_h that also belong to that neighborhood.

Various kernels can be shown to satisfy assumptions (H1) and (H2). For instance, both the compactly supported kernel $K(t) \propto (1-t)^3 \mathbb{1}_{[0,1]}(t)$ and the exponential kernel $K(t) \propto e^{-t}\mathbb{1}_{[0,\infty)}(t)$ satisfy our assumptions. (H3) is an assumption on the smoothness of the random curves. Intuitively, (H3) corresponds to assuming that the probablility law P does not favor curves that are too irregular or wiggly. (H4) is a regularity assumption on the functional p_h: essentially, under the above assumptions on K, (H4) corresponds to assuming that the functional p_h does not have flat "ridges" in regions where it is positive.

Remark 1. A sufficient condition for (H4) to hold is that p_h is a Morse functional. The following Proposition provides a sufficient condition under which p_h is a Morse functional.

Proposition 1. *Suppose that P has density p with respect to the Lebesgue measure and p is supported on a finite-dimensional compact domain $S_c \subset \mathbb{R}^d$. Suppose furthermore that p and ∂S_c, the boundary of S_c, satisfy*

- *∂S_c is smooth enough so that the normal vector $n(x)$ exists for any $x \in \partial S_c$*
- *p is continuous on $\mathbb{R}^d$*
- *p is twice differentiable in the interior of S_c,* $\mathrm{int}(S_c)$
- *∇p is not vanishing on ∂S_c.*

Then, for h sufficiently small, all the critical points of p_h in $\mathrm{int}(S_c)$ *are non-degenerate and there are no non-trivial critical points outside of* $\mathrm{int}(S_c)$.

In order to simplify the discussion, from now on we focus on the shifted random curves $X - X(0)$; however, with a little abuse of notation, we will keep using the letter X to mean $X - X(0)$. This choice is just made for convenience as it significantly simplifies the proofs of many of the results that we present. Following this notational convention, X thus belongs P-almost surely to the space $H_0^1([0,1]) = \{x : [0,1] \to \mathbb{R} \text{ such that } \|x'\|_{L^2} < \infty \text{ and } x(0) = 0\}$. Poincaré inequality ensures that the semi-norm $\|x'\|_{L^2}$ is in fact a norm on $H_0^1([0,1])$ and $\|x\|_{L^2} \leq C_p\|x'\|_{L^2}$ with $C_p = 1$ (i.e. $H_0^1([0,1])$ can be continuously embedded in $L^2([0,1])$). In the following, we denote $\|x\|_{H_0^1} = \|x'\|_{L^2}$ for $x \in H_0^1$. Moreover, to alleviate the notation, from now on we denote $L^2 = L^2([0,1])$ and $H_0^1 = H_0^1([0,1])$. If the curves were not shifted so that $X(0) = 0$, then they would belong P-almost surely to $H^1 = H^1([0,1]) = \{x : [0,1] \to \mathbb{R} \text{ such that } \|x\|_{L^2} + \|x'\|_{L^2} < \infty\}$, which can still be continuously embedded in L^2.

The main goal of this section is to show that the L^2 gradient flow associated to p_h is well-defined. In particular, we establish the following facts:

1. the L^2 gradient flow associated to p_h is a flow in H_0^1
2. for any initial value in H_0^1, there exists exactly one trajectory of such flow which is a solution to the initial value problem of equation (1.2)

3. for any initial value in H_0^1, the unique solution of the initial value problem of equation (1.2) converges to a critical point of p_h as $t \to \infty$ and the convergence is with respect to the L^2 norm
4. all the non-trivial critical points of p_h are in H_0^1, the support of P.

These facts guarantee that the clusters described in equation (1.3) exist and are well-defined.

Remark 2. In general, in an infinite dimensional Hilbert space, the trajectory of the solution of an ordinary differential equation such as the one of equation (1.2) may not converge as $t \to \infty$. In fact, such trajectory can be entirely contained in a closed and bounded set without converging to any particular point of that set. To guarantee the convergence of the gradient flow trajectories, one needs that (see Jost, 2011)

1. the trajectories satisfy some compactness property
2. the functional of interest (in our case p_h) is reasonably well-behaved: for instance it is smooth, with isolated critical points.

In L^2, compactness is a delicate problem: no closed bounded ball in L^2 is compact. However, any closed and bounded H^1 ball is compact with respect to the L^2 norm (and so is any closed and bounded H_0^1 ball). In fact, H^1 can be compactly embedded in L^2 (see, for instance, Chapter 5.7 of Evans, 1998), which means that every bounded set in H^1 is totally bounded in L^2 and H^1 can be continuously embedded in L^2. Since H_0^1 is a closed subspace of H^1, H_0^1 can also be compactly embedded in L^2. From a theoretical point of view, L^2 is strictly larger than H^1. However, H^1 is dense in L^2.

The remainder of our discussion focuses on the main results of this section, which concern the computation of the derivatives of p_h and their properties, the existence, the uniqueness, and the convergence of the solution of the initial value problem of equation (1.2).

Before we state our results, let us recall that for a functional random variable $X \sim P$ valued in $L^2([0,1])$ the expected value of X is defined as the element $E_P X \in L^2([0,1])$ such that $E_P \langle X, y\rangle_{L^2} = \langle E_P X, y\rangle_{L^2}$ for all $y \in L^2([0,1])$ (Horváth and Kokoszka, 2012). Furthermore, the expectation commutes with bounded operators. Also, recall that for a functional F mapping a Banach space B_1 into another Banach space B_2, the Frechét derivative of F at a point $a \in B_1$ is defined, if it exists, as the bounded linear operator DF such that $\|F(a+\delta) - F(a) - DF(\delta)\|_{B_2} = o(\|\delta\|_{B_1})$. The most common case in this paper sets $B_1 = L^2$, $B_2 = \mathbb{R}_+$, and $F = p_h$. Because DF is a bounded linear operator, if B_1 is also an Hilbert space then the Riesz representation theorem guarantees the existence of an element $\nabla F(a) \in B_1$ such that, for any $b \in B_1$, $DF(b) = \langle b, \nabla F(a)\rangle_{B_1}$. The element $\nabla F(a)$ corresponds to the gradient of F at $a \in B_1$. In this way, the gradient $\nabla F(a)$ and the first derivative operator DF at $a \in B_1$ can be identified. In the following, with a slight abuse of notation, we will use DF both to mean the functional gradient of the operator F (which is an element of B_1) and its Frechét derivative (which is a bounded linear operator from B_1 to B_2). It will be clear from the context whether we are referring to the derivative

operator or to the functional gradient. Note that higher order Frechét derivatives can be similarly identified with multilinear operators on B_1 (see, for example, Ambrosetti and Prodi, 1995).

Recall that, by assumption, the function $K_h(\|X - x\|_{L^2}^2)$ is bounded from above by a constant K_0. Furthermore, it is three times differentiable and its first Frechét derivative at x is

$$DK_h(\|X - x\|_{L^2}^2) = 2K_h'\left(\|X - x\|_{L^2}^2\right)(x - X). \tag{3.2}$$

The second Frechét derivative at x corresponds to the symmetric bilinear operator

$$\begin{aligned} D^2K_h(\|X - x\|_{L^2}^2)(z_1, z_2) &= 2K_h'\left(\|X - x\|_{L^2}^2\right)\langle z_1, z_2\rangle_{L^2} \\ &+ 4K_h''\left(\|X - x\|_{L^2}^2\right)\langle x - X, z_1\rangle_{L^2}\langle x - X, z_2\rangle_{L^2} \end{aligned} \tag{3.3}$$

for $z_1, z_2 \in L^2$.

Remark 3. Any bounded bilinear operator B on L^2 can be represented as a bounded linear operator from L^2 to L^2. In fact, let z_1 be any element of L^2; then, $B(z_1, \cdot)$ is a bounded linear operator from L^2 to $\mathbb{R}$. By the Riesz representation theorem, one can define $B(z_1) \in L^2$ by letting $\langle B(z_1), z_2\rangle_{L^2} = B(z_1, z_2)$ for any $z_2 \in L^2$. The operator norm of B is then defined by

$$\|B\| = \sup_{\{v\ :\|v\|_{L^2}=1\}} \|B(v)\|_{L^2}. \tag{3.4}$$

It is straightforward to check that both derivatives correspond to bounded linear operators under assumption (H1). The following Lemma provides the first and the second Frechét derivatives of p_h.

Lemma 1. *Under assumption (H1) the Frechét derivative of $p_h : L^2 \to \mathbb{R}$ at x corresponds to the L_2 element*

$$Dp_h(x) = 2E_P\, K_h'\left(\|X - x\|_{L^2}^2\right)(x - X). \tag{3.5}$$

The second Frechét derivative of p_h at x corresponds to the symmetric bilinear operator

$$\begin{aligned} D^2p_h(x)(z_1, z_2) = E_P\ \big[&4K_h''\left(\|X - x\|_{L^2}^2\right)\langle x - X, z_1\rangle_{L^2}\langle x - X, z_2\rangle_{L^2} \\ &+2K_h'\left(\|X - x\|_{L^2}^2\right)\langle z_1, z_2\rangle_{L^2}\big]. \end{aligned} \tag{3.6}$$

Furthermore, both derivatives have bounded operator norm for any $x \in L^2([0, 1])$.

We state without proof the following standard Lemma.

Lemma 2. *Let $v \in C_c^\infty([0, 1])$ be a compactly supported infinitely differentiable function. Suppose $f \in L^2([0, 1])$ is such that $\langle f, v'\rangle_{L^2} = L(v)$ for any such v, where $L \in L^2([0, 1])^*$ is a bounded linear operator. Then the weak first derivative f' of f exists, $f' \in L^2([0, 1])$, and $\|f'\|_{L^2} = \|L\|_{(L^2)^*}$. Moreover, $\langle f', v\rangle_{L^2} = -L(v)$ for any $v \in C_c^\infty([0, 1])$ and therefore for any $v \in L^2([0, 1])$.*

The following Proposition shows that the L^2 gradient of p_h is an element of H_0^1. Intuitively, this means that if the starting point of the initial value problem of equation (1.2) is in H_0^1 (and a solution exists for that starting point), then we should expect that the path π_x only visits elements of H_0^1, i.e. the L^2 gradient flow associated to p_h is a H_0^1 flow.

Proposition 2. *For any $x \in H_0^1$, the L^2 gradient of p_h at x, $Dp_h(x)$, is an element of H_0^1 such that for any $y \in L^2$,*

$$\langle Dp_h(x)', y\rangle_{L^2} = E_P\left[-2K_h'(\|X - x\|_{L^2}^2)\langle x' - X', y\rangle_{L^2}\right]. \tag{3.7}$$

Proposition 2 also implies that the equation $\frac{d}{dt}\pi_x(t) = Dp_h(\pi_x(t))$ is meaningful when restricted to H_0^1. The next Lemma, Lemma 3, establishes that Dp_h is locally Lipschitz under the H_0^1 norm. The subsequent Lemma, Lemma 4, guarantees that a solution of the problem (if it exists) is necessarily bounded. These two Lemmas allow us to claim that if the starting point $\pi_x(0) = x$ is an element of H_0^1, then the initial value problem of equation (1.2) has a unique solution in H_0^1. This claim is summarized in Proposition 3.

Lemma 3. *Under (H1), the L^2 gradient of p_h corresponds to a locally Lipschitz map in $H_0^1([0,1])$.*

Lemma 4. *The following two results hold under (H1) and (H2)*

1. *Suppose that $p_h(\pi_x(0)) \geq \delta > 0$. If $\|\pi_x(t)\|_{H_0^1} \geq K_2N_1/\delta$, then $\langle Dp_h(\pi_x(t)), \pi_x(t)\rangle_{H_0^1} \leq 0$.*
2. *Let $M > 0$. If $\|X\|_{H_0^1} \leq M$ a.s., then $\langle Dp_h(\pi_x(t)), \pi_x(t)\rangle_{H_0^1} \leq 0$ as soon as $\|\pi_x(t)\|_{H_0^1} > M$.*

The intutive interpretation of Lemma 4 is that a trajectory π_x that is a solution to the initial value problem of equation (1.2) cannot wander too far from the origin in H_0^1. In fact, if the H_0^1 norm of π_x increases too much, then the path π_x is eventually pushed back into the closed and bounded H_0^1 ball of radius K_2N_1/δ (or radius M if one makes the stronger assumption that the probability law P of the random curves is completely concentrated on the H_0^1 ball of radius M). This "push-back" effect is captured by the condition $\langle Dp_h(\pi_x(t)), \pi_x(t)\rangle_{H_0^1} \leq 0$. By combining Lemma 3 and Lemma 4, we obtain the following

Proposition 3. *Under assumptions (H1), (H2), and (H3), the initial value problem $\pi_x'(t) = Dp_h(\pi_x(t))$ with $x = \pi_x(0) \in H_0^1$ has a unique solution in H_0^1 with respect to the H_0^1 topology. Moreover, if $\|x\|_{H_0^1} \leq R$, then $\|\pi_x(t)\|_{H_0^1} \leq C_1$ for all $t \geq 0$, where $C_1 = C_1(R, K_2, N_1, p_h(x))$.*

Remark 4. Proposition 3 establishes the existence and the uniqueness of a solution to the initial value problem of equation (1.2) in the H_0^1 topology. The initial value problem can be solved uniquely in the L^2 topology as well. In fact, it is easily verified that, because D^2p_h is bounded, then the first derivative of p_h, $Dp_h : L^2 \to L^2$, is uniformly Lipschitz with respect to the L^2 norm. Thus, one only has to show that the H_0^1 flow π_x of Proposition 3 solved in the H_0^1

topology corresponds to the L^2 gradient flow associated to p_h. To verify this, one needs to check that the H_0^1 solution also satisfies the initial value problem of equation (1.2) under the L^2 norm. Specifically, consider the H_0^1 solution π_x of Proposition 3 with any $\pi_x(0) = x \in H_0^1$. The path π_x is continuously differentiable as a map from $\mathbb{R}_+$ to H_0^1. It suffices to check that $\pi_x(t)$ is continuously differentiable as a map from $\mathbb{R}_+$ to L^2 as well. This is easily established using Poincaré inequality since

$$\begin{aligned} &\|\pi_x(t+\delta) - \pi_x(t) + Dp_h(\pi_x(t))\|_{L^2} \\ &\leq C_p \|\pi_x(t+\delta) - \pi_x(t) + Dp_h(\pi_x(t))\|_{H_0^1} = o(\delta), \end{aligned} \tag{3.8}$$

where the Poincaré constant is $C_p = 1$ for the pair (L^2, H_0^1). It is clear from equation (3.8) and the definition of Frechét derivative that the H_0^1 solution π_x also satisfies the initial value problem of equation (1.2) under the L^2 norm. Thus, π_x is the unique L^2 solution of the intial value problem of equation (1.2).

The following Theorem, based on Proposition 3, guarantees the convergence of π_x to a critical point of p_h as $t \to \infty$. The statement about the convergence strongly relies on the compact embedding of H_0^1 in L^2, the boundedness of the first two derivatives of p_h, and assumption (H4).

Theorem 1. *Assume (H1), (H2), (H3), and (H4) hold. Let π_x be the H_0^1 solution of the initial value problem of equation (1.2) with $x = \pi_x(0) \in H_0^1$. Let $C_1 > 0$ be such that $\|\pi_x(t)\|_{H_0^1} \leq C_1$ for all $t \geq 0$. Then there exists a unique $\pi_x(\infty) \in L^2$ such that $\|\pi_x(\infty)\|_{H_0^1} \leq C_1$, $\lim_{t\to\infty} \|\pi_x(t) - \pi_x(\infty)\|_{L^2} = 0$, and $Dp_h(\pi_x(\infty)) = 0$.*

The results above show that the L^2 gradient flow on p_h is well-defined and its trajectories converge to critical points of p_h that are in H_0^1 whenever the starting point $x = \pi_x(0)$ is an element of H_0^1. We conclude this section with the following Lemma which states that all the non-trivial critical points of p_h belong to H_0^1: thus, even though the functional p_h "spreads" the probability law P of the random curves outside of its support H_0^1 (in fact, it is easily seen that there exists points $x \in L^2$ that are not in H_0^1 with $p_h(x) > 0$), all of its non-trivial critical points still lie in the support of P.

Lemma 5. *Assume (H1), (H2), and (H3) hold. Let $x \in L^2$ be a critical point of p_h such that $p_h(x) > 0$ (i.e. x is a nontrivial critical point of p_h). Then $x \in H_0^1$. Furthermore, if $\|X\|_{H_0^1} \leq M$ P-almost surely, then all the nontrivial critical points of p_h are contained in $B_{H_0^1}(0, M)$.*

Note that the stronger assumption that $P(\|X\|_{H_0^1} \leq M) = 1$ is a functional analogue of the boundedness assumption which is frequently made with finite-dimensional data.

4. Finite-dimensional adaptivity

If $X \sim P$ is a functional random variable and P is supported in a finite-dimensional subspace of L^2 of dimension d, there is a d-dimensional orthonor-

mal system that one can use to represent X in terms of a d-dimensional vector of Fourier coefficients. In this case X is just a d-dimensional random vector, hence the statistical model is *multivariate* and not really a functional one. Suppose now that P admits a proper density p which is supported on the span of this d-dimensional orthonormal system. Then, in this case, one could use a traditional kernel density estimator $\hat{p}_h^{\mathrm{KDE}}$ of p and estimate its population clusters based on its gradient flow. This section clarifies that the gradient flow of $p_h(x) = E_P K_h(\|X - x\|_{L^2}^2)$ (which is a functional defined on L^2) and the gradient flow of $E_P\, \hat{p}_h^{\mathrm{KDE}}$ (which is a function defined on $\mathbb{R}^d$) are equivalent. Since the two gradient flows are equivalent, the population clustering induced by p_h and $E_P\, \hat{p}_h^{\mathrm{KDE}}$ is the same. This is essentially a consequence of the isometric isomorphism between any d-dimensional finite-dimensional subspace of L^2 and $\mathbb{R}^d$.

Keeping this in mind, we can now move on to the mathematical details. For the remainder of this section, we assume that the distribution of the random function X is supported on some compact subset S_c of a finite dimensional vector space. In other words,

$$P(X \in S_c) = 1, \tag{4.1}$$

where S_c is a compact subset of a finite-dimensional subspace $S \subset L^2$. We discuss two insightful facts.

1. Under some mild extra assumptions on the finite dimensional distribution of X, it is shown in Lemma 7 that p_h, as a functional from L^2 to $\mathbb{R}_+$, is a Morse functional. This provides an important sufficient condition under which (H4) holds.
2. If the functional random variable X admits a finite-dimensional distribution on S_c, it is natural to ask whether the L^2 gradient flow on p_h corresponds to the finite-dimensional gradient flow associated to the expectation of a kernel density estimator of the density of X on S. This section provides a positive answer to this question. Furthermore, we show that such finite-dimensional gradient flow is entirely contained in S.

Suppose that the probability law P of the functional random variable X is supported on a compact subset S_c of a finite-dimensional space $S \subset L^2$. If this is the case, there exists $\delta > 0$ such that if $0 < h \leq \delta$, then p_h is a Morse function on the interior of S_c (see Remark 1 and Proposition 1). Moreover, as implied by Lemma 6 and Lemma 7 of this section, the trajectories of the L^2 gradient flow associated to p_h are all contained in S and they end at critical points of p_h that belong to S_c. It is natural to ask whether the L^2 gradient flow on p_h corresponds to the finite-dimensional gradient flow associated to some pseudo-density on S. This section answers this question and shows that, if X admits a density function p (when X is viewed as a finite-dimensional random vector in S_c), then the L^2 gradient flow associated to p_h corresponds to the gradient flow associated to the expectation of a kernel density estimator of p with bandwidth h.

Let $S = \text{span}\{f_1 \dots f_d\}$ be a linear subspace of L^2. Without loss of generality, assume that the f_i's form an orthonormal basis of S equipped with the L^2 norm and that $X \in S_c$ almost surely. Then, X admits the decomposition $X = a_1 f_1 + \ldots + a_d f_d$ for some random coefficients $\{a_i\}_{i=1}^d$. Let $\tilde{X} = [a_1, \ldots, a_d]^T$ and suppose that the distribution of $\tilde{X}$ has density $p : \mathbb{R}^d \to \mathbb{R}_+$ with respect to the Lebesgue measure. We have the following

Lemma 6. *Assume (H1) and (H2) hold and $P(X \in S_c) = 1$. If $x = \pi_x(0) \in S$, then $\pi_x(t) \in S$ for any $t \geq 0$. Furthermore, all the non-trivial critical points of p_h belong to S.*

For the rest of this section, let us replace assumption (H4) with

(H4') X is an element of S_c with probability 1, $X \sim P$ admits density p on S_c, and p satisfies the assumptions of Proposition 1.

Consider $x = x_1 f_1 + \cdots + x_d f_d \in S$. Let $\tilde{x} = [x_1, \ldots, x_d]^T \in \mathbb{R}^d$. Define $\tilde{p}_h(\tilde{x}) : \mathbb{R}^d \to \mathbb{R}_+$ to be $\tilde{p}_h(\tilde{x}) = E_P K_h(\|\tilde{X} - \tilde{x}\|_2^2)$, where $\|\cdot\|_2$ denotes the standard Euclidean norm. Note that $\frac{1}{h^d}\tilde{p}_h(\tilde{x})$ is the expectation of a standard finite dimensional kernel density estimator at $\tilde{x}$. Since $\|X - x\|_{L^2}^2 = \|\tilde{X} - \tilde{x}\|_2^2$, it is clear that $p_h(x) = \tilde{p}_h(\tilde{x})$. To see the connection between the functional gradient $Dp_h(x)$ and $\nabla\tilde{p}_h(\tilde{x})$, the gradient of $\tilde{p}_h$ at $\tilde{x}$, note that the random variable

$$\begin{aligned} \langle Dp_h(x), f_i\rangle_{L^2} &= \langle E_P\, 2K_h'(\|X - x\|_{L^2}^2)(x - X), f_i\rangle_{L^2} \\ &= E_P\, 2K_h'(\|X - x\|_{L^2}^2)\langle x - X, f_i\rangle_{L^2} \\ &= 2E_P K_h'(\|\tilde{X} - \tilde{x}\|_2^2)(x_i - a_i) \end{aligned} \tag{4.2}$$

agrees with the i-th component of the gradient of $\tilde{p}_h$ at $\tilde{x}$. This equivalence implies that the gradient flow (with starting points in the subspace S) on p_h and $\tilde{p}_h$ coincide (note that scaling the $\tilde{p}_h$ by h^{-d} does affect the associated gradient flow). Furthermore, there exists a $\delta > 0$ depending on p such that $\tilde{p}_h(\tilde{x})$ is a Morse function for $0 < h \leq \delta$ (see Remark 1). Therefore, all the non-trivial critical points of $\tilde{p}_h$ are separated in $\mathbb{R}^d$. In light of Lemma 6, all the non-trivial critical points of p_h are thus separated in S (and in L^2).

Next, we have the following Lemma which guarantees that if p is a Morse density on S_c, then the non-trivial critical points of p_h are non-degenerate for h sufficiently small and they all belong to S_c (a critical point x^* of p_h is non-degenerate if $D^2 p_h(x^*)$ is an isomorphism from L^2 to L^2).

Lemma 7. *Under assumption (H1) (H2) and (H4'), all the non-trivial critical points of p_h lie in S_c and are non-degenerate for h sufficiently small. Thus, for sufficiently small h, (H4) holds.*

In the finite-dimensional case considered in this section, we can say more about the behavior of the L^2 gradient flow on p_h. In particular, we can characterize the solutions to the initial value problem of equation (1.2) also for the case in which the starting point $x = \pi_x(0)$ does not belong to the support of P (which is, in this case, $S_c \subset L^2$). In fact, let x be an element of L^2 which does not belong to S. The Gram-Schmidt orthogonalization process guarantees that

there exists $S' \supset S$ such that $x = \pi_x(0) \in S'$ and $S' = \text{span}\{f_1, \dots f_d, f_{d+1}\}$, where f_{d+1} is orthogonal to $\{f_i\}_{i=1}^d$ and $\|f_{d+1}\|_{L^2} = 1$. The following Lemma guarantees that the gradient ascent path originating from x is entirely contained in S'. Its proof is identical to that of Lemma 6.

Lemma 8. *Assume (H1) and (H2) hold and that* $P(X \in S_c) = 1$. *Suppose* $x = \pi_x(0) \in S'$. *Then* $\pi_x(t) \in S'$ *for all* $t \geq 0$.

Remark 5. In the finite-dimensional setting of this section (in particular under assumption (H4')), and for h sufficiently small, the basin of attraction of a saddle point of p_h is negligible: in fact, from the above disussion, it is clear that if the random function $X \sim P$ is valued in a compact subset S_c of a finite-dimensional linear subspace S of L^2 and P has a proper Morse density p on S_c, then the basin of attraction of any saddle point of p_h is neglibible for h sufficiently small (since p_h is Morse on $\text{int}(S_c)$ for h small enough). Stated more precisely, for h sufficiently small, if $x_0^* \in \text{int}(S_c)$ is a saddle point of p_h then $P(\{x \in S : \lim_{t\to\infty} \|\pi_x(t) - x_0^*\|_{L^2} = 0\}) = 0$.

5. Statistical relevance of the estimated local modes

The empirical counterpart of $p_h(x) = E_P K_h(\|X - x\|_{L^2}^2)$ is the functional $\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^n K_h(\|X_i - x\|_{L^2}^2)$, where $\{X_i\}_{i=1}^n$ are i.i.d functional random variables with probability law P. The critical points of $\hat{p}_h$ can be found, for example, by using a functional version of the mean-shift algorithm (see Hall and Heckman, 2002; Ciollaro et al., 2014). In this section, we provide a statistical algorithm to detect whether a critical point of $\hat{p}_h$ corresponds to a local maximum of p_h. This algorithm provides two insights for functional mode clustering.

1. For finite-dimensional clustering problems, if the underlying density p is a Morse function, then the basin of attraction of a saddle point of p has null probability content as it corresponds to a manifold of lower dimension. In functional data clustering, however, the structure of the functional space is more complicated in the sense that there is no guarantee that the probability content of the basin of attraction of a saddle point of p_h is negligible, even if p_h is a Morse function. However, in analogy with the finite-dimensional case, clusters associated to non-degenerate local modes should generally be considered more informative as opposed to clusters associated with saddle points.
2. Several results in the previous section are derived under assumption (H4), which essentially states that the relevant critical points of p_h are well-behaved. Without assuming (H4), the algorithm provides a simple way to classify well-behaved local modes of p_h by analyzing $\hat{p}_h$. Thus, informative clusters can still be revealed in a less restrictive setting.

Since the local modes of $\hat{p}_h$ that correspond to non-degenerate local modes of p_h provide the greatest insight about the population clustering, we refer to these local modes as "significant" local modes. In the following, we derive a

procedure that allows us to discriminate the significant local modes from the non-significant ones.

Before giving the definition non-degeneracy for a critical point of a functional defined on an Hilbert space (L^2 in our case), it is convenient to adopt the convention that a linear operator from an Hilbert space to itself can be associated to a bilinear form on the Hilbert space and vice versa. For example if $T : L^2 \to L^2$ is a linear operator, then it can be associated to a bilinear form by letting $T(v, w) = \langle Tv, w\rangle_{L^2}$.

Definition 1. *Let $T : L^2 \to L^2$ be a bounded linear operator. T is said to be self-adjoint if $\langle Tv, w\rangle = \langle v, Tw\rangle$. T is said to be positive (respecively negative) definite if $\langle Tv, v\rangle > 0$ (respectively < 0) for all $v \neq 0$. Furthermore, T is said to be an isomorphism if both T and T^{-1} are bounded.*

Definition 2. *Let $f : L^2 \to \mathbb{R}$ be twice continuously differentiable with bounded third derivative. Suppose x^* is a critical point of f, i.e. $Df(x^*) = 0$. Then, x^* is said to be a non-degenerate local maximum (respectively minimum) if $D^2 f(x^*)$ is a negative (respectively positive) definite isomorphism on L^2.*

It is a known fact that for any x, the second derivative of f, $D^2 f(x)$, is a self-adjoint linear operator. Furthermore, the following Lemma follows as a simple consequence of the fact that the second derivative of f at a non-degenerate local maximum is a self-adjoint negative-definite isomorphism.

Lemma 9. *Suppose that x^* is a non-degenerate local maximum of f. Then there exist $\delta > 0$ such that*

$$\sup_{\|v\|_{L^2}=1} D^2 f(x^*)(v, v) \leq -\delta. \tag{5.1}$$

Let now $f_1, f_2 : L^2 \to \mathbb{R}_+$ be twice continuously differentiable with bounded third derivative. Consider the following abstract setting for f_1 and f_2.

(C1) The non-trivial critical points of f_1 and f_2 are all in H_0^1.

(C2) For $i = 1, 2$, if $x \in H_0^1$ then $Df_i(x) \in H^1$. Moreover,

$$\begin{cases} \pi_i'(t) = Df_i(\pi_i(t)) \\ \pi_i(0) \in H_0^1, \end{cases} \tag{5.2}$$

have H_0^1 solutions whose trajectories admit a convergent subsequence in L^2.

(C3) For $\ell = 0, 1, 2$, let η_ℓ denote

$$\eta_\ell = \sup_{x \in B_{H_0^1}(0,M)} \|D^\ell f_1(x) - D^\ell f_2(x)\|, \tag{5.3}$$

where $\|\cdot\|$ stands for the appropriate norms. Also, for $i = 1, 2$ and $k = 0, 1, 2, 3$, let

$$\beta_k = \sup_{x \in L^2} \|D^k f_i(x)\| < \infty. \tag{5.4}$$

Remark 6. Of course, the results that we obtain here are most useful for the particular case where

$$\begin{aligned} f_1(x) &= p_h(x) = E_P\, K_h(\|X - x\|_{L^2}^2) \\ f_2(x) &= \hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(\|X_i - x\|_{L^2}^2) \end{aligned} \tag{5.5}$$

and $X_1, \ldots, X_n \sim P$ are i.i.d. functional random variables valued in H_0^1. In this case, Lemma 5 and Proposition 3 provide sufficient conditions for (C1) and (C2), respectively. The boundedness for β_k is ensured by (H1), and the probability bounds in Appendix B guarantee that η_l converges to 0 as the sample size n increases.

Lemma 10. *Suppose conditions (C1), (C2) and (C3) hold. Let x_2^* be a non-degenerate local maximum of f_2 such that $\|x_2^*\|_{H_0^1} \leq M$. By Lemma 9, there exists $\delta(x_2^*) > 0$ such that $\sup_{\|u\|_{L^2}=1} D^2 f_2(x_2^*)(u,u) := -\delta(x_2^*) < 0$. If $\eta_1 \leq \delta^2(x_2^*)/(8\beta_3)$ and $\eta_2 \leq \delta(x_2^*)/8$, there exists $x_1^* \in B_{L^2}(x_2^*, \delta(x_2^*)/(2\beta_3))$, such that*

1. *x_1^* is a unique local maximum of f_1 in $B_{L^2}(x_2^*, \delta(x_2^*)/(2\beta_3))$*
2. *$\sup_{\|u\|_{L^2}=1} D^2 f_1(x_1^*)(u,u) \leq -3\delta(x_2^*)/8$*
3. *$\|x_1^* - x_2^*\|_{L^2} \leq 8\eta_1/\delta(x_2^*)$.*

Consider $f_1(x) = p_h$, $f_2(x) = \hat{p}_h(x)$ as in equation (5.5). For any $\alpha \in (0,1)$, we can derive a procedure based on Lemma 10 which allows us to classify non-degenerate local modes of $\hat{p}_h$ as significant and construct an L^2 neighbor around them with the property that the probability that each of such neighbors contains a non-degenerate local mode of p_h is at least $1-\alpha$ for n large enough. The procedure is summarized in Display 1 and its statistical guarantees are described in Proposition 4.

Learning non-degenerate local modes

Input: data, $X_1, \ldots, X_n$; kernel function, K; bandwidth, $h > 0$; significance level $\alpha \in (0,1)$.

Output: a set $\hat{\mathcal{R}}$ of significant local modes of $\hat{p}_h$.

1. Compute $\hat{p}_h$ and determine the set of non trivial local max of $\hat{p}_h$, $\hat{\mathcal{C}}$ (here non-trivial means $\hat{x}^* \in \hat{\mathcal{C}} \Rightarrow \hat{p}_h(\hat{x}^*) > 0$).
2. If $\hat{x}^* \in \hat{\mathcal{C}}$ is such that $\delta(\hat{x}^*) := -\sup_{\|u\|_{L^2}=1} D^2\hat{p}_h(\hat{x}^*)(u,u) \geq \max\{\sqrt{8\beta_3 C_1(\alpha)}, 8C_2(\alpha)\}$ where

$$C_1(\alpha) = \left(\frac{125MK_1^2K_2}{2n}\right)^{\frac{1}{3}} + \left(\frac{25K_1^2\log(2\log(n)/\alpha)}{4n}\right)^{\frac{1}{2}}$$

and

$$C_2(\alpha) = \left(\frac{125MK_2^2K_3}{4n}\right)^{\frac{1}{3}} + \left(\frac{25K_2^2\log(2\log(n)/\alpha)}{8n}\right)^{\frac{1}{2}}$$

then classify $\hat{x}^*$ as a significant local mode of $\hat{p}_h$. Here $\beta_3 = 12K_3$.

Display 1

Proposition 4. *Consider $f_1(x) = p_h$, $f_2(x) = \hat{p}_h(x)$. Assume (H1) and (H2) hold and $P(\|X\|_{H_0^1} \leq M) = 1$ for some known $M > 0$. Let $\hat{\mathcal{R}}$ denote the set of points classified by the algorithm of Display 1. Then, for large enough n, with probability $1 - \alpha$ the following holds for all $\hat{x}^* \in \hat{\mathcal{R}}$:*

1. *the random ball $B_{L^2}(\hat{x}^*, \delta(\hat{x}^*)/(2\beta_3))$ contains a unique non-degenerate local mode x^* of p_h*
2. *$\|x^* - \hat{x}^*\|_{L^2} \leq 8C_1(\alpha)/\delta(\hat{x}^*)$.*

Let $\mathcal{R}$ denote the set of non-degenerate local modes of p_h. Consider the map $\Phi : \hat{\mathcal{R}} \to \mathcal{R}$ by letting

$$\Phi(\hat{x}^*) = B_{L^2}(\hat{x}^*, \delta(\hat{x}^*)/(2\beta_3)) \cap \mathcal{R} \cap B_{L^2}(\hat{x}^*, \log(n)\, C_1(\alpha)/\delta(\hat{x}^*)). \tag{5.6}$$

According to Proposition 4, with probability $1 - \alpha$, for every $\hat{x}^* \in \hat{\mathcal{R}}$, there exists a unique $x^* \in \mathcal{R}$ contained in the right hand side of equation (5.6). In other words, with probability $1 - \alpha$, Φ is a well-defined map. Under suitable assumptions on $p_h(x)$, more can be said.

Proposition 5. *Assume that (H1) and (H2) hold and that $P(\|X\|_{H_0^1} \leq M) = 1$ for some known $M > 0$. Suppose further that p_h has finitely many non-degenerate local modes. Let $\mathcal{R}$ denote the collection of non-trivial local maxima of p_h. Then, with probability converging to 1 as $n \to \infty$, every $x^* \in R$ has a unique preimage of Φ in $\hat{\mathcal{R}}$.*

Remark 7. Under the assumptions of Proposition 4 and 5, one can conclude that with probability converging to $1 - \alpha$, the map $\Phi : \hat{\mathcal{R}} \to R$ is bijective. In other words, the algorithm of Display 1 is consistent.

6. From theory to applications

So far, all the results have been developed in an infinite-dimensional functional space. In this section, we connect the theory that we developed to practical applications and, in particular, we address the following challenges.

1. Complete functional data can never be observed: a functional datum is always observed on a discrete grid. For example, let $\{X_i\}_{i=1}^n$ be an i.i.d sample from a distribution P on H_0^1 and let $\{t_j\}_{j=1}^m$ be a set of equally spaced design points. In practice, only noisy measurements of the X_i's at $\{t_j\}_{j=1}^m$ are available. It is therefore important to design procedures that work with discretized curves.
2. While the theory is developed in an infinite-dimensional functional space, in practice any functional clustering method relies on the use of only finitely many basis functions. However, a flexible algorithm for functional data clustering should be asymptotically consistent with the infinite-dimensional theory.

One way to accomplish these two tasks at the same time is to apply a projection method. As shown later in this section, projections onto a linear space introduce small L^2 perturbations to the functional data and to the pseudo-density.

Nonetheless, the procedure that we describe is tolerant to such perturbations (see Corollary 1 for more details).

Before turning to the technical arguments, let us describe the following simple example which motivates the projection approach.

Example 1. *Consider the simple model* $y = X(t)+\epsilon$, *where* $X \sim P$ *is a random function and* ϵ *is a random variable independent of* X. *Instead of observing* n *complete random function samples* $\{X_i\}_{i=1}^n$, *one only observes the discrete noisy measurements* $\{y_{ij}\}$, *where* $y_{ij} = X_i(t_j) + \epsilon_{ij}$. *Here,* $\{t_j\}_{j=1}^m$ *is a set of equally spaced design points for the samples and the measurement errors* $\epsilon_{ij} \sim N(0, \sigma)$ *are independent of* $\{X_i\}_{i=1}^n$.

In Gasser and Müller (1984), for example, if one assumes further that the random function X *is bounded in* H^2, *i.e.*

$$\int_0^1 |X''(t)|dt \leq M_2 \quad P\text{-almost surely}, \tag{6.1}$$

it is shown that there exists a kernel W *so that an approximation of* X *can be constructed as*

$$\tilde{X}(t) = \sum_{j=1}^m \frac{y_{ij}}{b} \int_{t_{j-1}}^{t_j} W\left(\frac{t-u}{b}\right) du. \tag{6.2}$$

If b *is chosen to be of order* $m^{-1/5}$, *the above estimator also satisfies*

$$E(\|X - \tilde{X}\|_{L^2}^2 | X) \leq C(M_2, W)m^{-4/5} \tag{6.3}$$

and

$$E(\|X' - \tilde{X}'\|_{L^2}^2 | X) \leq C(M_2, W)m^{-2/5}, \tag{6.4}$$

where $C(M_2, W)$ *is a constant only depending on* M_2 *and* W, *and the expectation is taken with respect to* ϵ *only.*

As shown in the example, with noisy discrete measurements of the functional datum X, one can construct an approximation

$$\tilde{X} \in \text{span}\left\{\int_{t_{j-1}}^{t_j} W\left(\frac{t-u}{b}\right) du\right\}_{j=1}^m, \tag{6.5}$$

where b is of order $O(m^{-1/5})$. This approximation corresponds to a perturbed version of the underlying complete functional datum. The perturbation vanishes asymptotically as the number of discrete measurements m goes to infinity. This motivates the following assumption.

(H5) The collection $\{\tilde{X}_i\}_{i=1}^n$ of the i.i.d approximations of $\{X_i\}_{i=1}^n$ based on the equally spaced design points $\{t_j\}_{j=1}^m \subset [0,1]$ is such that $\{\tilde{X}_i\}_{i=1}^n \subset H_0^1$ and

$$E(\|X_i - \tilde{X}_i\|_{L^2} | X_i) \leq \phi(m), \tag{6.6}$$

where $\phi(m)$ does not depend on $\{X_i\}_{i=1}^n$ and $\phi(m) \to 0$ as $m \to \infty$.

Recall that in our theoretical results, the sample version of the pseudo density takes the form

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(\|X_i - x\|_{L^2}^2) \tag{6.7}$$

When the only available functional data are $\{\tilde{X}_i\}_{i=1}^n$ instead of $\{X_i\}_{i=1}^n$, one should consider

$$\tilde{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(\|\tilde{X}_i - x\|_{L^2}^2). \tag{6.8}$$

The following simple Lemma is useful to characterize the aforementioned L^2 perturbation and allows us to derive Corollary 1.

Lemma 11. *Let $\{X_i\}_{i=1}^n$ be an i.i.d. sample of functional data. Under assumptions (H1), (H2), (H3), and (H5), for $l = 0, 1, 2$,*

$$P\left(\sup_{x\in L^2} \|D^l\hat{p}_h(x) - D^l\tilde{p}_h(x)\| \geq \epsilon\right) \leq \frac{2^l K_{l+1}\phi(m)}{\epsilon} \tag{6.9}$$

where $\|\cdot\|$ stands for the appropriate L^2 operator norm.

Corollary 1. *Consider a modified version of the algorithm of Display 1, where $\hat{p}_h$ is replaced by $\tilde{p}_h$ and $C_1(\alpha), C_2(\alpha)$ are replaced by*

$$\begin{aligned}
\tilde{C}_1(\alpha) &= \left(\frac{125MK_1^2K_2}{2n}\right)^{\frac{1}{3}} + \left(\frac{25K_1^2\log(4\log(n)/\alpha)}{4n}\right)^{\frac{1}{2}} + \frac{8K_2\phi(m)}{\alpha} \\
\tilde{C}_2(\alpha) &= \left(\frac{125MK_2^2K_3}{4n}\right)^{\frac{1}{3}} + \left(\frac{25K_2^2\log(4\log(n)/\alpha)}{8n}\right)^{\frac{1}{2}} + \frac{16K_3\phi(m)}{\alpha}.
\end{aligned} \tag{6.10}$$

Let $\tilde{\mathcal{R}}$ be the significant local modes learned by this modified version of the algorithm. Then, the following statements are true.

1. *The probability that all random balls $B_{L^2}(\tilde{x}^*, \delta(\tilde{x}^*)/(2\beta_3))$ with $\tilde{x}^* \in \tilde{\mathcal{R}}$ contain a unique non-degenerate local mode x^* of p_h and that $\|x^* - \tilde{x}^*\|_{L^2} \leq 8\tilde{C}_1(\alpha)/\delta(\tilde{x}^*)$ is at least $1-\alpha$ for sufficiently large n.*
2. *Consider the map $\Phi : \tilde{\mathcal{R}} \to \mathcal{R}$ such that*

$$\Phi(\tilde{x}^*) = B_{L^2}(\tilde{x}^*, \delta(\tilde{x}^*)/(2\beta_3)) \cap \mathcal{R} \cap B(\tilde{x}^*, \log n\, \tilde{C}_1(\alpha)/\delta(\tilde{x}^*)), \tag{6.11}$$

where $\mathcal{R}$ denotes the collection of non-degenerate local modes of p_h. Suppose further that p_h has finitely many non-trivial local modes and that they are non-degenerate. Assume that $\{\tilde{X}_i\}_{i=1}^n \in B_{H_0^1}(0, M)$. Then, with probability converging to 1 as $n \to \infty$, every $x^* \in \mathcal{R}$ has a unique preimage in $\tilde{\mathcal{R}}$ with respect to Φ.

Remark 8. Let $\tilde{S} = \text{span}\{\int_{t_{j-1}}^{t_j} W(\frac{t-u}{b})du\}_{j=1}^m$, with $b = O(m^{-1/5})$. Although $\{\tilde{X}_i\}_{i=1}^n \subset \tilde{S}$, $\tilde{p}_h(x)$ as defined in equation (6.8) is still a functional on L^2.

It is desirable to have a method that does not use infinitely many L^2 basis functions to compute a non-degenerate local mode $\tilde{x}^*$ of $\tilde{p}_h$ and $\delta(\tilde{x}^*) := -\sup_{\|u\|_{L^2}=1} D\tilde{p}_h(\tilde{x}^*)(u,u)$. Lemma 6, together with the assumption that $\{\tilde{X}_i\}_{i=1}^n \subset \tilde{S}$, ensures that all the non-trivial critical points of $\tilde{p}_h(x)$ belong to $\tilde{S}$. Equation (A.31) of Appendix A shows that for any $v \in (\tilde{S})^{\perp}$,

$$D^2\tilde{p}_h(x)(v,v) = \frac{2}{n}\sum_{i=1}^{n} K_h'(\|\tilde{X}_i - x\|_{L^2}^2)\|v\|_{L^2}^2 < 0. \tag{6.12}$$

Therefore, in analogy to the results of Section 4, in order to classify the significant local modes of $\tilde{p}_h$ it is not required to consider infinitely many L^2 basis functions.

Remark 9. The assumptions of Example 1 do not immediately guarantee that $\{\tilde{X}_i\}_{i=1}^n \subset B_{H_0^1}(0,M)$ as we assume in the second claim of Corollary 1. However, simple computations show that

$$P\left(\tilde{X}_i \in B_{H_0^1}(0,M) \quad \forall i \in \{1,\dots,n\}\right) \geq 1 - Cnm^{-\frac{2}{5}} \tag{6.13}$$

for some positive constant C. Therefore, as long as $n = o(m^{\frac{2}{5}})$, the consistency result (the second claim) of Corollary 1 still holds. Projection using regression estimators introduces some bias to the pseudo density. However, as long as the observation grid is sufficiently fine (m is large enough), then the bias is of lower order. For example, if $n = o(m^{2/5})$, then $\phi(m) = O(m^{-4/5}) = o(n^{-2})$, which is of lower order in (6.10). Thus, in this case, one can simply focus on the projected version of the observed functional data.

7. Simulations and applications

In this section, we apply the methodology that we discussed on two simulated functional datasets and on a real dataset. In the following examples, we describe two practical ways to select the bandwidth and we show that both lead to meaningful estimated clusters. We use the mean-shift algorithm (Fukunaga and Hostetler, 1975; Cheng, 1995) with the exponential kernel to identify the local modes and cluster the data.

The mean-shift algorithm is a recursive algorithm that is equivalent to gradient ascent with a particular choice of adaptive step-size. In particular, the mean-shift algorithm approximates the gradient ascent path of $\hat{p}_h$ starting at a point $x = x_0$ by means of the recursive update

$$x \leftarrow \frac{\sum_{i=1}^n K\left(\frac{\|X_i - x\|_{L^2}^2}{h}\right) X_i}{\sum_{i=1}^n K\left(\frac{\|X_i - x\|_{L^2}^2}{h}\right)}. \tag{7.1}$$

In practice, one typically takes $x_0 \in \{X_1,\dots,X_n\}$.

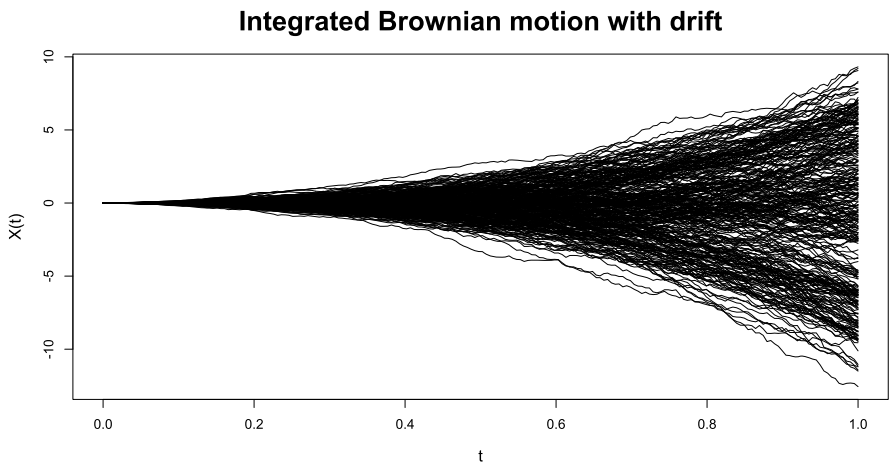

FIG 1. *Integrated Brownian motion trajectories with drift.*

In what follows, the constants $C_1(\alpha)$ and $C_2(\alpha)$ of the algorithm of Display 1 are determined by using the nonparametric Bootstrap. Recall that $C_1(\alpha)$ and $C_2(\alpha)$ are simply the $1-\alpha$ quantiles of the random variables η_1 and η_2 of equation (5.3). The expressions of $C_1(\alpha)$ and $C_2(\alpha)$ reported in Display 1 allow us to deduce the convergence rate of the procedure, but they represent rather conservative bounds. In practice, it is advisable to determine this quantities in a data-driven way (for example, use non-parametric bootstrap to find the quantiles of equation (5.3)).

7.1. Integrated Brownian motion

DATA Let W_i denote a realization of the standard Brownian motion process on $t \in [0,1]$. We generate i.i.d. trajectories of the integrated Brownian motion process with drift

$$X_i(t) = \sigma \int_0^t W_i(s)\, ds + b_i t^2, \tag{7.2}$$

where $\sigma = 2$ and

$$b_i = \begin{cases} 0 \text{ for } i = 1, \dots, 100 \\ 5 \text{ for } i = 101, \dots, 200 \\ -7 \text{ for } i = 201, \dots, 300. \end{cases} \tag{7.3}$$

The curves above are generated on a grid of 200 equally-spaced points in $[0,1]$ and they are displayed in Figure 1. Because of the variability of the process and the noise level σ, it is hard to distinguish the presence of the three distinct clusters in the observed trajectories.

PROCEDURE We run the mean-shift algorithm using the exponential kernel and a set of candidate bandwidths corresponding to the $10, 15, 20, 25, \dots 100$

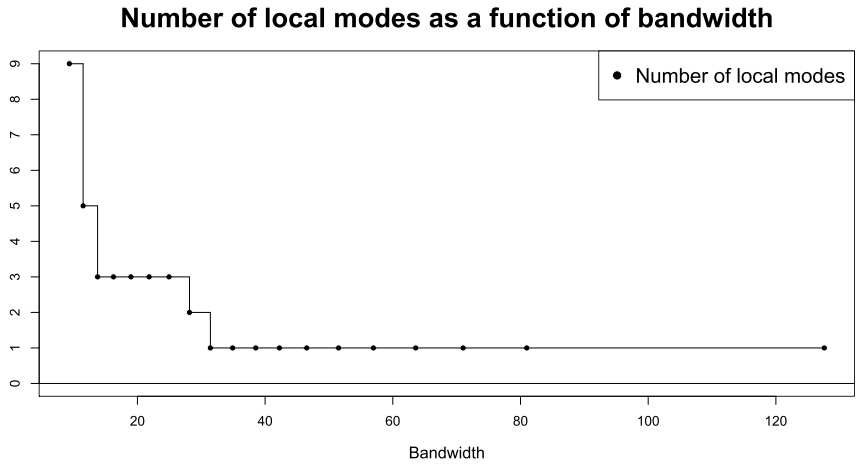

FIG 2. *Number of estimated local modes as a function of the bandwidth in the integrated Brownian motion example.*

percentiles of the distribution of all pairwise L^2 distances between the trajectories $X_1, \ldots, X_{300}$.

RESULTS Figure 2 depicts how the number of estimated local modes (hence the number of estimated clusters) varies as the bandwidth increases: when the bandwidth is too small (*overclustering*), $\hat{p}_h$ has a large numbers of noninformative local modes, whereas when the bandwidth is too large (*underclustering*) all the data points are eventually merged in a unique cluster corresponding to the unique mode of $\hat{p}_h$.

Frequently, one observes a relatively large intermediate range of bandwidths where the clustering structure is stable. In Figure 2, we notice the presence of such range. There, the number of clusters is stable and equal to 3. This gives a heuristic to choose h in practice.

Figure 3, depicts the estimated clustering structure when h is chosen within the aforementioned range. Note that the estimated local modes (continuous light blue lines) are very closed to the quadratic drift functions of equations (7.2) and (7.3) (broken light blue lines).

7.2. Two-dimensional linear space with added noise

DATA We generate 2000 i.i.d. pairs of coefficients (α_i, β_i) from the bivariate Normal mixture

$$\frac{1}{2}\mathcal{N}\left(\begin{pmatrix}0\\0.2\end{pmatrix}, \begin{pmatrix}0.03^2 & 0\\0 & 0.03^2\end{pmatrix}\right) + \frac{1}{2}\mathcal{N}\left(\begin{pmatrix}0\\-0.2\end{pmatrix}, \begin{pmatrix}0.03^2 & 0\\0 & 0.03^2\end{pmatrix}\right). \quad (7.4)$$

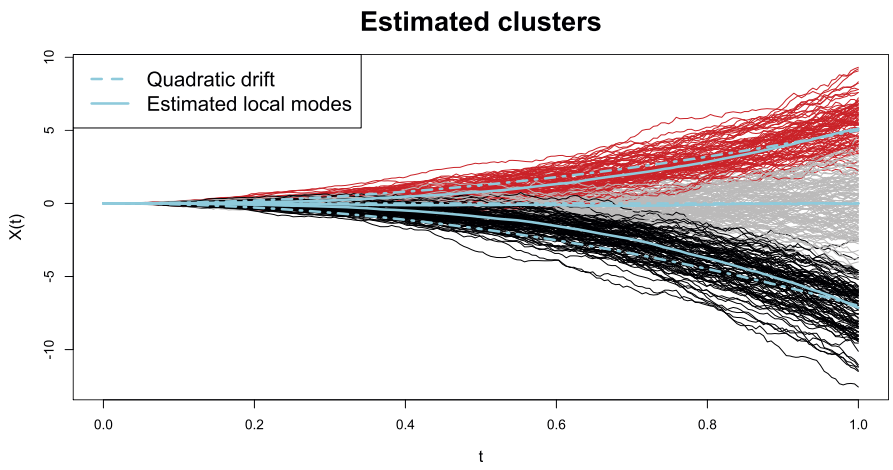

FIG 3. *Estimated local modes (continuous light blue lines) and estimated clusters (red, grey, and black) in the integrated Brownian motion example. The broken lines (quadratic drifts) correspond to the true local modes.*

These random coefficients are used to generate noiseless functional data according to

$$X_i(t) = \alpha_i \frac{1}{\sqrt{2}} \sin(8t\pi) + \beta_i \frac{1}{\sqrt{2}} \sin((8t-1)\pi). \tag{7.5}$$

Noisy measurements of these curves are then obtained from the model

$$Y_i(t_j) = X_i(t_j) + \epsilon_{i,j}, \tag{7.6}$$

where $\{t_j, j = 1, \ldots, 500\}$, are equally-spaced points in $[0, 1]$ and $\epsilon_{i,j}$ are i.i.d. draws from $\mathcal{N}(0, 0.5^2)$.

Figure 4 depicts two noisy observations of these curves while Figure 5 is a pairs scatterplot representing the coefficients of the projection of the Y_i's on an 8-dimensional cosine orthonormal basis. Because of the very low signal-to-noise ratio of the Y_i's, neither of these plots clearly provides conclusive evidence about the presence of two distinct clusters.

PROCEDURE The goal is once again to choose h and recover the hidden clustering structure. We proceed as follows.

1. For all $i = 1, \ldots, 2000$, we obtain $\tilde{X}_i$, an estimate of X_i, by using an orthogonal projection estimator on the canonical cosine orthonormal basis. We use Generalized Cross Validation to optimally pick the number of basis functions (which is 8 in this case).
2. We randomly split the set of $\tilde{X}_i$'s in two subsamples of size 1000 each.
3. We form a set of candidate values for h.

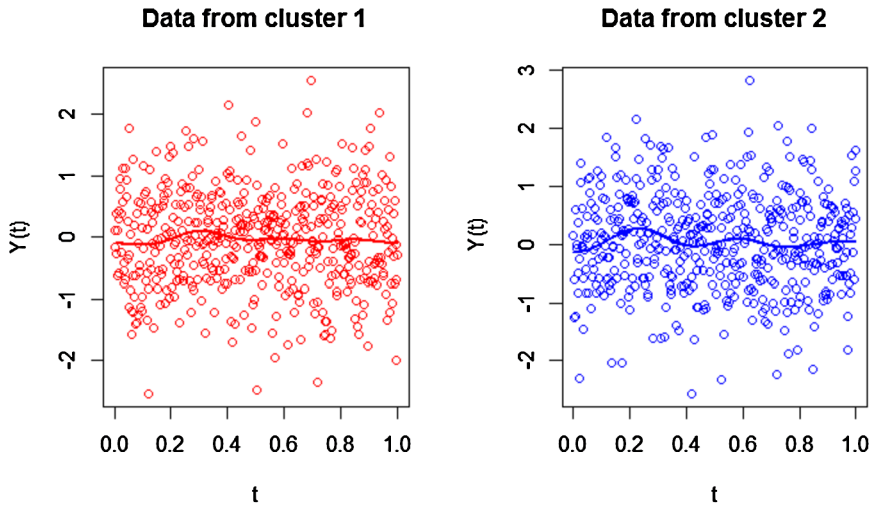

FIG 4. *Examples of noisy functional data in the two-dimensional linear space example. The hollow dots represent the noisy observations, i.e. the Y_i's. The continuous lines represent the orthogonal projection estimate $\tilde{X}_i$ of the corresponding noiseless functional datum X_i.*

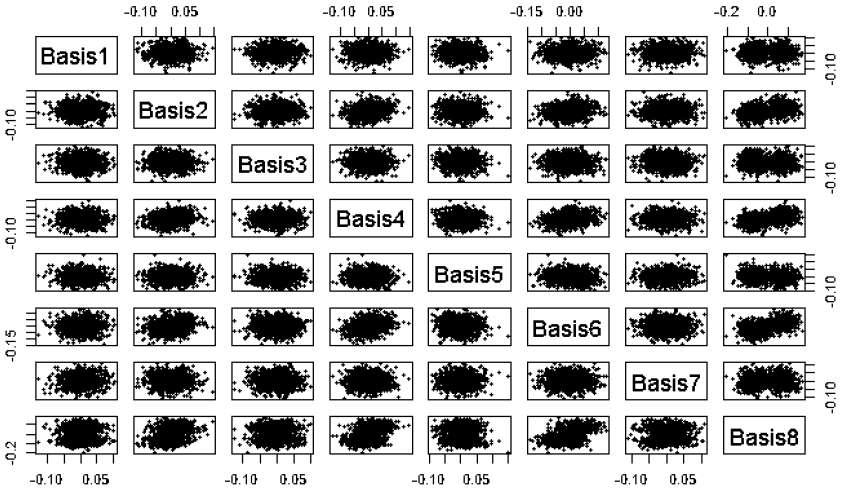

FIG 5. *Pairs plot of the orthogonal projection coefficients.*

4. For each candidate value of h, the first subsample of $\tilde{X}_i$'s is used to estimate the functional modes and the corresponding clusters.
5. For each estimated clustering structure, the second subsample of $\tilde{X}_i$'s is used to test for the clusters' significance ($\alpha = 0.10$).

Thanks to the availability of a test for the significance of the local modes, we have a principled way of selecting the bandwidth and an alternative to the heuristic

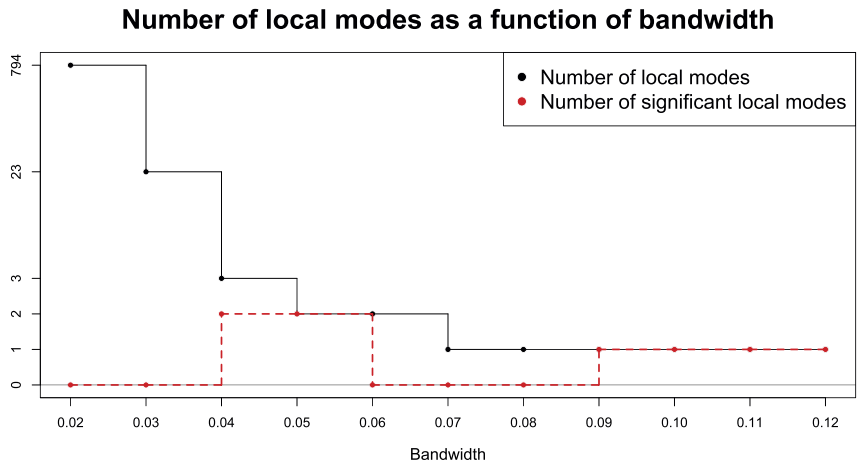

FIG 6. *Number of estimated local modes and number of significant estimated local modes as a function of the bandwidth in the two-dimensional linear space example.*

proposed in the previous application. In particular, h should be chosen as the bandwidth which reveals the largest amount of statistically significant structure in the data, i.e. as the bandwidth that maximizes the number of significant clusters.

RESULTS The final result of this procedure is summarized in Figure 6. In this case, the final bandwidth is $h = 0.05$ and, despite the low signal-to-noise ratio of the observed data, we detect the presence of two distinct significant clusters.

Note once again that, for too large values of the bandwidth, we observe a single cluster corresponding to the entire sample.

7.3. Neural activity curves

Figure 7 100 out of 1000 curves which correspond to recordings of neural activity at 32 equally-spaced time points. These data come from a behavioral experiment performed at the Andrew Schwartz's *Motorlab* (University of Pittsburgh) on a macaque monkey. The monkey performs a center-out and out-center target-reaching task with 26 targets in a virtual 3D environment and these curves represent voltage changes over time in the monkey's neurons.

It is known that in this dataset there are three distinct clusters (see Taylor, Tillery Helms and Schwartz, 2002 for more details). In this example, we show that our procedure allows us to recover the true clustering structure, even if we did not have any a priori information about the true number of clusters.

We follow the same strategy outlined in the previous example regarding the two-dimensional linear space spanned by the sine functions. While we only show 100 curves in Figure 7, the clustering procedure is run on the entire

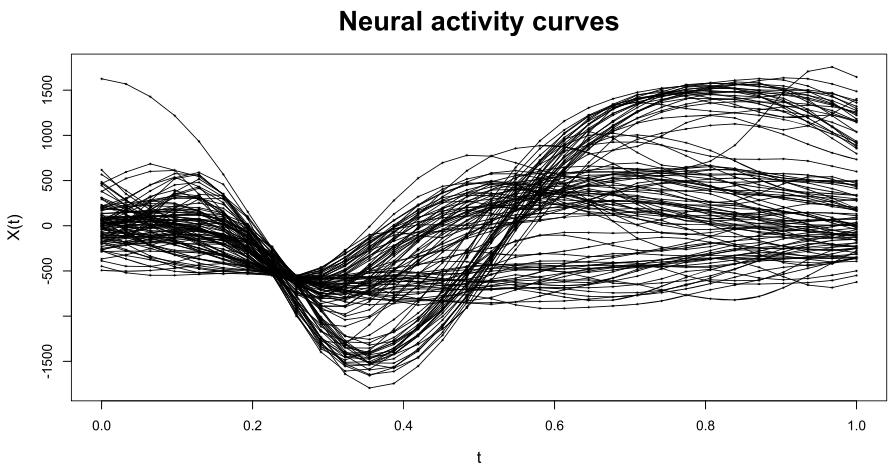

FIG 7. *A subset of 100 curves representing neural activity over time.*

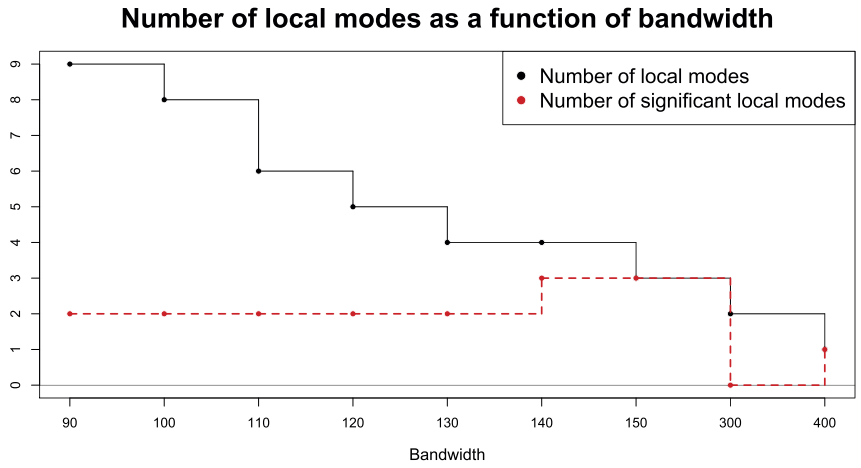

FIG 8. *Number of estimated local modes and number of significant estimated local modes as a function of the bandwidth in the neural activity example.*

dataset (1000 curves). Figure 8 displays the number of clusters and the number of significant clusters as the bandwidth is varied. Once again, the bandwidth is chosen to be the value which maximizes the number of significant local modes ($\alpha = 0.10$). In this case, this corresponds to $h = 140$ or $h = 150$, both of which produce three significant clusters. Finally, Figure 9 displays the same subset of curves of Figure 7, this time colored by the cluster membership, together with the three local modes associated to the clusters.

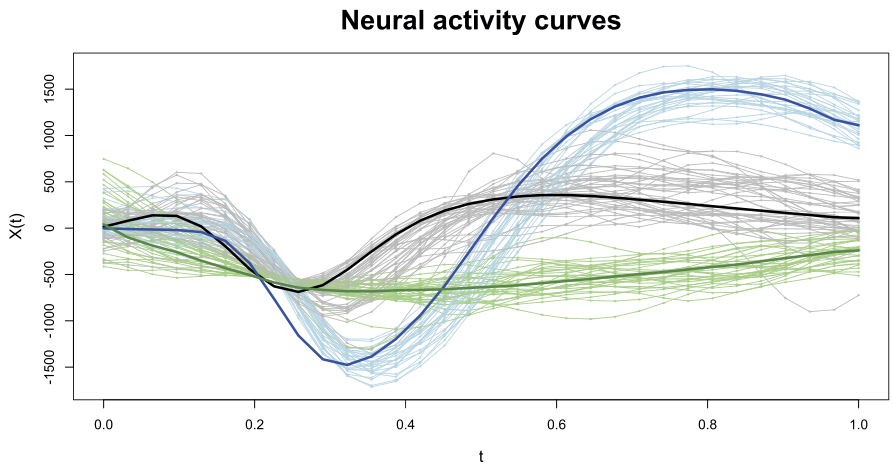

FIG 9. *Estimated local modes and clusters in the neural activity curves (only a subset of 100 curves shown). The estimated local modes correspond to the three thicker lines.*

8. On the choice of the pseudo-density functional

It is well-known that the Lebesgue measure does not exist in infinite-dimensional spaces. As a consequence, a proper density function for a functional random variable cannot generally be defined (Delaigle and Hall, 2010). Developing a theory of modal clustering for functional data necessarily requires a choice of a surrogate notion of density that substitutes the probability density function associated with the data. A pseudo-density should satisfy some basic differentiability properties, so that one can study the associated gradient flow. While we explicitly choose to use the functional p_h of equation (3.1), one could in principle work with a different functional. The choice of the pseudo-density is not an easy one, however.

First of all, with particular emphasis on the setting that we consider, we point out that, while tempting, one cannot naively assume that the pseudo-density is L^2 differentiable (or even just continuous) and also vanishes as the H_0^1 norm diverges.

Fact 1. *Let $p : L^2([0,1]) \to \mathbb{R}_+$ be a pseudo-density for the functional random variable X valued in $H_0^1([0,1])$ such that p is L^2 continuous and $p(x) \to 0$ as $\|x\|_{H_0^1} \to \infty$. Then $p = 0$ everywhere on $H_0^1([0,1])$.*

Proof. Consider the sequence of functions $x_n(t) = n^{-1}\sin(n^2 t)$ for $n \geq 1$ and $t \in [0,1]$. Clearly, $x_n \in H_0^1([0,1]$ for any $n \geq 1$. Notice that $\|x_n\|_{L^2} \to 0$ and $\|x_n\|_{H_0^1} \to \infty$ as $n \to \infty$. Thus, by assumption, $p(x_n) \to p(0) = 0$ as $n \to \infty$. Consider now $z_n = y + x_n$ where $y \in H_0^1([0,1])$. We have $\|z_n\|_{L^2} \to \|y\|_{L^2}$ as $n \to \infty$, hence $p(z_n) \to p(y)$ as $n \to \infty$. However $\|z_n\|_{H_0^1} \geq \|x_n\|_{H_0^1} - \|y\|_{H_0^1}$. Hence $\|z_n\|_{H_0^1} \to \infty$ as $n \to \infty$. We thus have $p(z_n) \to p(y) = 0$ as $n \to \infty$ for any $y \in H_0^1([0,1])$, implying that p is null on $H_0^1([0,1])$. □

The argument above shows that requiring a pseudo-density to be L^2 continuous and to vanish outside of H_0^1 necessarily leads to an uninteresting scenario for modal clustering, despite the fact that these two requirements apparently sound reasonable at first and carry some resemblance with the standard assumptions that are made on density functions in finite-dimensional problems.

Secondly, analyzing the asymptotic regime where $h = h_n \to 0$ makes little sense even in the most well-understood situations. In fact, let us consider the following two settings in which one typically chooses $h = h_n \to 0$ as $n \to \infty$.

1. If the law P of X is supported on a finite-dimensional space and admits a density p, then the bias of $\hat{p}_h$ is easy to compute and one can choose $h_n \to 0$ to optimally balance the bias-variance trade-off as it is usually done in density estimation. However, if p is defined over a finite-dimensional vector space S, the gradient flow is not well-defined outside of S. As discussed in Section 6, all of the observed functional data are generally reconstructed from noisy discrete measurements. As a result, neither the observed discrete measurements nor the reconstructed functional data are in S, and the gradient flow with respect to p starting at any of these elements is not well-defined.
2. If the law of P of X is supported on an infinite-dimensional space (for example X is a diffusion process), then some authors (see, for instance, Gasser, Hall and Presnell, 1998 and Ferraty, Kudraszow and Vieu, 2012) suggest to implicitly define a pseudo-density by assuming a particular factorization of the small-probability function associated to P. In particular, it is assumed that

$$P(\|X - x\| \leq h) = p(x)\phi(h) + o(\phi(h)) \tag{8.1}$$

as $h \to 0$ for some pseudo-density functional p depending only on the center of the ball x and some function ϕ depending only on the radius of the ball h. In this second case, p is non-zero only on its domain S_1, which is typically taken to be a compact subset of the infinite-dimensional functional space. Compact subsets of L^2 are singular in the sense that any L^2 closed ball is not compact. As a result, the pseudo-density p is also singular, hence not L^2 differentiable. One might then assume that p is differentiable with respect to norm induced by S_1 and study the gradient flow associated to p using the S_1 topology. In light of the first point just given, one should then assume that S_1 is the closure of an open set of an infinite-dimensional functional subspace. This leads to an even more serious problem: closed bounded balls in S_1 are not compact under the S_1 topology. The lack of compactness implies that the gradient ascent paths are not guaranteed to converge.

On the other hand, the pseudo-density functional $p_h(x) = E_P K(\frac{\|X-x\|_{L^2}^2}{h})$ with $h > 0$ is a natural candidate to develop a theory of modal clustering of smooth random curves in a density-free setting. Furthermore, the functional p_h corresponds to the functional discussed by Hall and Heckman (2002), who

proposed a mode-finding algorithm for functional data and had the intuition that their algorithm was approximating a gradient flow on the estimator $\hat{p}_h$.

Of course, from a practical point of view, one still has to choose a value for h. The next section describes two bandwidth selection strategies.

9. Bandwidth selection

The choice of the bandwidth is generally a difficult task in nonparametric problems. Furthermore, the difficulty increases as the dimension of the model increases (see for instance Scott, 2015).

For multivariate data, the behavior of the topological structure of $\hat{p}$ (the estimator of the underlying density function) often exhibits a phase-transition: for small values of h (*undersmoothing*) the estimated density generates many irrelevant clusters (*overclustering*), while for large values of h (*undersmoothing*) $\hat{p}$ generates few uninformative clusters (*underclustering*) which eventually merge into a single cluster once h becomes large enough. Interestingly, one can usually identify a relatively broad range of intermediate values of h for which the number of clusters associated to $\hat{p}$ is stable (see for instance, Genovese et al., 2016).

As we illustrate in Section 7, $\hat{p}_h$ tends to behave in a very similar way when h varies (see also Figure 2). This leads to a simple yet effective heuristic for bandwidth selection: h should be chosen in the range where the clustering structure is stable.

If a test for the significance of the estimated local modes of $\hat{p}_h$ is available, then one can use a complementary and more principled approach: h should be chosen so to reveal the largest amount of statistically significant structure in the data (Genovese et al., 2016). In particular, in the context of nonparametric modal clustering for functional data, h should be chosen to maximize the number of significant local modes of $\hat{p}_h$ using, for example, the test proposed in Section 5. This second approach to bandwidth selection is illustrated by means of two applications in Section 7 (see Figure 6 and Figure 9).

10. Discussion and conclusions

In this paper, we provide a general theoretical background for clustering of functional data based on pseudo-densities. We show that clusters of functional data can be characterized in terms of the basins of attraction of the critical points of a pseudo-density functional, both at the population and at the sample level. Our theory can be generalized to different functional spaces, as long as the chosen pseudo-density functional is sufficiently smooth and the range of the functional random variable X can be compactly embedded in a larger space to guarantee the compactness of the gradient flow trajectories. Because of the need of a compact embedding, one has to consider two non-equivalent topologies at the same time (in our case the L^2 and the H^1 topologies): from a statistical viewpoint, this means that the data need to be smoother than the ambient space in which they are embedded.

Besides compactness, there is another element that makes the theory of population clustering in the functional data setting more challenging when compared to the finite-dimensional case. This is the fact that the basin of attraction of a saddle point of p_h is not necessarily negligible. While in the finite-dimensional setting the basin of attraction of a saddle point of the Morse density function p is a manifold whose dimension is strictly smaller than the dimension of the domain of p (and therefore its probability content is null), the same property is not necessarily satisfied by a pseudo-density functional in the infinite-dimensional and density-free setting that we consider. Nonetheless, in analogy to the finite-dimensional case, one expects that clusters that are associated to the local modes of p_h are more informative than those associated to the saddle points of the same functional.

It becomes natural to ask whether it is possible to derive a statistical procedure that marks a local mode of $\hat{p}_h$ (and its associated empirical cluster) as significant whenever it corresponds to a non-degenerate local mode of p_h. We provide a consistent algorithm to achieve this task that can be applied to real data, such as noisy measurements of random curves on a grid. When only noisy and discretized versions of the functional data are observed, one typically projects the noisy observations onto a linear space and use the resulting projections (which correspond to perturbed versions of the partially observed underlying functional data) for subsequent statistical analyses. We show that the asymptotic properties of the proposed algorithm are preserved with noisy and discretized functional data, as long as the size of the discretization grid grows suitably fast with the sample size.

The algorithm that we propose can also be used to appropriately select the bandwidth parameter in practice: h should be chosen as the bandwidth that maximizes the number of significant local modes (and therefore the number of significant clusters) in the data. We demonstrate the this bandwidth selection criterion performs well both on simulated and real data in Section 7. The logic behind this rule is that one wants to uncover as much structure as possible based on the observed data, while at the same time being confident that the uncovered structure corresponds to actual population features.

Finally, it should be mentioned that kNN (k-Nearest Neighbors) bandwidth parameters have gained a growing popularity in nonparametric FDA. kNN bandwidths have the advantage of being more sensitive to local features of the empirical distribution of the data. The kNN bandwidth at location $x \in L^2$ is defined as

$$H_{n,k}(x) = \min \left\{ h > 0 : \sum_{i=1}^{n} \mathbb{1}_{B_{L^2}(x,h)}(X_i) = k \right\}, \tag{10.1}$$

i.e. $H_{n,k}(x)$ is the smallest positive real number h such that the L^2 ball of radius h centered at $x \in L^2$ contains exactly k sample observations (see, for instance, Kudraszow and Vieu, 2013). It is clear from the definition that $H_{n,k}$ is a random variable that depends on the sample and the sample size. Extending our theory to kNN bandwidths is a challenging task because the population functional becomes much more involved. Specifically, the population pseudo-

density corresponding to a kNN bandwidth is

$$p_{H_n,k}(x) = E_P\, K\left(\frac{\|X-x\|_{L^2}^2}{H_{n,k}(x)}\right). \tag{10.2}$$

Future work shall investigate the extent to which our theory can be generalized to these types of local bandwidths and the practical advantages that these may have in the context of nonparametric clustering based on local modes and gradient flows.

Appendix A: Proofs of the results

Proof of Proposition 1. For convenience, we prove the result under the additional assumption that K is compactly supported on $[0,1]$. The proof that we present can be easily extended to exponentially decaying kernel functions and this extra assumption can be safely removed.

Consider the set of assumptions of the Proposition. Furthermore, let $\bar{K}_2 = \sup_{x\in\mathbb{R}^d} \|\nabla^2 p(x)\|_2$ and $\underline{K}_1 = \inf_{x\in\partial S_c} \|\nabla p(x)\|_2$. Note that since ∂S_c is compact, $\underline{K}_1 > 0$. Consider now the set $S_\epsilon = \{x \in S_c : d(x, \partial S_c) \geq \epsilon\}$. Then, there exists a set Ω such that $S_{2\epsilon} \subset \Omega \subset S_\epsilon$ and $\partial\Omega$ is also smooth. As a result, if $x \in \partial\Omega$, then $\epsilon \leq d(x,\partial S) \leq 2\epsilon$ and $\inf_{x\in S\cap\Omega^c} \|\nabla p(x)\|_2 \geq \underline{K}_1/3$. Since p is Morse on Ω and twice continuously differentiable on $\mathrm{int}(S_c)$, then standard mollification results guarantee that there exist $\eta > 0$ and $h_1 > 0$ such that if $0 < h \leq h_1$, then $\sup_{x\in\Omega} \|\nabla^{(i)} p_h(x) - \nabla^{(i)} p(x)\|_2 \leq \eta$ for $i = 0,1,2$. Then, Lemma 16 of Chazal et al. (2014) guarantees that p_h is Morse on Ω. It is only left to show that if $x \in \Omega^c$ is such that $\mathcal{L}(B_{\mathbb{R}^d}(x,h) \cap S_c) > 0$ then $\nabla p_h \neq 0$. Consider $h < \frac{\underline{K}_1}{6\bar{K}_2}$ and let $n(\cdot)$ denote the outward normal vector to S_c with unitary norm. We have

$$\begin{aligned}
\nabla p_h(x) &= \nabla_x \int_{\mathbb{R}^d} K\left(\frac{\|y-x\|_2^2}{h}\right) p(y)\, dy \\
&= \int_{S_c\cap B_{\mathbb{R}^d}(x,h)} \nabla_x K\left(\frac{\|y-x\|_2^2}{h}\right) p(y)\, dy \\
&= \int_{S_c\cap B_{\mathbb{R}^d}(x,h)} -\nabla_y K\left(\frac{\|y-x\|_2^2}{h}\right) p(y)\, dy \\
&= \int_{S_c\cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) \nabla_y p(y)\, dy \\
&\quad - \int_{\partial S_c\cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) n(y) p(y)\, dy \\
&\quad - \int_{S_c\cap \partial B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) n(y) p(y)\, dy.
\end{aligned} \tag{A.1}$$

Note that $p(y) = 0$ if $y \in \partial S_c$ and, since K is compactly supported on $[0,1]$,

$K(\frac{\|y-x\|_{L^2}^2}{h}) = 0$ if $y \in \partial B_{\mathbb{R}^d}(x,h)$. Hence, the last two integrals on the boundaries are null. Now, since ∇p is $\bar{K}_2$-Lipschitz, we have

$$\begin{aligned}
\|\nabla p_h(x)\|_2 &= \left\| \int_{S_c \cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) \nabla p(y)\, dy \right\|_2 \\
&\geq -\left\| \int_{S_c \cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) \nabla p(y)\, dy \right. \\
&\quad \left. - \int_{S_c \cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) \nabla p(x)\, dy \right\|_2 \\
&\quad + \left\| \int_{S_c \cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) \nabla p(x)\, dy \right\|_2 \\
&\geq -\bar{K}_2 h \int_{S_c \cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) dy \\
&\quad + \|\nabla p(x)\|_2 \int_{S_c \cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) dy \\
&\geq \left(\frac{\underline{K}_1}{3} - \bar{K}_2 h\right) \int_{S_c \cap B_{\mathbb{R}^d}(x,h)} K\left(\frac{\|y-x\|_2^2}{h}\right) dy > 0
\end{aligned} \tag{A.2}$$

since $-\bar{K}_2 h > -\underline{K}_1/6$ and $\mathcal{L}(B_{\mathbb{R}^d}(x,h) \cap S_c) > 0$. This shows that p_h has no non-trivial critical points outside of Ω. □

Proof of Lemma 1.

$$\begin{aligned}
&|K_h(\|X-(x+\delta)\|_{L^2}^2) - K_h(\|X-x\|_{L^2}^2) \\
&- \langle DK_h(\|X-x\|_{L^2}^2), \delta\rangle_{L^2}| \\
&\leq \sup_{s\in[0,1]} \frac{1}{2} |D^2 K_h(\|X-(x+s\delta)\|_{L^2}^2)(\delta,\delta)|
\end{aligned} \tag{A.3}$$

Now, using the bounds on the derivatives of K_h and equation (3.3), we have

$$\begin{aligned}
&|D^2 K_h(\|X-(x+s\delta)\|_{L^2}^2)(\delta,\delta)| \\
&\leq 4\left|K_h''(\|X-(x+s\delta)\|_{L^2}^2)\right| \|X-(x+s\delta)\|_{L^2}^2 \|\delta\|_{L^2}^2 \\
&+ 2\left|K_h'(\|X-(x+s\delta)\|_{L^2}^2)\right| \|\delta\|_{L^2}^2 \\
&\leq 4K_2 \|\delta\|_{L^2}^2 = o(\|\delta\|_{L^2}).
\end{aligned} \tag{A.4}$$

Taking the expectation and applying Jensen's inequality in equation (A.3) yields

$$\begin{aligned}
&|E_P\, K_h(\|X-(x+\delta)\|_{L^2}^2)| - E_P\, K_h(\|X-x\|_{L^2}^2) \\
&- E_P\, \langle DK_h(\|X-x\|_{L^2}^2), \delta\rangle_{L^2}| = o(\|\delta\|_{L^2})
\end{aligned} \tag{A.5}$$

which implies that

$$\begin{aligned}\langle Dp_h(x), \cdot \rangle_{L^2} &= E_P \langle DK_h(\|X - x\|_{L^2}^2, \cdot \rangle_{L^2} \\ &= E_P \langle 2K_h' \left(\|X - x\|_{L^2}^2\right)(x - X), \cdot \rangle_{L^2} \\ &= \langle E_P\, 2K_h' \left(\|X - x\|_{L^2}^2\right)(x - X), \cdot \rangle_{L^2}.\end{aligned} \tag{A.6}$$

Thus, by definition, equation (3.5) is established. It is clear from assumption (H1) that $\|Dp_h(x)\|_{L^2} \leq 2K_1$. In order to derive $D^2 p_h(x)$, a similar computation is used. The Taylor expansion of $F(x) = K_h'(\|X - x\|_{L^2}^2)(x - X)$ as a function of x gives

$$\|F(x + \delta) - F(x) - DF(x)(\delta)\|_{L^2} \leq \sup_{s \in [0,1]} \frac{1}{2} \|D^2 F(x + s\delta)(\delta, \delta)\|_{L^2} \tag{A.7}$$

where

$$\begin{aligned}DF(x)(\delta) = {} & 2K_h'' \left(\|X - x\|_{L_2}^2\right) \langle x - X, \delta \rangle_{L_2} (x - X) \\ & + K_h' \left(\|X - x\|_{L_2}^2\right) \delta.\end{aligned} \tag{A.8}$$

Furthermore,

$$\begin{aligned}&D^2 F(x + s\delta)(\delta_1, \delta_2) \\ &= 4K_h''' \left(\|X - x\|_{L_2}^2\right) \langle x - X, \delta_1 \rangle_{L_2} \langle x - X, \delta_2 \rangle_{L_2} (x - X) \\ &+ 2K_h'' \left(\|X - x\|_{L_2}^2\right) \langle \delta_1, \delta_2 \rangle_{L_2} (x - X) \\ &+ 2K_h'' \left(\|X - x\|_{L_2}^2\right) \langle x - X, \delta_1 \rangle_{L_2} \delta_2 \\ &+ 2K_h'' \left(\|X - x\|_{L_2}^2\right) \langle x - X, \delta_2 \rangle_{L_2} \delta_1\end{aligned} \tag{A.9}$$

By assumption (H1), $\sup_{s \in [0,1]} \|D^2 F(x + s\delta)(\delta, \delta)\|_{L^2} \leq 6K_3 \|\delta\|_{L^2}^2$. Thus,

$$\begin{aligned}&\|E_P F(x + \delta) - E_P F(x) - E_P\, DF(x)(\delta)\|_{L^2} \\ &\leq E_P \|F(x + \delta) - F(x) - DF(x)(\delta)\|_{L^2} \leq 3K_3 \|\delta\|_{L^2}^2,\end{aligned} \tag{A.10}$$

and the claim then easily follows. □

Proof of Proposition 2.

$$\begin{aligned}&\langle Dp_h(x), v' \rangle_{L^2} = \left\langle E_P\, 2K_h' \left(\|X - x\|_{L^2}^2\right)(x - X), v' \right\rangle_{L^2} \\ &= E_P\, 2K_h' \left(\|X - x\|_{L^2}^2\right) \langle x - X, v' \rangle_{L^2} \\ &= E_P\ -2K_h' \left(\|X - x\|_{L^2}^2\right) \langle x' - X', v \rangle_{L^2} \\ &\leq E_P\ -2K_h' \left(\|X - x\|_{L^2}^2\right) \|x' - X'\|_{L^2} \|v\|_{L^2} \\ &\leq 2\, K_2 (\|x'\|_{L^2} + E_P \|X'\|_{L^2}) \|v\|_{L^2} \\ &\leq 2K_2 (\|x'\|_{L^2} + N_1) \|v\|_{L^2}\end{aligned} \tag{A.11}$$

where the second equality holds by integration by parts. An application of Lemma 2 with $L(v) = E_P\, -2K' \left(\|X - x\|_{L^2}^2\right) \langle x' - X', v \rangle_{L^2}$ yields $\|Dp_h(x)\|_{H_0^1} = \|Dp_h(x)'\|_{L^2} \leq 2K_1 (\|x'\|_{L^2} + N_1)$ and therefore $Dp_h(x) \in H_0^1$. □

Proof of Lemma 3. Let $\|x\|_{H_0^1}, \|y\|_{H_0^1} \leq L < \infty$. It suffices to show that $\exists 0 < C(L) < \infty$ such that $\|Dp_h(x) - Dp_h(y)\|_{H_0^1} \leq C(L)\|x - y\|_{H_0^1}$. Equivalently, one has to show that $\|Dp_h(x)' - Dp_h(y)'\|_{L^2} \leq C(L)\|x' - y'\|_{L^2}$. By Lemma 2 and Proposition 2 we have that, for any $v \in L^2$,

$$\begin{aligned}
&\langle Dp_h(x)' - Dp_h(y)', v\rangle_{L^2} \\
&= 2E_P\, K_h'\left(\|X - x\|_{L^2}^2\right)\langle x' - X', v\rangle_{L^2} \\
&- K_h'\left(\|X - y\|_{L^2}^2\right)\langle y' - X', v\rangle_{L^2} \\
&= 2E_P\left[K_h'\left(\|X - x\|_{L^2}^2\right) - K_h'\left(\|X - y\|_{L^2}^2\right)\right]\langle x' - X', v\rangle_{L^2} \\
&+ 2E_P\, K_h'\left(\|X - y\|_{L^2}^2\right)\langle x' - y', v\rangle_{L^2}
\end{aligned} \tag{A.12}$$

Since $\frac{d}{dt}K_h'(t^2) = 2K_h''(t^2)t \leq 2K_3$ by assumption (H1), $K_h'(t^2)$ is Lipschitz with Lipschitz constant not larger that $2K_3$. Therefore,

$$\begin{aligned}
&\left|K_h'\left(\|X - y\|_{L^2}^2\right) - K_h'\left(\|X - x\|_{L^2}^2\right)\right| \\
&\leq 2K_2\left|\|X - y\|_{L^2} - \|X - x\|_{L^2}\right| \\
&\leq 2K_2\|x - y\|_{L^2} \leq 2K_2C_p\|x' - y'\|_{L^2}.
\end{aligned} \tag{A.13}$$

We have

$$\begin{aligned}
&E_P\left[K_h'\left(\|X - y\|_{L^2}^2\right) - K_h'\left(\|X - x\|_{L^2}^2\right)\langle x' - X', v\rangle_{L^2}\right] \\
&\leq 2K_2C_p\|x' - y'\|_{L^2}\left(\|x'\|_{L^2} + E_P\|X'\|_{L^2}\right)\|v\|_{L^2} \\
&\leq 2K_2C_p\|x' - y'\|_{L^2}(L + N_1)\|v\|_{L^2}
\end{aligned} \tag{A.14}$$

and

$$E_P\, K_h'\left(\|X - y\|_{L^2}^2\right)\langle x' - y', v\rangle_{L^2} \leq K_2\|x' - y'\|_{L^2}\|v\|_{L^2}. \tag{A.15}$$

By putting together equations (A.14) and (A.15) we then have the following bound for equation (A.12):

$$\langle Dp_h(x)' - Dp_h(y)', v\rangle_{L^2} \leq C(L)\|x' - y'\|_{L^2}\|v\|_{L^2}, \tag{A.16}$$

with $C(L) = 2\left[K_3(L + N_1) + K_2\right]$, which obviously implies $\|Dp_h(x)' - Dp_h(y)'\|_{L^2} \leq C(L)\|x' - y'\|_{L^2}$. □

Proof of Lemma 4. Lemma 2 and Proposition 2 allow us to write

$$\begin{aligned}
&\langle Dp_h(\pi_x(t)), \pi_x(t)\rangle_{H_0^1} = \langle Dp_h(\pi_x(t))', \pi_x(t)'\rangle_{L^2} \\
&= 2E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right)\langle \pi_x(t)' - X', \pi_x(t)'\rangle_{L^2} \\
&= 2E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right)\left(\|\pi_x(t)'\|_{L^2}^2 - \langle X', \pi_x(t)'\rangle_{L^2}\right) \\
&\leq 2E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right)\|\pi_x(t)'\|_{L^2}^2 \\
&- 2E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right)\|X'\|_{L^2}\|\pi_x(t)'\|_{L^2},
\end{aligned} \tag{A.17}$$

where the last inequality follows because (H2) guarantees that $K_h'(t^2) \leq 0$.

For the first claim, assumption (H2) and $p_h(\pi_x(t)) \geq p_h(\pi_x(0)) \geq \delta$ imply

$$\begin{aligned} &E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right) \leq -E_P\, K_h\left(\|X - \pi_x(t)\|_{L^2}^2\right) \\ &= -p_h(\pi_x(t)) \leq -\delta. \end{aligned} \tag{A.18}$$

Thus, if $\|\pi_x(t)\|_{H_0^1} \geq K_2 N_1/\delta$,

$$\langle Dp_h(\pi_x(t)), \pi(t)\rangle_{H_0^1} \leq -2\delta\|\pi_x(t)\|_{H_0^1}^2 + 2K_2N_1\|\pi_x(t)\|_{H_0^1} \leq 0. \tag{A.19}$$

For the second part, equation (A.17) gives

$$\begin{aligned} &\langle Dp_h(\pi(t)), \pi(t)\rangle_{H_0^1} \\ &\leq 2E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right)\|\pi_x(t)'\|_{L^2}^2 \\ &- 2E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right)\|X'\|_{L^2}\|\pi_x(t)'\|_{L^2} \\ &\leq 2E_P\, K_h'\left(\|X - \pi_x(t)\|_{L^2}^2\right)\}\left(\|\pi_x(t)'\|_{L^2}^2 - M\|\pi_x(t)'\|_{L^2}\right). \end{aligned} \tag{A.20}$$

Thus, $\langle Dp_h(\pi_x(t)), \pi_x(t)\rangle_{H_0^1} \leq 0$ as soon as $\|\pi_x(t)'\|_{L^2} = \|\pi_x(t)\|_{H_0^1} > M$. □

Proof of Proposition 3. Proposition 2 and Lemma 3 guarantee the existence and uniqueness of a local solution under the H_0^1 norm from the standard theory of ordinary differential equations. Some extra work is needed to extend the local solution to a global one. We provide a complete proof in three steps which builds on Theorem 3.10 of Hunter and Nachtergaele (2001) (their Theorem 3.10 holds more generally on Banach spaces, see for instance Schechter, 2004) and the authors' subsequent remark concerning the extension of the local solution to a global one.

Step 1. In this step, we show that if the solution $\pi_x(t)$ exists for any time interval $[0, T]$, then there exists $C_1 > 0$ such that $\|\pi_x(t)\|_{H_0^1} \leq C_1$.
If $p_h(x) = 0$, then x is a trivial local minimum of $p_h(x)$. As a result, $Dp_h(\pi_x(0)) = 0$ and $\pi_x(t) = \pi_x(0)$ for all t. Thus, in this case it suffices to take $C_1 = R$. Suppose instead that $p_h(x) = \delta > 0$. Consider $g(t) = \|\pi_x(t)\|_{H_0^1}^2$. Clearly, $\frac{d}{dt}g(t) = 2\left\langle \pi_x(t), \frac{d}{dt}\pi_x(t)\right\rangle_{H_0^1} = 2\langle \pi_x(t), Dp_h(\pi_x(t))\rangle_{H_0^1}$. Note that $g(0) \leq R^2$. Take $C_1 = \max\{R, K_2N_1/\delta\}$. Fix an arbitrary $\epsilon > 0$ and suppose that there exists T' such that $0 \leq T' \leq T$ and $g(T') \geq C_1^2 + \epsilon$. Then, there must exist $0 \leq t^* \leq T'$ such that

$$g'(t^*) = 2\left\langle \pi_x(t^*), Dp_h(\pi_x(t^*))\right\rangle_{H_0^1} > 0 \tag{A.21}$$

and $g(t^*) \in (C_1^2, C_1^2 + \epsilon)$. This is a contradiction because, by Lemma 4, if $\|\pi_x(t^*)\|_{H_0^1} > K_2N_1/\delta$ then $\langle \pi_x(t^*), Dp_h(\pi_x(t^*))\rangle_{H_0^1} \leq 0$.

Step 2. Let $\pi_x : [0, T_1] \to H_0^1$ be the local solution of the ordinary differential equation $\pi_x'(t) = Dp_h(\pi_x(t))$ with $\pi_x(0) = x$. Suppose that $\|\pi_x(t)\|_{H_0^1} \leq C_1$ if $t \leq T_1$. Given $C_2 > C_1$, we show that there exists $T_2 > 0$ such that the solution can be uniquely extended to $\pi_x : [0, T_1 + T_2] \to H_0^1$ with $\|\pi_x(t)\|_{H_0^1} \leq C_2$ if $t \leq T_1 + T_2$. To see this, consider the ordinary

differential equation $\frac{d}{dt}\phi(t) = Dp_h(\phi(t))$ with $\phi(0) = \pi_x(T_1)$. Note that $\|\phi(0)\|_{H_0^1} = \|\pi_x(T_1)\|_{H_0^1} \leq C_1 < C_2$ by assumption. Also, let $N > 0$ be such that

$$\sup_{x \in B_{H_0^1}(0, M_2)} \|Dp_h(x)\|_{H_0^1} \leq N. \tag{A.22}$$

Now, by the Picard-Lindelöf theorem on Banach spaces, if one takes $T_2 = (C_2 - C_1)/N$ then the solution ϕ exists on $[0, T_2]$ and $\phi(t) \in B_{H_0^1}(\pi_x(T_1), C_2 - C_1)$. Consider the extension $\pi_x(t)$ given by

$$\pi_x(t) = \begin{cases} \pi_x(t) & \text{if } t \leq T_1 \\ \phi(t - T_1) & \text{if } T_1 \leq t \leq T_1 + T_2. \end{cases} \tag{A.23}$$

The newly defined π_x is well-defined and continuous. Since

$$\frac{d}{dt}\pi_x(t) = \frac{d}{dt}\phi(t - T_1) = Dp_h(\phi(t - T_1)) = Dp_h(\pi_x(t)) \tag{A.24}$$

if $t \in [T_1, T_1 + T_2]$, the new π_x is an extension of the solution. Furthermore, clearly $\pi_x(t) \in B_{H_0^1}(0, C_2)$ for $t \in [0, T_1 + T_2]$. The uniqueness of the extended solution follows from the fact that Dp_h is Lipschitz on $B_{H_0^1}(0, C_2)$.

Step 3. Since $\|x\|_{H_0^1} \leq R$, by Picard's theorem there exists a local solution $\pi_x(t) : [0, T_1] \to H_1^0$ and $\|\pi_x(t)\|_{H_0^1} \leq C_1$. Step 2 guarantees that the solution can be uniquely extended to $[0, T_1 + T_2]$. Step 1 then implies that such extended solution π_x satisfies $\|\pi_x(t)\|_{H_0^1} \leq C_1$ for all $t \in [0, T_1 + T_2]$. By Step 2 again, the extended solution π_x can be extended again to the larger time interval $[0, T_1 + 2T_2]$ and, once again, by Step 1 the extended solution is entirely contained in the H_0^1 ball of radius C_1. By iterating this procedure, one sees that the unique solution π_x can be extended to all of $\mathbb{R}_+$ and $\|\pi_x(t)\|_{H_0^1} \leq C_1$ for all $t \geq 0$. □

Proof of Theorem 1. Since p_h is a bounded functional and both Dp_h and $D^2 p_h$ are bounded operators on L^2, it is clear that

$$\lim_{t \to \infty} \|Dp_h(\pi_x(t))\|_{L^2} = 0 \tag{A.25}$$

(see Lemma 7.4.4 in Jost, 2011). Furthermore, since $\|\pi_x(t)\|_{H_0^1} \leq C_1$ for all $t \geq 0$ and closed H_0^1 balls are compact with respect to the L^2 norm, there exist $\{\pi(t_k)\}_{k=1}^{\infty}$ such that $\lim_{k \to \infty} \|\pi_x(t_k) - \pi_x(\infty)\|_{L^2} \to 0$ for some $\pi_x(\infty) \in L^2$. By the continuity of $Dp_h : L^2 \to L^2$, one also has that $Dp_h(\pi_x(\infty)) = 0$.

Recall that by assumption (H4), all the non-trivial critical points of p_h are isolated. Hence, for any non-trivial critical point of p_h, one can find a L^2 neighborhood around it in which there are no other critical points of p_h. Let $\delta_1 > 0$ be the radius of such neighborhood around $\pi_x(\infty)$. Suppose now that the sequence $\{\pi_x(t)\}_{t \geq 0}$ does not converge to $\pi_x(\infty)$ in the L^2 sense. Then, there exists $\delta_2 > 0$ and a subsequence $\{\pi_x(s_k)\}_{k \geq 1}$ such that $\|\pi_x(\infty) - \pi_x(s_k)\|_{L^2} \geq \delta_2$ for all $k \geq 1$.

Without loss of generality, one can assume that $\|\pi_x(t_k) - \pi_x(\infty)\|_{L^2} \leq \delta_1/3$ and that $t_k < s_k < t_{k+1}$ for all k. But then, by the continuity of the path π_x, there exists r_k such that $t_k \leq r_k \leq s_k$ and $\|\pi_x(\infty) - \pi_x(r_k)\|_{L^2} = \min\{\delta_1, \delta_2\}/2$ for all $k \geq 1$. Since $\|\pi_x(r_k)\|_{H_0^1} \leq C_1$, $\{\pi_x(r_k)\}_{k\geq 1}$ also has a subsequence which converges with respect to the L^2 norm as well. Without loss of generality assume that $\pi_x(r_k) \to \tilde{\pi}_x(\infty)$ in L^2 sense. By the continuity of $Dp_h(x)$, $\tilde{\pi}_x(\infty)$ is also a critical point of p_h. But then, $\|\pi_x(\infty) - \tilde{\pi}_x(\infty)\|_{L^2} = \min\{\delta_1, \delta_2\}/2 < \delta_1$, which is a contradiction. This establishes the uniqueness of $\pi_x(\infty)$ and concludes the proof. □

Proof of Lemma 5. By assumption, $Dp_h(x) = 2E_P\, K_h'(\|X - x\|_{L^2}^2)(x - X) = 0$ and $E_P\, K_h'(\|X - x\|_{L^2}^2) \leq -E_P\, K_h(\|X - x\|_{L^2}^2) = -p_h(x) < 0$. Thus,

$$x = \frac{E_P\, K_h'(\|X - x\|_{L^2}^2)X}{E_P\, K_h'(\|X - x\|_{L^2}^2)}. \tag{A.26}$$

Note that, by assumption (H2), $E_P\, K_h'(\|X - x\|_{L^2}^2) \leq -E_P\, K_h(\|X - x\|_{L^2}^2) < 0$. Therefore, it suffices to show that $E_P\, K_h'(\|X - x\|_{L^2}^2)X \in H_0^1$. We have

$$\begin{aligned} &\langle E_P\, K_h'(\|X - x\|_{L^2}^2)X, v'\rangle_{L^2} = E_P\, K_h'(\|X - x\|_{L^2}^2)\langle X, v'\rangle_{L^2} \\ &= E_P\, K_h'(\|X - x\|_{L^2}^2)\langle -X', v\rangle_{L^2} \leq K_2 N_1 \|v\|_{L^2}. \end{aligned} \tag{A.27}$$

Thus, $E_P\, K_h'(\|X - x\|_{L^2}^2)X \in H_0^1$ by Lemma 2. For the second claim of the Lemma, suppose that $\|X\|_{H_0^1} \leq M$ P-almost surely. Then, any x which is a non-trivial critical point of $p(x)$ satisfies equation (A.26). As a result, for any $v \in C_c^\infty([0,1])$,

$$\begin{aligned} \langle x, v'\rangle_{L^2} &= \frac{E_P K_h'(\|X - x\|_{L^2}^2)\langle X, v'\rangle_{L^2}}{E_P K_h'(\|X - x\|_{L^2}^2)} \\ &= \frac{E_P\, K_h'(\|X - x\|_{L^2}^2)\langle -X', v\rangle_{L^2}}{E_P\, K_h'(\|X - x\|_{L^2}^2)} \\ &\leq \frac{E_P\, K_h'(\|X - x\|_{L^2}^2)\|X\|_{H_0^1}\|v\|_{L^2}}{E_P\, K_h'(\|X - x\|_{L^2}^2)} \\ &\leq M\|v\|_{L^2}. \end{aligned} \tag{A.28}$$

By Lemma 2, it follows that $\|x\|_{H_0^1} \leq M$ and the proof is complete. □

Proof of Lemma 6. In light of Proposition 2, for the first claim it suffices to show that if $x \in S$, then $Dp_h(x) \in S$. Note that S is a closed subspace of L^2. As a result, there exists another subspace $S^\perp \subset L^2$ which is the orthogonal complement of S. Let $g \in S^\perp$, so that $\langle X, g\rangle_{L^2} = 0$ almost surely. Then,

$$\langle Dp_h(x), g\rangle_{L^2} = 2E_P\, K_h'(\|X - x\|_{L^2}^2)\langle x - X, g\rangle_{L^2} = 0, \tag{A.29}$$

and thus $Dp_h(x) \in S$. The second claim is established in a similar way as in Lemma 5. □

Proof of Lemma 7. By Lemma 6, if x^* is a non-trivial critical point then $x^* \in S$. If one views $D^2 p_h(x^*)$ as a linear operator from L^2 to L^2, it is sufficient to show that $D^2 p_h(x^*)$ is an isomorphism (i.e. a continuous map from L^2 to L^2 such that its inverse is also continuous). Note first that for any $v \in L^2$

$$\begin{aligned} D^2 p_h(x^*)(v) = E_P \left[4K_h''(\|X - x^*\|_{L^2}^2)\langle x^* - X, v\rangle_{L^2}(x^* - X) \right. \\ \left. + 2K_h'(\|X - x^*\|_{L^2}^2)v \right]. \end{aligned} \tag{A.30}$$

Observe that

1. If $v \in S$, then $D^2 p_h(x^*)(v) \in S$. One can use a similar computation as in equation (4.2) to show that $D^2 p_h(x^*)(v) = D^2 \tilde{p_h}(\tilde{x}^*)(\tilde{v})$, where $\tilde{v}$ is the vector in R^d corresponding to v.
2. Suppose $v \in S^\perp$. Since $\langle x^* - X, v\rangle_{L^2} = 0$ a.s., $D^2 p_h(x^*)(v) \in S^\perp$. More specifically,
$$D^2 p_h(x^*)(v) = 2E_P\, K_h'(\|X - x^*\|_{L^2}^2)v. \tag{A.31}$$

Thus, S and $S^\perp$ are invariant subspaces of $D^2 p_h(x^*)$. In order to see that $D^2 p_h(x^*)$ is indeed an isomorphism, it is therefore enough to show that it is isomorphism on both S and $S^\perp$ separately. Under assumption (H4'), p is a Morse density on S_c and there exists $h > 0$ small enough so that p_h is also a Morse function on the interior of S_c (see Remark 1). Then, x^* is in S_c by Proposition 1 and since $D^2 p_h(x^*)$ is equivalent to $\nabla^2 \tilde{p}_h(\tilde{x}^*)$ (the Hessian of $\tilde{p}_h$ at $\tilde{x}^*$), for h small enough $D^2 p_h(x^*)$ is an isomorphism on S. Since x^* is a non-trivial critical point of p_h, $p_h(x^*) = \delta > 0$. By (H2), $E_P\, K_h'(\|X - x^*\|_{L^2})) \le -\delta < 0$. According to equation (A.31), $D^2 p_h(x^*)$ acts on $S^\perp$ by multiplying every vector in $S^\perp$ by $2E_P\, K_h'(\|X - x^*\|_{L^2})$ and hence $D^2 p_h(x^*)$ is clearly an isomorphism on $S^\perp$. □

Proof of Lemma 9. Denote $T = -D^2 f(x^*)$ for simplicity. Then, T is a positive definite isomorphism on L^2. Thus, there exists $C > 0$ such that $\|T^{-1}\| \le C$ where $\|\cdot\|$ here denotes the operator norm. Also, it is straightforward to check that T induces a well-defined inner product $\langle\cdot,\cdot\rangle_T$ on L^2 by $\langle v, w\rangle_T = \langle Tv, w\rangle_{L^2}$. Now, for any $v \in L^2$ we have

$$\begin{aligned} \|v\|_{L^2}^2 &= \langle v, v\rangle_{L^2} = \langle T(T^{-1}v), v\rangle_{L^2} = \langle T^{-1}v, Tv\rangle_{L^2} = \langle T^{-1}v, v\rangle_T \\ &\le \|T^{-1}v\|_T \|v\|_T = \|v\|_T \sqrt{\langle T^{-1}v, T^{-1}v\rangle_T} \\ &= \|v\|_T \sqrt{\langle T(T^{-1}v), T^{-1}v\rangle_{L^2}} \le \|v\|_T \sqrt{\|T^{-1}v\|_{L^2}\|v\|_{L^2}} \\ &= \|v\|_T \sqrt{C}\|v\|_{L^2}. \end{aligned} \tag{A.32}$$

This implies that $\|v\|_{L^2}^2 \le C\|v\|_T^2$, and thus

$$\begin{aligned} \sup_{\|v\|_{L^2}=1} Df^2(x^*)(v, v) &= \sup_{\|v\|_{L^2}=1} -T(v, v) \\ &= \sup_{\|v\|_{L^2}=1} -\|v\|_T^2 \le -1/C. \end{aligned} \tag{A.33}$$

Therefore, by taking $\delta = 1/C$ the claim of the Lemma follows. □

Proof of Lemma 10. Let $\delta = \delta(x_2^*)$. The proof is in three steps.

Step 1. Suppose that $\eta_1 \leq \delta^2/8\beta_3$. Then, if $\epsilon = \delta/2\beta_3$, the solution of the initial value problem $\pi_1'(t) = Df_1(\pi_1(t))$, with $\pi_1(0) = x_2^*$ is contained in $B_{L^2}(x_2^*, \epsilon)$. In fact, suppose that the trajectory π_1 is not contained in $B_{L^2}(x_2^*, \epsilon)$. Since $\pi_1(0) = x_2^*$, there must exist $t_0 > 0$ such $\|\pi_1(t_0) - x_2^*\| = \epsilon$. Denote $\pi(t_0) = x_0$. Then since $Df_2(x_2^*) = 0$, a Taylor expansion implies that

$$\begin{aligned}
&f_1(x_0) \leq f_1(x_2^*) + \langle Df_1(x_2^*), x_0 - x_2^* \rangle_{L^2} \\
&+ \frac{1}{2} D^2 f_1(x_2^*)(x_0 - x_2^*, x_0 - x_2^*) \\
&+ \frac{1}{6}\beta_3 \|x_0 - x_2^*\|_{L^2}^3 \\
&\leq f_1(x_2^*) + \langle Df_2(x_2^*), x_0 - x_2^* \rangle_{L^2} \\
&+ \|Df_1(x_2^*) - Df_2(x_2^*)\|_{L^2} \|x_0 - x_2^*\|_{L^2} \\
&+ \frac{1}{2} D^2 f_2(x_2^*)(x_0 - x_2^*, x_0 - x_2^*) \\
&+ \frac{1}{2} \|D^2 f_1(x_2^*) - D^2 f_2(x_2^*)\|_{L^2} \|x_0 - x_2^*\|_{L^2}^2 \\
&+ \frac{1}{6}\beta_3 \|x_0 - x_2^*\|_{L^2}^3 \\
&\leq f_1(x_2^*) + \eta_1 \epsilon - \frac{1}{2}\delta\epsilon^2 + \frac{1}{2}\eta_2\epsilon^2 + \frac{1}{6}\beta_3\epsilon^3 \\
&\leq f_1(x_2^*) + \frac{1}{4}\delta\epsilon^2 - \frac{1}{2}\delta\epsilon^2 + \frac{1}{16}\delta\epsilon^2 + \frac{1}{12}\delta\epsilon^2 \\
&< f_1(x_2^*),
\end{aligned} \tag{A.34}$$

which is a contradiction because $f_1(\pi_1(t))$ is an non-decreasing function of t.

Step 2. By condition (C2), π_1 admits a convergent subsequence in L^2. Thus there is a subsequence $\{t_k\}_{k=1}^{\infty}$ and a critical point x_1^* such that $\|\pi_1(t_k) - x_1^*\|_{L^2} \to 0$ as $k \to \infty$ and $x_1^* \in B_{L^2}(x_2^*, \epsilon)$. In order to show that x_1^* is a non-degenerate local maximum in $B_{L^2}(x_2^*, \epsilon)$, consider $\eta_2 \leq \delta/8$. Given any $\|u\|_{L^2} = 1$, for any $x \in B_{L^2}(x_2^*, \epsilon)$ one has

$$\begin{aligned}
&D^2 f_1(x)(u,u) \\
&\leq D^2 f_1(x_2^*)(u,u) + |D^2 f_1(x_2^*)(u,u) - D^2 f_1(x)(u,u)| \\
&\leq D^2 f_2(x_2^*)(u,u) + |D^2 f_2(x_2^*)(u,u) - D^2 f_1(x_2^*)(u,u)| \\
&+ \beta_3 \|x_2^* - x\|_{L^2} \\
&\leq -\delta + \eta_2 + \beta_3\epsilon = -\frac{3}{8}\delta.
\end{aligned} \tag{A.35}$$

Therefore $\sup_{\|u\|_{L^2}=1} D^2 f_1(x)(u,u) \leq -3\delta/8$ and $Df_1(x_1^*)$ is negative definite. If one views $-Df_1(x_1^*)$ as a linear operator from L^2 to L^2,

then by the Lax-Milgram theorem, it is an isomorphism and hence x_1^* is a non-degenerate local maximum. Moreover, x_1^* is the unique maximum in $B_{L^2}(x_2^*, \epsilon)$: suppose that y_1^* is another local maximum of f_1 in $B_{L^2}(x_2^*, \epsilon)$; then, $Df_1(x_1^*) = 0$ and $Df_1(y_1^*) = 0$, and by equation (A.35), $\sup_{\|u\|_{L^2}=1} D^2 f_1(y_1^*)(u,u) \leq -3\delta/8$. A Taylor expansion shows that

$$\begin{aligned} f_1(x_1^*) &\leq f_1(y_1^*) + \frac{1}{2} D^2 f_1(y_1^*)(x_1^* - y_1^*, x_1^* - y_1^*) \\ &\quad + \frac{1}{6}\beta_3 \|x_1^* - y_1^*\|^3 \\ &\leq f_1(y_1^*) - \frac{3}{16}\delta \|x_1^* - y_1^*\|^2 + \frac{1}{6}\epsilon\beta_3 \|x_1^* - y_1^*\|^2 \\ &\leq f_1(y_1^*) - \frac{5}{48}\delta \|x_1^* - y_1^*\|^2 \end{aligned} \tag{A.36}$$

and by symmetry, $f_1(y_1^*) \leq f_1(x_1^*) - \frac{5}{48}\delta\|x_1^* - y_1^*\|^2$, which is a contradiction unless $y_1^* = x_1^*$.

Step 3. Now it is only left to show that $\|x_1^* - x_2^*\|_{L^2} \leq C\eta_1$. Since $Df_1(x)$ is a twice continuously differentiable function, a Taylor expansion around x_2^* allows us to write

$$\begin{aligned} &\langle Df_1(x_2^*), \cdot\rangle_{L^2} = \langle Df_1(x_1^*), \cdot\rangle_{L^2} + D^2 f_1(x_1^*)(x_2^* - x_1^*, \cdot) \\ &+ \int_0^1 \frac{1}{2} D^3 f_1(x_1^* + s(x_2^* - x_1^*))(x_2^* - x_1^*, x_2^* - x_1^*, \cdot)\, ds. \end{aligned} \tag{A.37}$$

Note that one can replace $Df_1(x_1^*)$ by $Df_2(x_2^*)$ as both of them are 0. Apply this identity to $x_2^* - x_1^*$, then

$$\begin{aligned} &\langle Df_1(x_2^*) - Df_1(x_2^*), x_2^* - x_1^*\rangle_{L^2} \\ &\leq D^2 f_1(x_1^*)(x_2^* - x_1^*, x_2^* - x_1^*) + \frac{1}{2}\beta_3 \|x_1^* - x_2^*\|_{L^2}^3 \\ &\leq -\frac{3}{8}\delta\|x_1^* - x_2^*\|^2 + \frac{1}{4}\delta\|x_1^* - x_2^*\|^2 \\ &\leq -\frac{1}{8}\delta\|x_1^* - x_2^*\|^2. \end{aligned} \tag{A.38}$$

This is equivalent to

$$\|x_1^* - x_2^*\|^2 \leq \frac{8}{\delta}\|Df_1(x_2^*) - Df_2(x_2^*)\|\|x_1^* - x_2^*\|. \tag{A.39}$$

Taking $C = 8$ completes the step. □

Proof of Proposition 4. First of all note that since $P(\|X\|_{H_0^1} \leq M) = 1$, lemma 5 ensures that all the non trivial critical points of $f_1(x) = p_h(x)$ and $f_2(x) = \hat{p}_h(x)$ are contained in $B_{H_0^1}(0, M)$. Let

$$\eta_1 = \sup_{x \in B_{H_0^1}(0,M)} \|D\hat{p}_h(x) - Dp_h(x)\|_{L^2} \tag{A.40}$$

and

$$\eta_2 = \sup_{x \in B_{H_0^1}(0,M)} \|D^2 \hat{p}_h(x) - D^2 p_h(x)\|. \tag{A.41}$$

Consider the events $A = \{\eta_1 \leq C_1(\alpha)\}$ and $B = \{\eta_2 \leq C_2(\alpha)\}$ where $C_1(\alpha)$ and $C_2(\alpha)$ are defined in Display 1. We can then use the uniform exponential inequalities on the first and second derivatives of Lemma 13 and Lemma 15 of $\hat{p}_h$ to ensure $P((A \cap B)^c) = P(A^c + B^c) \leq P(A^c) + P(B^c) \leq \alpha$ for n large enough (which will be justified later in the proof). For now, under the event $A \cap B$, for each point $\hat{x}^*$ marked by the algorithm of Display 1, i.e, $\hat{x}^* \in \hat{\mathcal{R}}$ we have

$$\delta^2 \geq 8\beta_3 C_1(\alpha) \geq 8\beta_3 \eta_1 \tag{A.42}$$

and

$$\delta \geq 8C_2(\alpha) \geq 8\eta_2, \tag{A.43}$$

hence the assumptions of Lemma 10 are satisfied. Furthermore, Lemma 10 ensures that the ball $B_{L^2}(\hat{x}^*, \delta(\hat{x}^*)/(2\beta_3))$ contains a unique non-degenerate local mode x^* of p_h and that $\|\hat{x}^* - x^*\|_{L^2} \leq 8\eta_1/\delta(\hat{x}^*) \leq 8C_1(\alpha)/\delta(\hat{x}^*)$ under the event $A \cap B$.

To justify $P(A^c) = P(\eta_1 \geq C_1(\alpha)) \leq \alpha/2$, consider the inequality of Lemma 13. We have

$$\begin{aligned} &P\left(\sup_{x \in B_{H_0^1}(0,M)} \|D\hat{p}_h(x) - Dp_h(x)\|_{L^2} \geq \epsilon \right) \\ &\leq C \exp\left(-\frac{4n\epsilon^2}{25K_1^2} + \frac{10MK_2}{\epsilon} \right). \end{aligned} \tag{A.44}$$

Let $a = 4/(25K_1^2)$, $b = 10MK_2$, $d = \log\left(\frac{\alpha}{2C}\right) < 0$. Take

$$\epsilon = \left(\frac{b}{a}\right)^{\frac{1}{3}} n^{-\frac{1}{3}} + \left(\frac{-d}{a}\right)^{\frac{1}{2}} n^{-\frac{1}{2}}. \tag{A.45}$$

Then,

$$\begin{aligned} &-\frac{4n\epsilon^2}{25K_1^2} + \frac{10MK_2}{\epsilon} \\ &= -an\epsilon^2 + \frac{b}{\epsilon} \leq -an\left(\left(\frac{b}{a}\right)^{\frac{1}{3}} n^{\frac{1}{3}} + \left(\frac{-d}{a}\right)^{\frac{1}{2}} n^{\frac{1}{2}} \right)^2 + \frac{b}{\left(\frac{b}{a}\right)^{\frac{1}{3}} n^{-\frac{1}{3}}} \\ &\leq -an\left(\left(\frac{b}{a}\right)^{\frac{2}{3}} n^{-\frac{2}{3}} + \frac{-d}{a} n^{-1} \right) + a^{\frac{1}{3}} b^{\frac{2}{3}} n^{\frac{1}{3}} \\ &= -a^{\frac{1}{3}} b^{\frac{2}{3}} n^{\frac{1}{3}} + d + a^{\frac{1}{3}} b^{\frac{2}{3}} n^{\frac{1}{3}} = d = \log\left(\frac{\alpha}{2C}\right). \end{aligned} \tag{A.46}$$

With this particular choice of $\epsilon = C_1(\alpha)$ it then follows that

$$\begin{aligned} &P\left(\sup_{x\in B_{H_0^1}(0,M)} \|D\hat{p}_h(x) - Dp_h(x)\|_{L^2} \geq \epsilon\right) \\ &\leq Ce^d = C\frac{\alpha}{2C} = \alpha/2. \end{aligned} \tag{A.47}$$

An almost identical argument is used to justify $P(B) = P(\eta_2 \geq C_2(\alpha)) \leq \alpha/2$. □

Proof of Proposition 5. Taking $f_1(x) = \hat{p}_h(x)$ and $f_2(x) = p_h(x)$, the goal is to apply Lemma 10 for all non-trivial critical points of p_h. It is worth to mention that, under the given assumption, $\mathcal{R}$ is a finite set and $\mathcal{R} \subset B_{H_0^1}(0, M)$. As a result, there exists a γ such that

$$-\gamma := \sup_{x^*\in\mathcal{C}} \sup_{\|u\|_{L^2}=1} D^2 p_h(x^*)(u,u) < 0. \tag{A.48}$$

According to Lemmas 13 and 15, for $l = 1, 2$ there exist constants $0 < H_l < \infty$ and $0 < h_l < \infty$ depending only on K_1, K_2 and K_3 and M such that

$$P\left(\sup_{x\in B_{H_0^1}(0,M)} \|D^l\hat{p}_h(x) - D^l p_h(x)\| \geq \frac{H_l}{n^{1/3}}\right) \leq C\exp\left(-h_l n^{1/3}\right). \tag{A.49}$$

Let η_l, $l = 1, 2$ be defined as in (C3). Let $F_n := \{\eta_1 \leq \frac{H_1}{n^{1/3}}\} \cap \{\eta_2 \leq \frac{H_2}{n^{1/3}}\}$. Then, $P(F_n) \to 1$. The rest of the argument follows by assuming that F_n holds.

Suppose that, for large n, $H_1 n^{-1/3} \leq \gamma^2/(8\beta_3)$ and $H_2 n^{-1/3} \leq \gamma/8$. Then, for all $x^* \in \mathcal{R}$, one has

$$\begin{aligned} &\eta_1 \leq H_1 n^{-1/3} \leq \gamma^2/(8\beta_3) \leq \delta(x^*)^2/(8\beta_3) \\ &\eta_2 \leq H_2 n^{-1/3} \leq \gamma/8 \leq \delta(x^*)/8, \end{aligned} \tag{A.50}$$

where as before

$$-\delta(x^*) := \sup_{\|u\|_{L^2}=1} D^2 p_h(x^*)(u,u) < 0. \tag{A.51}$$

One can apply Lemma 10 to all x^* to conclude that there exists a $\hat{x}^*$ such that

1. $\hat{x}^*$ is the unique local maximum of $\hat{p}_h$ in $B_{L^2}(x^*, \delta(x^*)/(2\beta_3))$;
2. $\delta(\hat{x}^*) := -\sup_{\|u\|_{L^2}=1} D^2\hat{p}_h(\hat{x}^*)(u,u) \geq 3\delta(x^*)/8 \geq 3\gamma/8$;
3. $\|x^* - \hat{x}^*\|_{L^2} \leq 8\eta_1/\delta(x^*)$.

The following three steps complete the proof.

step 1. In this step, one shows that $\hat{x}^* \in \hat{\mathcal{R}}$. According to item 2. in the first paragraph, $-\delta(\hat{x}^*) := \sup_{\|u\|_{L^2}=1} D^2\hat{p}_h(\hat{x}^*)(u,u) \leq -3\gamma/8$. Thus,

$$-\sup_{\|u\|_{L^2}=1} D^2\hat{p}_h(\hat{x}^*)(u,u) \geq 3\gamma/8 \geq \max\{\sqrt{8\beta_3 C_1(\alpha)}, 8C_2(\alpha)\} \tag{A.52}$$

because both $C_1(\alpha)$ and $C_2(\alpha)$ are of order $O(n^{-1/3})$.

step 2. One shows that $\Phi(\hat{x}^*) = x^*$, where Φ is defined in equation (5.6). Then, according to equation (5.6), it suffices to show that

$$x^* \in B_{L^2}(\hat{x}^*, \delta(\hat{x}^*)/(2\beta_3)) \cap B(\hat{x}^*, \log(n)C_1(\alpha)/\delta(\hat{x}^*)). \tag{A.53}$$

From item 3. in the first paragraph, $\|\hat{x}^* - x^*\| \leq 8\eta_1/\delta(x^*)$. Thus it suffices to show that

$$\begin{aligned} &B(\hat{x}^*, 8\eta_1/\delta(x^*)) \subset \\ &B_{L^2}(\hat{x}^*, \delta(\hat{x}^*)/(2\beta_3)) \cap B(\hat{x}^*, \log(n)C_1(\alpha)/\delta(\hat{x}^*)). \end{aligned} \tag{A.54}$$

This is equivalent to

$$8\eta_1/\delta(x^*) \leq \delta(\hat{x}^*)/(2\beta_3) \text{ and } 8\eta_1/\delta(x^*) \leq \log(n)C_1(\alpha)/\delta(\hat{x}^*). \tag{A.55}$$

The first inequality of (A.55) is clear because

$$\delta(\hat{x}^*) \geq 3\gamma/8 \quad \text{and} \quad 8\eta_1/\delta(x^*) \leq 8H_1 n^{-1/3}/\gamma = O(n^{-1/3}). \tag{A.56}$$

The second one holds for large n because

$$8\eta_1/\delta(x^*) \leq 8H_1 n^{-1/3}/\gamma \tag{A.57}$$

while

$$\begin{aligned} &\log(n)C_1(\alpha)/\delta(\hat{x}^*) \\ &\geq \log(n)C_1(\alpha)/\beta_2 = C(\alpha, K_1, K_2, K_3, M)n^{-1/3}\log(n). \end{aligned} \tag{A.58}$$

step 3. To complete the argument, it suffices to show that if $\Phi(\hat{y}^*) = x^*$ for some $\hat{y}^* \in \hat{\mathcal{R}}$, then $\hat{y}^* = \hat{x}^*$. Since $\hat{y}^* \in \hat{\mathcal{R}}$, by the algorithm of Display 1,

$$\delta(\hat{y}^*) \geq \max\{\sqrt{8\beta_3 C_1(\alpha)}, 8C_2(\alpha)\}. \tag{A.59}$$

Thus, $\delta(\hat{y}^*) \geq c(\alpha, M, K_1, K_2, K_3)n^{-1/6}$ for some $c(\alpha, MK_1, K_2, K_3) > 0$ independent of n. As a result, since $\Phi(\hat{y}^*) = x^*$,

$$\|x^* - \hat{y}^*\|_{L^2} \leq \log(n)C_1(\alpha)/\delta(\hat{y}^*) = O(\log(n)n^{-1/6}). \tag{A.60}$$

Then, for large n

$$\|\hat{y}^* - x^*\|_{L^2} \leq \gamma/(2\beta_3) \leq \delta(x^*)/(2\beta_3) \tag{A.61}$$

and therefore $\hat{y}^* \in B(x^*, \delta(x^*)/(2\beta_3))$. According to item 1. in the first paragraph, $\hat{x}^*$ is the unique local maximum of $\hat{p}_h$ in $B(x^*, \delta(x^*)/(2\beta_3))$. It thus follows that $\hat{y}^* = \hat{x}^*$. □

Proof of Lemma 11. We discuss the case $l = 1$. Only the constants differ in the remaining cases. For any $x \in L^2$

$$\begin{aligned}
&\|D\hat{p}_h(x) - D\tilde{p}_h(x)\| \\
\leq &\frac{2}{n}\sum_{i=1}^{n}\left\|K_h'(\|X_i - x\|^2)_{L^2}(x - X_i) - K_h'(\|\tilde{X}_i - x\|_{L^2})(x - \tilde{X}_i)\right\| \\
\leq &\frac{2}{n}\sum_{i=1}^{n}K_2\|X_i - x - (\tilde{X}_i - x)\| \\
= &\frac{2}{n}\sum_{i=1}^{n}K_2\|X_i - \tilde{X}_i\|
\end{aligned} \tag{A.62}$$

Thus,

$$E\left(\sup_{x\in L^2}\|D\hat{p}_h(x) - D\tilde{p}_h(x)\| \,\middle|\, X_1, \ldots, X_n\right) \leq 2K_2\phi(m), \tag{A.63}$$

where $\phi(m)$ does not depend on X_i. Therefore, this implies

$$E\left(\sup_{x\in L^2}\|D\hat{p}_h(x) - D\tilde{p}_h(x)\|\right) \leq 2K_2\phi(m). \tag{A.64}$$

As a result,

$$P\left(\sup_{x\in L^2}\|D\hat{p}_h(x) - D\tilde{p}_h(x)\| \geq \epsilon\right) \leq \frac{2K_2\phi(m)}{\epsilon}. \tag{A.65}$$

□

Proof of Corollary 1. The argument for the first part is almost the same as the one in Proposition 4, except that in this case one makes use of the fact that

$$\begin{aligned}
&P\left(\sup_{x\in B_{H^1}(0,M)}\|Dp_h(x) - D\tilde{p}_h(x)\| \geq \tilde{C}_1(\alpha)\right) \\
\leq &P\left(\sup_{x\in B_{H^1}(0,M)}\|Dp_h(x) - D\hat{p}_h(x)\| \geq C_1(\alpha/2)\right) \\
+ &P\left(\sup_{x\in B_{H^1}(0,M)}\|D\hat{p}_h(x) - D\tilde{p}_h(x)\| \geq \frac{8K_2\phi(m)}{\alpha}\right) \\
\leq &\alpha/4 + \alpha/4 = \alpha/2,
\end{aligned} \tag{A.66}$$

and that

$$P\left(\sup_{x\in B_{H^1}(0,M)}\|D^2p_h(x) - D^2\tilde{p}_h(x)\| \geq \tilde{C}_2(\alpha)\right) \leq \alpha/2. \tag{A.67}$$

The argument for the second part is the same as the one in Proposition 5, except that equation (A.49) becomes

$$\begin{aligned} & P\left(\sup_{x\in B_{H_0^1}(0,M)} \|D^l\tilde{p}_h(x) - D^l p_h(x)\| \geq \frac{H_l}{n^{1/3}} + \sqrt{\phi(m)}\right) \\ \leq & P\left(\sup_{x\in B_{H_0^1}(0,M)} \|D^l\tilde{p}_h(x) - D^l\hat{p}_h(x)\| \geq \sqrt{\phi(m)}\right) \\ & + P\left(\sup_{x\in B_{H_0^1}(0,M)} \|D^l p_h(x) - D^l\hat{p}_h(x)\| \geq \frac{H_l}{n^{1/3}}\right) \\ \leq & C\left(\exp(-h_l n^{1/3}) + \sqrt{\phi(m)}\right). \end{aligned} \tag{A.68}$$

□

Appendix B: Additional results

Lemma 12. *Under the assumptions (H1), (H2), and (H3),*

$$\begin{aligned} & P\left(\sup_{x\in B_{H_0^1}(0,M)} |\hat{p}_h(x) - p_h(x)| \geq \epsilon\right) \\ & \leq C\exp\left(-\frac{32n\epsilon^2}{25K_0^2} + \frac{10MK_1}{\epsilon}\right) \end{aligned} \tag{B.1}$$

for ϵ sufficiently small.

Proof. By Chapter 7 of Shiryayev (1993), the covering number N_ϵ of the ball $B_{H_0^1}(0,M)$ satisfies $N_\epsilon \leq C\exp\left(\frac{M}{\epsilon}\right)$. Let $\epsilon' = \epsilon/(10K_1)$. For a fixed radius M, pick $\{x_k\}_{k=1}^{N_{\epsilon'}}$ such that if $x \in B_{H_0^1}(0,M)$ then there exists $\|x_k - x\|_{L^2} \leq \epsilon'$. Note that for any fixed $x \in B_{H_0^1}(0,M)$,

$$\begin{aligned} |\hat{p}_h(x) - \hat{p}_h(x_k)| &= \left|\frac{1}{n}\sum_{i=1}^n K_h(\|x - X_i\|_{L^2}^2) - K_h(\|x_k - X_i\|_{L^2}^2)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n \left|K_h(\|x - X_i\|_{L^2}^2) - K_h(\|x_k - X_i\|_{L^2}^2)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n K_1\|x - x_k\|_{L^2} = \frac{\epsilon}{10}. \end{aligned} \tag{B.2}$$

Thus, $|p_h(x) - p_h(x_k)| \leq E_P|\hat{p}_h(x) - \hat{p}_h(x_k)| \leq \frac{\epsilon}{10}$. Since for any x,

$$\begin{aligned} & |\hat{p}_h(x) - p_h(x)| \\ & \leq |\hat{p}_h(x) - \hat{p}_h(x_k)| + |\hat{p}_h(x_k) - p_h(x_k)| + |p_h(x_k) - p_h(x)| \\ & \leq |\hat{p}_h(x_k) - p_h(x_k)| + \frac{\epsilon}{5}, \end{aligned} \tag{B.3}$$

it follows that

$$\begin{aligned} & P\left(\sup_{x\in B_{H_0^1}(0,M)} |\hat{p}_h(x) - p_h(x)| \geq \epsilon\right) \\ & \leq P\left(\sup_{1\leq k\leq N_{\epsilon'}} |\hat{p}_h(x_k) - p_h(x_k)| \geq \frac{4\epsilon}{5}\right) \\ & \leq N_{\epsilon'} P\left(|\hat{p}_h(x_1) - p_h(x_1)| \geq \frac{4\epsilon}{5}\right) \\ & \leq C\exp\left(\frac{10MK_1}{\epsilon}\right)\exp\left(-\frac{32n\epsilon^2}{25K_0^2}\right), \end{aligned} \tag{B.4}$$

where the last step uses Hoeffding's inequality. □

Lemma 13. *Under the same assumptions of the last Lemma, for ϵ sufficiently small,*

$$\begin{aligned} & P\left(\sup_{x\in B_{H_0^1}(0,M)} \|D\hat{p}_h(x) - Dp_h(x)\|_{L^2} \geq \epsilon\right) \\ & \leq C\exp\left(-\frac{4n\epsilon^2}{25K_1^2} + \frac{10MK_2}{\epsilon}\right). \end{aligned} \tag{B.5}$$

Proof. The proof is very similar to that of the previous Lemma. Notice first that

$$\begin{aligned} & \left\|K_h'(\|x - X_i\|_{L^2}^2)(x - X_i) - K_h'(\|x_k - X_i\|_{L^2}^2)(x_k - X_i)\right\|_{L^2} \\ & \leq K_2\|x - x_k\|_{L^2}. \end{aligned} \tag{B.6}$$

By taking $\epsilon' = \epsilon/(10K_2)$ and using the same argument of the previous Lemma, we have

$$\begin{aligned} & P\left(\sup_{x\in B_{H_0^1}(0,M)} \|D\hat{p}_h(x) - Dp_h(x)\|_{L^2} \geq \epsilon\right) \\ & \leq N_{\epsilon'} P\left(\|D\hat{p}_h(x_1) - Dp_h(x_1)\|_{L^2} \geq \frac{4\epsilon}{5}\right). \end{aligned} \tag{B.7}$$

In oder to proceed, we need an Hoeffding-type inequality for Hilbert spaces. Specifically, using Lemma 1, one has

$$Dp_h(x_1) = \frac{2}{n}\sum_{i=1}^{n} E_P K_h'(\|X_i - x_1\|_{L^2}^2)(x_1 - X_i). \tag{B.8}$$

Now, if one denotes Z_i as

$$Z_i = 2K_h'(\|X_i - x_1\|_{L^2}^2)(x_1 - X_i) - 2E_P K_h'(\|X_i - x_1\|_{L^2}^2)(x_1 - X_i), \tag{B.9}$$

then

$$D\hat{p}_h(x_1) - Dp_h(x_1) = \frac{1}{n}\sum_{i=1}^{n} Z_i. \tag{B.10}$$

Thus, $E_P Z_i = 0$ as a L^2 function and $\|Z_i\|_{L^2} \leq 4K_1$.

Finally, by using the exponential inequality of the Corollary of Lemma 4.3 in Yurinskiĭ (1976),

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i\right\|_{L^2} \geq \frac{4\epsilon}{5}\right) \leq 2\exp\left\{-\frac{16n\epsilon^2}{50K_1^2}\left(1+\frac{1.62\epsilon}{\frac{1}{5}K_1}\right)^{-1}\right\} \tag{B.11}$$

and for ϵ sufficient small that $\left(1+\frac{1.62\epsilon}{K_1/5}\right) \leq 2$, one gets the desired result. □

Next we derive a similar result for the second derivative. Obtaining such a result is a little more difficult because the operator norm of a linear operator defined on a Hilbert space does not induce a Hilbert space structure. The following discussion and intermediate results are useful to circumvent this problem.

Definition 3. *Let $A : L^2 \to L^2$ be a linear operator. A is said to be a Hilbert-Schmidt operator on L^2 if*

$$\|A\|_{HS}^2 := \sum_{i=1}^{\infty} \|Ae_i\|_{L^2}^2 < \infty \tag{B.12}$$

where $\{e_i\}_{i=1}^{\infty}$ is an orthonormal basis of L^2.

Remark 10. The above definition is independent of the choice of the orthonormal basis. Furthermore, Hilbert-Schmidt operators form a Hilbert space with the following inner product: for two Hilbert-Schmidt operators A and B, the Hilbert-Schmidt inner product between A and B is defined as

$$\langle A, B\rangle_{HS} = \sum_{i=1}^{\infty} \langle Ae_i, Be_i\rangle_{L^2} \tag{B.13}$$

where $\{e_1\}_{i=1}^{\infty}$ is any orthonormal basis of L^2. Recall that the operator norm of bilinear operator A is defined as

$$\|A\| = \sup_{\{v\,:\,\|v\|_{L^2}=1\}} \|A(v)\|_{L^2} \tag{B.14}$$

A standard result guarantees that $\|A\| \leq \|A\|_{HS}$.

Lemma 14. *Let*

$$B(\cdot,\cdot) = 4K_h''(\|X-x\|_{L^2}^2)\langle x-X,\cdot\rangle_{L^2}\langle x-X,\cdot\rangle_{L^2}. \tag{B.15}$$

Then $\|B\|_{HS} \leq 4K_2$ P-almost surely.

Proof. Let $Y = 2\sqrt{K_h''(\|X-x\|_{L^2}^2)}(x-X)$, hence $Y \in L^2$. It is easily seen that $\|Y\|_{L^2} \leq 2\sqrt{K_2}$ P-almost surely by (H1). Consider $\bar{B}(v) = \langle Y, v\rangle_{L^2}$ and let $\{e_i\}_{i=1}^{\infty}$ be an orthonormal basis. We can write $Y = \sum_{i=1}^{\infty} y_i e_i$, where y_i

are random coefficients. Therefore, $\|Y\|_{L^2}^2 = \sum_{i=1}^{\infty} y_i^2 \leq 4K_2$ P-almost surely. Finally, $B : L^2 \to L^2$ can be expressed as $B(v) = \bar{B}(v)Y$ and

$$\begin{aligned}\|B\|_{HS}^2 &= \sum_{i=1}^{\infty} \|\bar{B}(e_i)Y\|_{L^2}^2 = \sum_{i=1}^{\infty} \|\langle Y, e_i\rangle_{L^2} Y\|_{L^2}^2 = \sum_{i=1}^{\infty} \|y_i Y\|_{L^2}^2 \\ &= \|Y\|_{L^2}^2 \sum_{i=1}^{\infty} y_i^2 = \|Y\|_{L^2}^4.\end{aligned} \tag{B.16}$$

This complete the proof. □

Lemma 15. *Under the same assumption of Lemma 12, for ϵ small enough so that $\left(1 + \frac{1.62\epsilon}{K_1/5}\right) \leq 2$, we have*

$$\begin{aligned}&P\left(\sup_{x \in B_{H_0^1}(0,M)} \|D^2\hat{p}_h(x) - D^2 p_h(x)\| \geq \epsilon\right) \\ &\leq C \exp\left(-\frac{n\epsilon^2}{25K_2^2} + \frac{10MK_3}{\epsilon}\right).\end{aligned} \tag{B.17}$$

Proof. If ϵ' is taken to be $\epsilon/(10K_3)$, one has

$$\begin{aligned}&P\left(\sup_{x \in B_{H_0^1}(0,M)} \|D^2\hat{p}_h(x) - D^2 p_h(x)\| \geq \epsilon\right) \\ &\leq N_{\epsilon'} P\left(\|D^2\hat{p}_h(x_1) - D^2 p_h(x_1)\| \geq \frac{4\epsilon}{5}\right).\end{aligned} \tag{B.18}$$

Let

$$B_i(\cdot,\cdot) = 4K_h''(\|X_i - x\|_{L^2}^2)\langle x - X_i, \cdot\rangle_{L^2}\langle x - X_i, \cdot\rangle_{L^2} \tag{B.19}$$

and

$$C_i(\cdot,\cdot) = 2K_h'(\|X_i - x\|_{L^2}^2)\langle \cdot, \cdot\rangle_{L^2}. \tag{B.20}$$

For any bilinear operator $T(v,w) = t\langle v, w\rangle$, where $t \in \mathbb{R}$, then $\|T\| = |t|$. Thus,

$$\begin{aligned}&\|D^2\hat{p}_h(x) - D^2 p_h(x)\| \\ &= \left\|\frac{1}{n}\sum_{i=1}^{n}(B_i - E_P(B_i)) + \frac{1}{n}\sum_{i=1}^{n}(C_i - E_P(C_i))\right\| \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^{n}(B_i - E_P(B_i))\right\| + \left\|\frac{1}{n}\sum_{i=1}^{n}(C_i - E_P(C_i))\right\| \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^{n}(B_i - E_P(B_i))\right\| \\ &+ 2\left|\frac{1}{n}\sum_{i=1}^{n} K_h'(\|x - X_i\|_{L^2}^2) - E_P K_h'(\|x - X\|_{L^2}^2)\right|.\end{aligned} \tag{B.21}$$

As a result,

$$\begin{aligned} &P\left(\|D^2\hat{p}_h(x_1) - p_h(x_1)\| \geq \frac{4\epsilon}{5}\right) \\ &\leq P\left(\left\|\sum_{i=1}^{n}\frac{1}{n}\left(B_i - E_p(B_i)\right)\right\| \geq \frac{2\epsilon}{5}\right) \\ &+ P\left(2\left|\frac{1}{n}\sum_{i=1}^{n}K_h'(\|x - X_i\|_{L^2}^2) - E_P K_h'(\|x - X\|_{L^2}^2)\right| \geq \frac{2\epsilon}{5}\right). \end{aligned} \tag{B.22}$$

Since $|K_h'(\|x - X_i\|_{L^2}^2)| \leq K_2$, Hoeffding's inequality implies

$$\begin{aligned} &P\left(\left|\frac{1}{n}\sum_{i=1}^{n}K_h'(\|x - X_i\|_{L^2}^2) - E_P K_h'(\|x - X\|_{L^2}^2)\right| \geq \frac{2\epsilon}{5}\right) \\ &\leq 2\exp\left(-\frac{8n\epsilon^2}{25K_2^2}\right). \end{aligned} \tag{B.23}$$

In order to apply the Corollary of Lemma 4.3 in Yurinskiĭ (1976) on the Hilbert-Schmidt operator norm, it suffices to check that $B_i - E_P(B_i)$ has bounded Hilbert-Schmidt norm. Lemma 14 guarantees that

$$\|B_i - E_P(B_i)\|_{HS} \leq 8K_2 \tag{B.24}$$

almost surely. Therefore, since $\|A\| \leq \|A\|_{HS}$ for any bilinear operator A, with small enough ϵ, then

$$\begin{aligned} &P\left(\left\|\sum_{i=1}^{n}\frac{1}{n}\left(B_i - E_p(B_i)\right)\right\| \geq \frac{2\epsilon}{5}\right) \\ &\leq P\left(\left\|\sum_{i=1}^{n}\frac{1}{n}\left(B_i - E_p(B_i)\right)\right\|_{HS} \geq \frac{2\epsilon}{5}\right) \\ &\leq 2\exp\left(-\frac{4n\epsilon^2}{50K_2^2}\left(1 + \frac{1.62\epsilon}{\frac{1}{5}K_1}\right)^{-1}\right) \\ &\leq 2\exp\left(-\frac{n\epsilon^2}{25K_2^2}\right). \end{aligned} \tag{B.25}$$

□

Acknowledgements

The authors are grateful to Prof. Domenico Marinucci, to an anonymous Associate Editor, and to two anonymous referees for their very constructive comments which led to a clearer and more complete version of this paper.

The authors are thankful to Andrew Schwartz, Valérie Ventura, and Sonia Todorova for sharing the neural activity recordings used in the application section of this paper.

M. Ciollaro and D. Wang wish to thank Giovanni Leoni and Xin Yang Lu for the very useful conversations leading to several improvements in the early versions of this paper.

References

Ambrosetti, A. and Prodi, G. (1995). *A Primer of Nonlinear Analysis* **34**. Cambridge University Press. MR1336591

Bongiorno, E. and Goia, A. (2015). A clustering method for Hilbert functional data based on the small ball probability. *arXiv preprint arXiv:1501.04308*.

Carreira-Perpiñán, M. Á. (2006). Fast nonparametric clustering with Gaussian blurring mean-shift. In *Proceedings of the 23rd International Conference on Machine Learning* 153–160.

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science* **30** 518–532. MR3432839

Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A. and Wasserman, L. (2014). Robust topological inference: distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197*.

Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17** 790–799.

Ciollaro, M., Genovese, C. R., Lei, J. and Wasserman, L. (2014). The mean-shift algorithm for mode hunting and clustering in infinite dimensions. *arXiv preprint arXiv:1408.1187*.

Comaniciu, D., Ramesh, V. and Meer, P. (2001). The variable bandwidth mean shift and data-driven scale selection. In *Proceedings of the Eighth IEEE International Conference on Computer Vision* **1** 438–445.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference* **147** 1–23. MR3151843

Delaigle, A. and Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics* **38** 1171–1193. MR2604709

Evans, L. C. (1998). *Partial Differential Equations*. American Mathematical Society. MR1625845

Ferraty, F., Kudraszow, N. and Vieu, P. (2012). Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. *Journal of Nonparametric Statistics* **24** 447–464. MR2921146

Ferraty, F. and Romain, Y. (2011). *The Oxford Handbook of Functional Data Analaysis*. Oxford University Press. MR2917982

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer. MR2229687

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory* **21** 32–40. MR0388638

Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society, Series B* **60** 681–691. MR1649539

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics* 171–185. MR0767241

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society, Series B* **78** 99–126. MR3453648

Goia, A. and Vieu, P. (2015). An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*. MR3477644

Hall, P. and Heckman, N. E. (2002). Estimating and depicting the structure of a distribution of random functions. *Biometrika* **89** 145–158. MR1888371

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer. MR2920735

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, With an Introduction to Linear Operators*. John Wiley & Sons. MR3379106

Hunter, J. K. and Nachtergaele, B. (2001). *Applied Analysis*. World Scientific. MR1829589

Jacques, J. and Preda, C. (2013). Functional data clustering: a survey. *Advances in Data Analysis and Classification* 1–25. MR3253859

Jost, J. (2011). *Riemannian Geometry and Geometric Analysis*. Springer. MR2829653

Kudraszow, N. L. and Vieu, P. (2013). Uniform consistency of kNN regressors for functional variables. *Statistics & Probability Letters* **83** 1863–1870. MR3069890

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer. MR2168993

Schechter, M. (2004). *An Introduction to Nonlinear Analysis*. Cambridge University Press. MR2127432

Scott, D. W. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons. MR3329609

Shiryayev, A. N. (1993). *Selected Works of A. N. Kolmogorov: Volume III: Information Theory and the Theory of Algorithms* **27**. Springer Science & Business Media. MR1228446

Taylor, D. M., Tillery Helms, S. I. and Schwartz, A. B. (2002). Direct cortical control of 3D neuroprosthetic devices. *Science* **296** 1829–1832.

Yurinskiĭ, V. V. (1976). Exponential inequalities for sums of random vectors. *Journal of Multivariate Analysis* **6** 473–499. MR0428401

Zhang, J.-T. (2013). *Analysis of Variance for Functional Data*. CRC Press.