Electronic Journal of Statistics
Vol. 12 (2018) 249–276
ISSN: 1935-7524
https://doi.org/10.1214/17-EJS1382

Empirical evolution equations

Susan Wei and Victor M. Panaretos

University of Minnesota and Ecole Polytechnique Fédérale de Lausanne
e-mail: susanwei@umn.edu; victor.panaretos@epfl.ch

Abstract: Evolution equations comprise a broad framework for describing the dynamics of a system in a general state space: when the state space is finite-dimensional, they give rise to systems of ordinary differential equations; for infinite-dimensional state spaces, they give rise to partial differential equations. Several modern statistical and machine learning methods concern the estimation of objects that can be formalized as solutions to evolution equations, in some appropriate state space, even if not stated as such. The corresponding equations, however, are seldom known exactly, and are empirically derived from data, often by means of nonparametric estimation. This induces uncertainties on the equations and their solutions that are challenging to quantify, and moreover the diversity and the specifics of each particular setting may obscure the path for a general approach. In this paper, we address the problem of constructing general yet tractable methods for quantifying such uncertainties, by means of asymptotic theory combined with bootstrap methodology. We demonstrates these procedures in important examples including gradient line estimation, diffusion tensor imaging tractography, and local principal component analysis. The bootstrap perspective is particularly appealing as it circumvents the need to simulate from stochastic (partial) differential equations that depend on (infinite-dimensional) unknowns. We assess the performance of the bootstrap procedure via simulations and find that it demonstrates good finite-sample coverage.

Keywords and phrases: Diffusion tensor imaging, gradient line, heat flow, integral curve, local principal curve, scale space, vector field.

Received August 2017.

Contents

1. Introduction

Given a Banach space $\mathcal{X}$ (a complete normed vector space) and a vector field v on $\mathcal{X}$, an *evolution equation* (Walker, 1980) is a model of the form

$$\dot{\gamma} = v(\gamma) \quad \text{and} \quad \gamma(0) = x_0 \in \mathcal{X}, \tag{1.1}$$

specifying a (differentiable) flow $\gamma : \mathbb{R} \to \mathcal{X}$ by relating its time derivative $\dot{\gamma}$ to the field v. A flow γ solving (1.1) is often called an *integral curve* of v, borrowing the established term from the case where $\mathcal{X}$ is Euclidean (solving a set of ordinary differential equations is colloquially referred to as "integrating" the system) (Lee, 1992). The abstract formulation (1.1) spans a broad and rich class of models in the physical sciences, encompassing ordinary as well as partial differential equations as special cases. In physics, integral curves are known as lines of force if v is a force field, or lines of flow if v is a velocity field of fluid flow (Zachmanoglou and Thoe, 1986, Chapter 2). In the study of dynamical systems, they are known as trajectories or orbits (Lee, 1992, Chapter 2). The typical treatment of evolution equations takes v to be known and studies questions as related to existence, uniqueness, and regularity of the induced integral curves (Lee, 2012, Chapter 17), as well as their stability under perturbations of initial conditions (Bellman, 2008, Chapters 2, 4).

Less explored is a setting where v is unknown and measured empirically under uncertainty. This setting can nevertheless be seen to encompass a surprising number of statistical and machine learning objects of interest, which can be formalized as solutions to empirical evolution equations. Important examples include filament estimation in galaxies (Genovese et al., 2009), diffusion tensor imaging tractography (Koltchinskii, Sakhanenko and Cai, 2007), fingerprint analysis (Huckemann, Hotz and Munk, 2008; Hill, Kendall and Thönnes, 2012), and structural economics (Vanhems, 2006). Less obvious examples include the widely used stochastic gradient descent algorithm for solving large-scale machine learning tasks (Bottou, 2010) can also be seen as a special case of integrating a (sampled) vector field. Given the empirical nature of the corresponding equations, one is compelled to consider methods for quantifying the induced uncertainty on its solution. The diversity of applied settings and methodological constructions can lead to a multitude of corresponding problem-specific approaches, but the elegant general specification begs the question whether a broad framework of uncertainty quantification that can be tractably ported to each specific context is feasible.

This question motivates us to study general evolution equations of the form (1.1) through the lens of sampling variation, when an unknown vector field v is replaced by a (potentially nonparametric) estimate $\hat{v}_n$ depending on sampled

data whose "sample size" (in a general sense) is n. This setting gives rise to what we term an *empirical evolution equation*:

$$\dot{\gamma} = \hat{v}_n(\gamma) \quad \text{and} \quad \gamma(0) = x_0 \in \mathcal{X}.$$

Its solution $\hat{\gamma}_n$ will be called the *empirical integral curve*. In order to establish valid general inference procedures based on the empirical integral curve, we investigate how the asymptotic (in n) behavior of the estimator $\hat{v}_n$ of v relates that of the estimator $\hat{\gamma}_n$ of γ (Section 2) when the state space $\mathcal{X}$ is a (potentially infinite dimensional) Banach space. We then exploit these results in order to provide asymptotically valid bootstrap procedures for the integral curve (Section 3). Such bootstrapping is particularly useful from a practical standpoint, as it allows us to bypass solving stochastic differential equations with unknown parameters to find the limiting distribution of the integral curve. Indeed, we develop bootstrap procedures that are valid uniformly over smoothing parameters involved in the construction of $\hat{v}_n$, thus correctly accounting for any uncertainties arising from data-dependent regularization (Section 4). We show how our results can be applied in the context of several modern statistical and machine learning problems (Section 5). These include finite-dimensional evolution equations arising in gradient line estimation, diffusion tensor tractography, and principal curve estimation; and an infinite dimensional setting describing anisotropic heat flow. Our work also elicits a noteworthy aspect of inference related to evolution equations: when the empirical evolution equation is based on a nonparametric estimate of the vector field, it appears necessary to adopt a scale-space view (Chaudhuri and Marron, 2000) if valid bootstrap procedures are desired (Section 4). The proofs of all formal statements are collected in Appendix B.

2. Asymptotic theory for the integral curve

We begin by investigating the limiting distribution of $r_n(\hat{\gamma}_n - \gamma_0)$ at some appropriate rate r_n, where the subscript "0" is used to indicate the *true* parameter values under (1.1). We will need to distinguish between two regimes in the asymptotics, depending on the behavior of the vector field estimator $\hat{v}_n$:

1. Regime 1: the process $\hat{v}_n$ converges weakly to a tight limiting process (Theorem 1).
2. Regime 2: $\hat{v}_n$ itself may fail to converge to a tight limit, but a related functional does (Theorem 2).

The limiting result under the first regime, derived in Section 2.2, can be interpreted as a delta method for the integral curve. We will show in Section 3 the first regime is particularly amenable to the construction of valid bootstrap procedures. The second regime, treated in Section 2.3, also accommodates tractable asymptotics, if at the cost of more stringent conditions on the vector field v_0. Of greater consequence, though, is the possible lack of consistent bootstrap procedures for the integral curve under this second regime.

2.1. Notation

We will make heavy use of the following notation and key notions in our subsequent development. Let E and F be Banach spaces, with E compact and Hausdorff. Let $C(E,F)$ be the Banach space of all continuous maps $E \to F$ equipped with the supremum norm $\|\cdot\|_\infty$. Let $L(E,F)$ denote the Banach space of all continuous linear maps $E \to F$ in the operator norm $\|f\|_{op} = \sup\left\{\|f(x)\|_F : \|x\|_E \le 1\right\}$. Next denote by $l^\infty(E,F)$ the Banach space of all uniformly norm-bounded maps $E \to F$ equipped with the supremum norm. Let $U \subset E$ be open. The function f is called *Fréchet differentiable*, with derivative Df, in case it is Fréchet differentiable at all points x in U. Let $C^1(U,F)$ denote the Banach space of continuously (Fréchet) differentiable maps $U \to F$ under the C^1 norm $\|f\|_{C^1} = \|Df\|_\infty + \|f\|_\infty$.

Let $\mathcal{X}$ be a Banach space, and $\mathcal{U} \subset \mathcal{X}$ be open. Let $v_0 \in C(\mathcal{U},\mathcal{X})$ be the vector field of interest. For well-definedness, we will assume henceforth that v_0 is locally Lipschitz on $\mathcal{U}$. This implies, by the Picard-Lindelof theorem (Nelson, 1969), that there is some interval $I \subset \mathbb{R}$ containing zero on which there exists a unique integral curve $\gamma_0 : I \to \mathcal{U}$ that solves (1.1). Throughout the paper, the asymptotic results hold for a fixed initial condition $x_0 \in \mathcal{U}$ and corresponding interval I.

2.2. Asymptotics in Regime 1

For parametric vector fields, one typically has a method for estimating the parameter defining v_0, along with a limit theorem for the estimator. For instance, Ramsay et al. (2007) and Brunel, Clairon and D'Alché-Buc (2014) treat precisely such a setting, proposing parametric vector field estimates for which they then derive asymptotics. Our results apply to the parametric setting too, though our eventual interest is in nonparametric vector field estimators. The current development lays the groundwork for the uniform-in-bandwidth result of Section 4, which is crucial to the nonparametric setting. We will assume that for some r_n it has been established that $r_n(\hat{v}_n - v_0)$ converges weakly to a tight limit, but will not circumscribe the type of estimator $\hat{v}_n$ can be, beyond this. Theorem 1 will require the following:

Assumption A1. The vector field v_0 is Fréchet differentiable at $\gamma_0(t)$ for all $t \in I$.

Note that when v_0 is C^1, it is both locally Lipschitz and Fréchet differentiable. The following result is derived using tools from Z-estimation theory (van der Vaart and Wellner, 1996).

Theorem 1. Suppose A1 holds. In addition, suppose the vector field estimate $\hat{v}_n$ is such that

$$r_n(\hat{v}_n - v_0) \xrightarrow{d} \mathbb{G} \text{ in } (l^\infty(\mathcal{U},\mathcal{X}), \|\cdot\|_\infty) \tag{2.1}$$

for some sequence $r_n \ge 0$, $r_n \to \infty$ and tight process $\mathbb{G}$ that satisfies

$$\|\mathbb{G}(\gamma_n(t)) - \mathbb{G}(\gamma_0(t))\|_{\mathcal{X}} \xrightarrow{\text{a.s.}} 0 \quad \forall t \in I \text{ as } \gamma_n \xrightarrow{\text{a.s.}} \gamma_0.$$

Then it holds that $r_n(\hat{\gamma}_n - \gamma_0) \xrightarrow{d} \xi$ in $(C^1(I,\mathcal{U}),\|\cdot\|_{C^1})$ where

$$d\xi(t) = Dv_0(\gamma_0(t))\xi(t)\,dt + \mathbb{G}(\gamma_0(t)) \quad \text{and} \quad \xi(0) = 0. \tag{2.2}$$

If $\mathbb{G}$ is a Gaussian process, then so is ξ.

Note the broadness of the result: $\mathcal{X}$ is assumed to be a general, potentially infinite-dimensional Banach space. The limiting distribution can then be interpreted as a stochastic partial differential equation governing the integral curve. The result simplifies considerable in the special case $\mathcal{X} = \mathbb{R}^d$. Since this is prominent in statistical applications, we present it in more detail. In this case, $v : \mathcal{U} \to \mathcal{X} = \mathbb{R}^d$ corresponds to the familiar notion of a classic vector field on $\mathbb{R}^d$. The associated integral curve γ has the property that the tangent vector of the curve at time t coincides with the value of the vector field v at position $\gamma(t)$. The Fréchet derivative of v_0 at $\gamma_0(t)$ is the familiar Jacobian matrix. Writing the vector field as $v_0 = (v^1, \ldots, v^d)$ where each coordinate v^i is a map from $\mathcal{U}$ to $\mathbb{R}$, define the Jacobian at time t to be

$$J_0(t) := Dv_0(\gamma_0(t)) = \begin{pmatrix} \frac{\partial v^1}{dx_1}(\gamma_0(t)) & \cdots & \frac{\partial v^1}{\partial x_d}(\gamma_0(t)) \\ \cdots & \ddots & \cdots \\ \frac{\partial v^d}{\partial x_1}(\gamma_0(t)) & \cdots & \frac{\partial v^d}{\partial x_d}(\gamma_0(t)) \end{pmatrix}.$$

With this notation in place, Theorem 1 in conjunction with standard theory on systems of linear differential equations reviewed in Appendix A yields:

Corollary 1. Suppose the conditions of Theorem 1 are satisfied. If the state space is $\mathcal{X} = \mathbb{R}^d$, then the limiting distribution ξ in (2.2) can be equivalently expressed as $\xi(t) = -V(t)\int_0^t V^{-1}(\tau)(\mathbb{G} \circ \gamma_0)(\tau)\,d\tau$, where the matrix $V(t) \in \mathbb{R}^{d\times d}$ is the solution of the deterministic linear equation

$$\dot{V} = J_0(t)V \quad \text{and} \quad V(0) = I_d.$$

If $J = J_0$ commutes, i.e. $J(t_1)J(t_2) = J(t_2)J(t_1)$ for all $t_1, t_2 \in I$, then closed-form solutions for V and its inverse exist: $V(t) = \exp\left\{\int_0^t J(\tau)\,d\tau\right\}$ and $V^{-1}(t) = \exp\left\{-\int_0^t J(\tau)\,d\tau\right\}$. In the more typical situation where J does not commute, an approximation such as the Magnus expansion (Blanes et al., 2009) can be used to estimate V, as discussed in Appendix A.

2.3. Asymptotics in Regime 2

When $\hat{v}_n$ is a nonparametric estimate depending on some regularization parameter h, condition (2.1) in Theorem 1 requiring the weak convergence of $r_n(\hat{v}_n - v_0)$ may not be satisfied. For instance, Theorem 2.2.3 in Bierens (1987) shows that when $\hat{v}_n$ is the standard Nadaraya-Watson kernel regression function estimator, the sequence $\sqrt{nh_n^d}(\hat{v}_n(x) - v(x))$ for distinct points $x_1, \ldots, x_k$ in $\mathbb{R}^d$ is

asymptotically independent. Thus the process $\hat{v}_n$ cannot have a tight limiting distribution.

Our interest in nonparametric estimators $\hat{v}_n$ leads us to seek an alternative derivation of the limiting distribution of the empirical integral curve $\hat{\gamma}_n$ under a relaxation of condition (2.1). We will discuss at the end of this section that we are likely, though, to lose the possibility of valid bootstrap procedures in the process. To remedy this, we point a way forward for nonparametric $\hat{v}_n$ by adopting a scale-space perspective in Section 4, redefining the targets to be the resolution-h vector field and corresponding resolution-h integral curve.

Going back to the second regime, consider the following assumptions:

Assumption A2. $Dv_0 : \mathcal{U} \to L(\mathcal{X}, \mathcal{X})$ is uniformly continuous.

Assumption A3. The estimates $\hat{v}_n$ and $D\hat{v}_n$ are uniformly consistent:

$$\sup_{x \in \mathcal{U}} \left\| \hat{v}_n(x) - v_0(x) \right\|_{\mathcal{X}} = o_P(1)$$

and $\sup_{x \in \mathcal{U}} \left\| D\hat{v}_n(x) - Dv_0(x) \right\|_{op} = o_P(1)$.

When $\mathcal{X} = \mathbb{R}^d$, Assumption A2 is satisfied if v_0 is uniformly differentiable. While Assumption A2 is largely made for technical reasons, Assumption A3 is reasonable, asking that $\hat{v}_n$ and the plug-in estimator for the derivative, $D\hat{v}_n$, are consistent estimators for their respective theoretical counterparts.

Theorem 2. Suppose A1 – A3 hold and $r_n \hat{\eta}_n \xrightarrow{d} \eta$ in $(C(I, \mathcal{X}), \|\cdot\|_\infty)$ for some sequence $r_n \geq 0$, $r_n \to \infty$ where $\hat{\eta}_n(t) = \int_0^t [\hat{v}_n(\gamma_0(s)) - v_0(\gamma_0(s))]\, ds$. Then it holds that $r_n(\hat{\gamma}_n - \gamma_0) \xrightarrow{d} \xi$ in $(C(I, \mathcal{U}), \|\cdot\|_\infty)$ where ξ solves the stochastic differential equation

$$d\xi(t) = Dv_0(\gamma_0(t))\xi(t)\, dt + d\eta(t) \quad \text{and} \quad \xi(0) = 0. \tag{2.3}$$

Theorem 2 can in fact be used to recover Theorem 1. To see this, note that if (2.1) holds, then $r_n \hat{\eta}_n(t) \xrightarrow{d} \eta(t) = \int_0^t (\mathbb{G}(\gamma_0(s))\, ds$. Since the derivative of η with respect to t is simply $\mathbb{G}(\gamma_0(t))$, both stochastic differential equations (2.2) and (2.3) characterize the same limiting process ξ.

When condition (2.1) does not hold, a valid bootstrap scheme for the integral curve is unlikely to be available. Heuristically speaking, since ξ results from a smooth operator applied to η, a valid bootstrap procedure for ξ would necessitate being able to bootstrap η. However, $\hat{\eta}_n$ is itself a smooth functional of $r_n(\hat{v}_n - v_0)$. It is difficult to envision a valid bootstrap procedure for η if $\hat{v}_n$ does not have a tight weak limit. Thus, we opt to work under the setting of Theorem 1 which readily affords valid bootstrap procedures, the subject of the next section. We shall this naturally leads to the consideration of a scale-space perspective for nonparametric vector field estimators, and related uniform-in-bandwidth bootstrap procedures.

3. Bootstrapping the integral curve

This section concerns asymptotic validity of bootstrap procedures (Efron, 1979, 1982) for the integral curve under the first regime, i.e. the setting of Theorem 1. In particular, we demonstrate bootstrap consistency for two standard bootstrap weights and vector field estimators that can be written as empirical processes relative to Donsker classes. Let $\{X_1, \ldots, X_n\} \subset \mathcal{U}$ be a random sample from a probability measure P on $\mathcal{U}$. The empirical measure is $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$ where δ_x is the measure that assigns mass 1 at x and zero elsewhere. We write, for a measurable function $f : \mathcal{U} \to \mathbb{R}$, $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(X_i)$.

Using the notation of Kosorok (2008), we use $\mathbb{P}_n^\circ$, where $f \mapsto \mathbb{P}_n^\circ f = \frac{1}{n} \sum_{i=1}^n W_{ni} f(X_i)$, to denote either of two bootstrapped empirical processes. The first is the *multinomial* bootstrapped empirical process where $(W_{n1}, \ldots, W_{nn})$ is a multinomial vector with probabilities $(1/n, \ldots, 1/n)$ independent of the data sequence $(X_1, \ldots, X_n)$. The second is the *multiplier* bootstrapped empirical process where $W_{ni} = (\psi_i/\bar{\psi})$ for i.i.d. positive weights $\psi_1, \ldots, \psi_n$, independent of the data sequence $(X_1, \ldots, X_n)$, with $0 < \mu = E\psi_1 < \infty$, $0 < \tau^2 = var(\psi_1) < \infty$, which satisfy $\|\psi_1\|_{2,1} = \int_0^\infty \sqrt{P(|\psi_1| > x)}\, dx < \infty$, and $\bar{\psi} = \frac{1}{n} \sum_{i=1}^n \psi_i$. The binomial and multiplier bootstrapped empirical processes were first studied by Praestgaard and Wellner (1993).

To rigorously define bootstrap consistency, we need to define the weak convergence of the conditional limit laws of bootstraps. For a metric space $(\mathbb{D}, d)$, let BL_1 denote the space of functions $f : \mathbb{D} \to \mathbb{R}$ with Lipschitz norm bounded by 1. If $\hat{Y}_n$ is a sequence of processes in $\mathbb{D}$ involving random weights W, the notation $\hat{Y}_n \xrightarrow[W]{p} Y$ for some tight process Y in $\mathbb{D}$ means $\sup_{h \in BL_1} |E_W h(\hat{Y}_n) - Eh(Y)| \xrightarrow{p} 0$. The subscript in the expectation indicates conditional expectation over the weights W given the remaining data. The result below is an application of theory on bootstrapped empirical processes presented in van der Vaart and Wellner (1996).

Theorem 3. Suppose $\mathcal{F} = \{f_x : x \in \mathcal{U}\}$ is a Donsker class of measurable functions $f_x : \mathcal{U} \to \mathbb{R}$, giving rise to vector fields

$$v_0(x) = Pf_x(X), \quad \hat{v}_n(x) = \mathbb{P}_n f_x(X), \quad \text{and} \quad \hat{v}_n^\circ(x) = \mathbb{P}_n^\circ f_x(X).$$

Let $c = \mu/\tau$ if $\mathbb{P}_n^\circ$ is the multiplier bootstrap and $c = 1$ if $\mathbb{P}_n^\circ$ is the multinomial bootstrap. If v_0 satisfies Assumption A1, then $\sqrt{n}(\hat{\gamma}_n - \gamma_0) \xrightarrow{d} \xi$ and $\sqrt{n}c(\hat{\gamma}_n^\circ - \hat{\gamma}_n) \xrightarrow[W]{p} \xi$ where

$$d\xi(t) = Dv_0(\gamma_0(t))\xi(t)\, dt + \mathbb{B}(\gamma_0(t)) \quad \text{and} \quad \xi(0) = 0,$$

and $\mathbb{B}$ is the standard Brownian bridge in in $l^\infty(\mathcal{U}, \mathcal{X})$.

4. Uniform-in-bandwidth asymptotics

In this section, we adopt a scale-space view (Chaudhuri and Marron, 2000). This bypasses the difficulty encountered for nonparametric estimators that fail

to verify (2.1), and thus not captured by Theorem 1. Specifically, rather than focusing on the one true underlying vector field v_0, we focus on $E\hat{v}_{n,h}(x)$ simultaneously for a range of h values in $H \subset (0,\infty)$. Certainly, $E\hat{v}_{n,h}(x)$ may be biased for v_0, but examining it may still be attractive for data analysis as different levels of smoothing invariably reveal different aspects of the truth.

Technical matters aside, there are first-principle advantages to the scale-space perspective. For example, in certain settings, v_0 may not be well defined to begin with, see e.g. Rinaldo and Wasserman (2010) and Chen, Genovese and Wasserman (2017a). In the gradient line example of Section 5.1, we further discuss this phenomenon in the context of functional density estimation. There is yet another advantage in the scale-space viewpoint: in many practical situations, a tuning parameter is chosen in a data-driven way or by means of "data snooping". This induces an additional layer of sampling variation that needs to be accounted for by bootstrap procedures that are valid uniformly in the choice of a non-trivial bandwidth (an issue discussed further after the statement of our result).

Let $v \in C(\mathcal{U} \times H, \mathcal{X})$ be the vector field of interest and define the evolution equation

$$\dot{\gamma}(t,h) = v(\gamma(t,h),h) \quad \text{and} \quad \gamma(0,h) = x_0 \quad \forall h \in H. \tag{4.1}$$

We assume throughout that v is locally Lipschitz on $\mathcal{U}\times H$ so that the solution of (4.1) is well-defined, i.e. there exists some $D = \{(t,h) \in \mathbb{R}\times H : t \in J_h \subset \mathbb{R}\}$ such that there is a unique integral curve $\gamma \in C^1(D,\mathcal{U})$ that solves (4.1). Throughout, fix the initial condition x_0 and the set D. We are now ready to present the uniform-in-bandwidth analogue of Theorem 1.

Assumption B1. The vector field v is Fréchet differentiable at $(\gamma(t,h),h)$ for all $(t,h) \in D$.

Theorem 4. Suppose B1 holds. In addition, suppose the vector field estimate $\hat{v}_n$ is such that

$$r_n(\hat{v}_n - v) \xrightarrow{d} \mathbb{G} \quad \text{in } (l^\infty(\mathcal{U}\times H,\mathcal{X}),\|\cdot\|_\infty)$$

for some sequence $r_n \geq 0, r_n \to \infty$ and tight process $\mathbb{G}$ that satisfies

$$\big\|\mathbb{G}(\gamma_n(t,h),h) - \mathbb{G}(\gamma(t,h),h)\big\|_{\mathcal{X}} \xrightarrow{\text{a.s.}} 0 \quad \forall (t,h) \in D \quad \text{as } \gamma_n \xrightarrow{\text{a.s.}} \gamma.$$

Then it holds that $r_n(\hat{\gamma}_n - \gamma) \xrightarrow{d} \xi$ in $(C^1(D,\mathcal{U}),\|\cdot\|_{C^1})$ where

$$d\xi(t,h) = Dv(\gamma(t,h),h)\xi(t,h)\,dt + \mathbb{G}(\gamma(t,h),h), \quad \text{and} \quad \xi(0,h) = 0 \quad \forall h \in H.$$

If $\mathbb{G}$ is a Gaussian process, then so is ξ.

The next result is analogous to Theorem 3, showing that vector fields which can be written as empirical processes of Donsker classes can be bootstrapped validly, uniform in bandwidth.

Theorem 5. Suppose $\mathcal{F} = \{f_{x,h} : x \in \mathcal{U}, h \in H\}$ is a Donsker class of measurable functions $f_{x,h} : \mathcal{U} \to \mathbb{R}$ giving rise to vector fields

$$v(x,h) = Pf_{x,h}(X), \quad \hat{v}_n(x,h) = \mathbb{P}_n f_{x,h}(X), \quad \text{and} \quad \hat{v}_n^\circ(x,h) = \mathbb{P}_n^\circ f_{x,h}(X).$$

Let $c = \mu/\tau$ if $\mathbb{P}_n^\circ$ is the multiplier bootstrap and $c = 1$ if $\mathbb{P}_n^\circ$ is the multinomial bootstrap. If v satisfies B1, then $\sqrt{n}(\hat{\gamma}_n - \gamma) \xrightarrow{d} \xi$ and $\sqrt{n}c(\hat{\gamma}_n^\circ - \hat{\gamma}_n) \xrightarrow[W]{p} \xi$ in $(C^1(D,\mathcal{U}), \|\cdot\|_{C^1})$ where

$$d\xi(t,h) = Dv(\gamma(t,h),h)\xi(t,h)\,dt + \mathbb{B}(\gamma(t,h),h) \quad \text{and} \quad \xi(0,h) = 0 \quad \forall h \in H,$$

and $\mathbb{B}$ is the standard Brownian bridge in in $l^\infty(\mathcal{U} \times H, \mathcal{X})$.

5. Applications

We now demonstrate the use of our results in a series of concrete situations. The first three concern finite-dimensional evolution equations. In Section 5.1, the gradient line of a probability density function is formulated as a solution to a finite-dimensional evolution equation. In Section 5.2, white matter fiber tracts in the brain are modeled as integral curves of the eigenvector field derived from the underlying water molecule diffusion tensor field. In Section 5.3, we consider a variation of the local principal curve, formulated as the integral curve of a local covariance field, and demonstrate it on a traffic flow dataset. We conclude with an infinite-dimensional evolution equation related to anisotropic heat diffusion in Section 5.4.

Based on the implications of Theorem 5, Algorithm 1 details the general construction of a bootstrapped confidence region for γ that is uniform simultaneously in time and bandwidth. This algorithm will be used throughout the first three examples in this section. Note that Theorem 3 can be similarly availed for a confidence region construction that is uniform only in time. A similar algorithm can be found in Chen, Genovese and Wasserman (2015) for constructing a confidence region using the bootstrap that is uniform over the one-dimensional ridge set of interest though not uniform for the smoothing bandwidth.

Algorithm 1 Confidence region for γ uniform in time and bandwidth.

Require: Data $X_1, \ldots, X_n \in \mathcal{U}$, set of bandwidths H, significance level α
1: **for each** $h \in H$ **do**
2: Estimate the integral curve from $\{X_1, \ldots, X_n\}$; denote this by $\hat{\gamma}_h$
3: Generate bootstrap samples $\{X_1^{*(b)}, \ldots, X_n^{*(b)}\}$ for $b = 1, \ldots, B$ using either the multinomial or multiplier bootstrap.
4: For each bootstrap sample, estimate the integral curve, call this $\hat{\gamma}_h^{(b)}$.
5: **end for**
6: For $b = 1, \ldots, B$, calculate $z_b = \sup_{0 \le t \le T, h \in H} \left\| \hat{\gamma}_h^{(b)}(t) - \hat{\gamma}_h(t) \right\|$.
7: Let $\hat{c}_\alpha$ be the α-upper quantile of $z_1, \ldots, z_B$.
8: **return** The set $\{u \in \mathcal{U} : \sup_{0 \le t \le T, h \in H} \|\hat{\gamma}_h(t) - u\| < \hat{c}_\alpha\}$

5.1. Gradient lines

Call γ a gradient line of a probability density function $f : \mathbb{R}^d \to \mathbb{R}^+$ if it is the solution to the evolution equation

$$\dot{\gamma}(t) = \nabla f(\gamma(t)) \quad \text{and} \quad \gamma(0) = x_0 \in \mathbb{R}^d. \tag{5.1}$$

Gradient lines are of interest in applications such as modal clustering (Chen, Genovese and Wasserman, 2016), filament estimation (Genovese et al., 2009), and Morse-Smale complex estimation (Chen, Genovese and Wasserman, 2017b). They play an important role in the computer vision literature where the ubiquitous mean-shift algorithm can be seen as a discrete approximation to the continuous system in (5.1) (see Arias-Castro, Mason and Pelletier (2016) for more details).

Gradient lines are typically estimated using the plug-in principle. Consider the kernel density estimator given by $\hat{f}_n(x,h) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)$ which is based on a random sample $X_1,\dots,X_n$ from the density f_0, where K is a kernel function with bandwidth $h \in H \subset (0,\infty)$. For simplicity, assume K is the standard multivariate Gaussian density. Following the scale-space view of Chaudhuri and Marron (2000), define the resolution-h theoretical counterpart of $\hat{f}_n(x,h)$ to be $f(x,h) \equiv E\hat{f}_n(x,h) = \frac{1}{h^d}EK\left(\frac{x-X_i}{h}\right)$. Let $\hat{f}_n^\circ(x,h) = \frac{1}{nh^d}\sum_{i=1}^n W_i K\left(\frac{x-X_i}{h}\right)$ be the bootstrapped kernel density estimator where the weights W_i correspond either to the multinomial bootstrap or the multiplier bootstrap defined in Section 3.

Before moving on, a discussion is warranted regarding this scale-space view. In its absence, it would be natural to consider the sequence $\sqrt{nh_n^d}(\hat{f}_n(x,h_n) - f_0(x))$, which converges in distribution for fixed x as $n \to \infty$ and $h_n \to 0$ at an appropriate rate. However, there is no equivalent statement for $\hat{f}_n(x,h_n)$ considered as a process over x. This is because for a distinct collection $\{x_1,\dots,x_k\}$, the Cramer-Wold device gives the joint convergence of

$$\sqrt{nh_n^d}\left[\hat{f}_n(x_1,h_n) - f_0(x_1),\dots,\hat{f}_n(x_k,h_n) - f_0(x_k)\right]^T$$

to a multivariate normal with a diagonal covariance matrix. Thus $\sqrt{nh_n^d}(\hat{f}_n(x,h_n) - f_0(x))$ cannot converge weakly over x to a tight limit, which would preclude the application of Theorem 3 to establish bootstrap consistency.

There is further reason to consider $f(x,h)$, rather than f_0, as the theoretical counterpart to $\hat{f}_n(x,h)$. In certain state spaces $\mathcal{X}$, it may not be possible to define a proper probability density function, e.g. when $\mathcal{X}$ is a space of functions as in functional data analysis. A common way to overcome this is to introduce a surrogate density or pseudo-density (Hall and Heckman, 2002; Delaigle and Hall, 2010; Ciollaro, Genovese and Wang, 2016), in much the same spirit as $f(x,h)$ is defined above.

Fix the initial condition x_0, and let the gradient lines of $f(x,h)$, $\hat{f}_n(x,h)$, and $\hat{f}_n^\circ(x,h)$ be denoted, respectively, $\gamma(t,h), \hat{\gamma}_n(t,h)$, and $\hat{\gamma}_n^\circ(t,h)$. Assume throughout that the vector field $\nabla f(x,h)$ is locally Lipschitz so that there exists some

$D = \{(t,h) \in \mathbb{R} \times H : t \in J_h \subset \mathbb{R}\}$ such that its gradient line γ is well-defined. The following justifies bootstrapping for the resolution-h gradient line. Note the requirement below that H be a compact subset of $(0,\infty)$ excludes the case $h \to 0$.

Theorem 6. Let $\mathcal{U}$ and H be compact subsets of $\mathbb{R}^d$ and $(0,\infty)$, respectively. If $f(x,h)$ is twice continuously differentiable, then $\sqrt{n}(\hat{\gamma}_n - \gamma)$ and $\sqrt{n}(\hat{\gamma}_n^\circ - \hat{\gamma}_n)$ converge weakly over $(t,h) \in D$ to the same mean-zero $\mathcal{U}$-valued Gaussian process.

We assess via a simulation study the finite-sample bootstrap coverage for gradient lines arising from a normal distribution. Consider a random sample $\{X_1,\ldots,X_n\}$ from the bivariate Gaussian density $f_0 : \mathbb{R}^2 \to \mathbb{R}^+$ with mean 0 and $\sigma_x^2 = 4, \sigma_y^2 = 9, \rho = 0.9$, i.e. covariance matrix $[4, 5.35; 5.35, 9]$. We calculate $\nabla \hat{f}_n(x,h) = \frac{1}{nh^2}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)\frac{X_i - x}{h^2}$. The theoretical counterpart to this, $\nabla f(x,h) = \nabla E\frac{1}{h^2}K((x-X_i)/h)$, was approximated using numerical integration in Matlab.

Fix the initial value $x_0 = (0.4, 0.5)$ and $I = [0,T]$ where $T = 0.1$ is the total length of flow. Figure 1 shows the gradient field ∇f_0 overlaid with its gradient line γ_0, as well as the resolution-h gradient lines $\gamma(t,h)$ for $h \in H = \{0.5, 0.6, \ldots, 0.9, 1\}$. We implement Algorithm 1 with the multinomial bootstrap. The number of bootstrap replications B was set to 200. We assessed coverage probabilities using 200 Monte Carlo iterations. The performance for different sample sizes n at desired 95% coverage is shown in Figure 2. We see that the nominal coverage is attained as n increases. There is sign of under-coverage for $n = 50$ but already for $n = 100$ the coverage becomes adequate.

5.2. Noisy principal eigenvector fields

Diffusion tensor imaging is a magnetic resonance imaging technique that can be used to map fiber tract structures in the brain (Basser, Mattiello and LeBihan, 1994). A 3×3 symmetric positive-definite matrix is measured at a location in $\mathbb{R}^3$ that captures the spatial covariation of water diffusion through tissue. The integral curves of the associated principal eigenvector field provide a continuous description of the diffusion tensor field. In fact, they form the basis of many tractography methods in diffusion tensor imaging (Mukherjee et al., 2008).

Since the diffusion tensor field is measured empirically with noise, so too is the principal eigenvector field derived from it. As in Koltchinskii, Sakhanenko and Cai (2007), we consider a model where the underlying principal eigenvector field v_0 is observed according to $V_i = v_0(X_i) + \epsilon_i, i = 1,\ldots,n$, where X_i is uniformly distributed in a bounded open set of $\mathbb{R}^d$, and ϵ_i's are mean-zero bounded random errors. We will employ the standard kernel regression estimator $\hat{v}_n(x,h) = \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)V_i$ to generate a plug-in estimate for the integral curve. To simplify matters, suppose K is the standard multivariate Gaussian density.

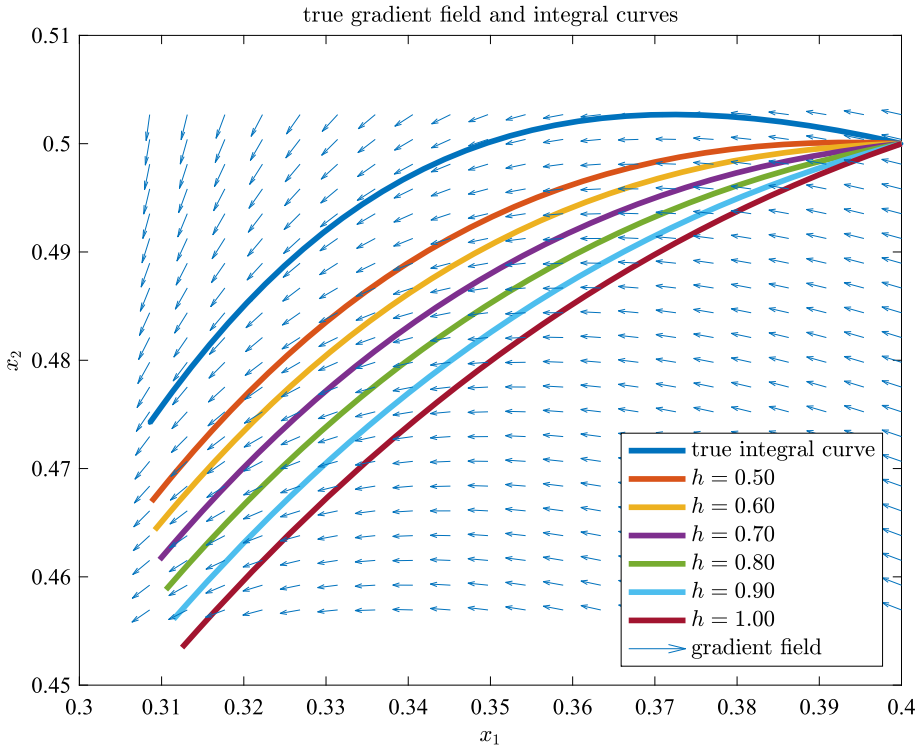

FIG 1. *Gradient field of the density f_0 corresponding to the bivariate Gaussian $N_2(0, [4, 5.35; 5.35, 9])$, overlaid with the gradient lines γ_0 and $\gamma(t, h)$ for various values of the bandwidth h. All gradient lines start at $(0.4, 0.5)$ and flow for $T = 0.1$ length of time.*

Our treatment will now begin to diverge from that of Koltchinskii, Sakhanenko and Cai (2007). Their analysis utilizes a result similar to Theorem 2, but for a Euclidean state space, to derive the limiting distribution of the empirical integral curve. Importantly, this result concerns γ_0, the integral curve of v_0. We will instead consider the scale-space perspective and make inference for a resolution-h integral curve. This is so that we can justify the bootstrapping procedure which would otherwise be difficult to ascertain if γ_0 were the object of interest instead. Define the theoretical counterpart of $\hat{v}_n(x, h)$ to be $v(x, h) \equiv E\hat{v}_n(x, h) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right) E(V_i|X_i)$. Let the bootstrapped estimator be $\hat{v}_n^\circ(x, h) = \frac{1}{nh^d} \sum_{i=1}^n W_i K\left(\frac{x-X_i}{h}\right) V_i$ where the weights W_i correspond either to the multinomial bootstrap or the multiplier bootstrap defined in Section 3.

Fix the initial condition x_0, and let the integral curves of the eigenvector fields $v(x, h)$, $\hat{v}_n(x, h)$, and $\hat{v}_n^\circ(x, h)$ be denoted, respectively, $\gamma(t, h), \hat{\gamma}_n(t, h)$, and $\hat{\gamma}_n^\circ(t, h)$. Throughout, assume v is locally Lipschitz so that there exists some $D = \{(t, h) \in \mathbb{R} \times H : t \in J_h \subset \mathbb{R}\}$ for which γ is well-defined. The

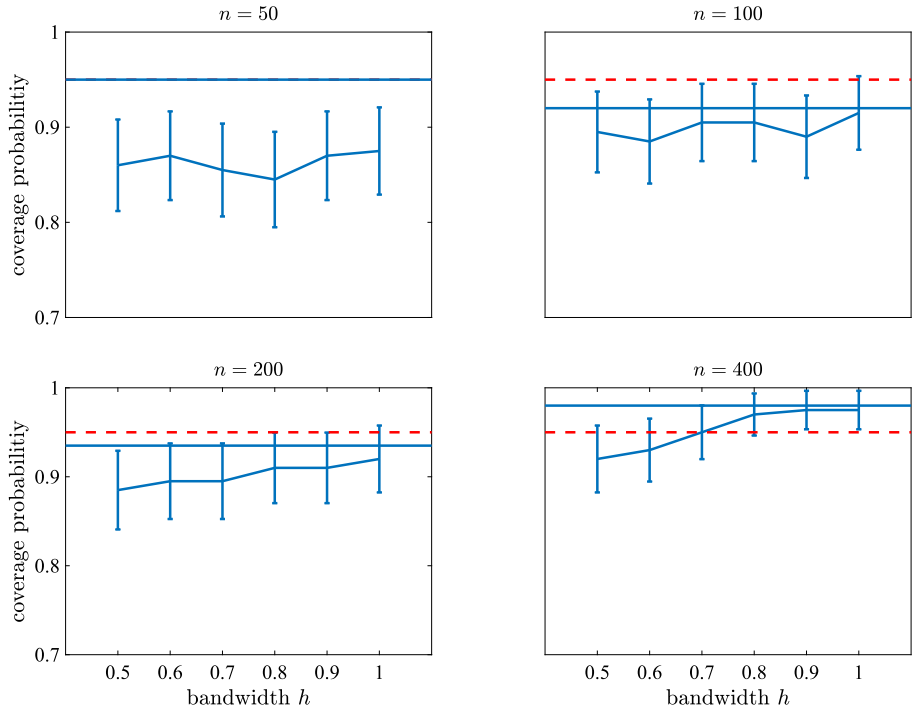

FIG 2. *Bootstrap performance for gradient line. Dashed horizontal line is the desired 95% coverage probability. Solid horizontal line is the coverage probability for bootstrapping $\gamma(t,h)$ over t and h simultaneously. The coverage for $\gamma(t,h)$ at fixed h is also displayed with Monte Carlo error (± 1 standard error) bars.*

following result justifies performing inference for the resolution-h integral curve via bootstrapping.

Theorem 7. Let $\mathcal{U}$ and H be compact subsets of $\mathbb{R}^d$ and $(0,\infty)$, respectively. If $v(x,h)$ is twice continuously differentiable, then $\sqrt{n}(\hat{\gamma}_n - \gamma)$ and $\sqrt{n}(\hat{\gamma}_n^\circ - \hat{\gamma}_n)$ converge weakly over $(t,h) \in D$ to the same mean-zero $\mathcal{U}$-valued Gaussian process.

Consider the following example from Section 4.2 in Koltchinskii, Sakhanenko and Cai (2007). We have noisy observations of a circular vector field v_0 on $\mathbb{R}^2$ given by $v_0(x_1,x_2) = (-x_2/\|x\|, x_1/\|x\|)$ and random errors ϵ_i distributed $\frac{1}{2}N(0,I_2)$. Fix the initial value at $x_0 = (3,0)$ and $I = [0,T]$ where $T = 10$ is the total length of flow. The top left panel of Figure 3 shows the vector field v_0 overlaid with the integral curve γ_0. The top right panel shows one realization of a corrupted vector field. The bottom row of Figure 3 displays smoothed vector fields $\hat{v}_n(x,h)$ for $h = 0.5$ and $h = 1$ and corresponding empirical integral curves $\hat{\gamma}_n(t,h)$.

Consider $H = \{0.1, 0.2, \ldots, 0.9, 1\}$ for the set of bandwidths. We implement Algorithm 1 with the multinomial bootstrap. The number of bootstrap replica-

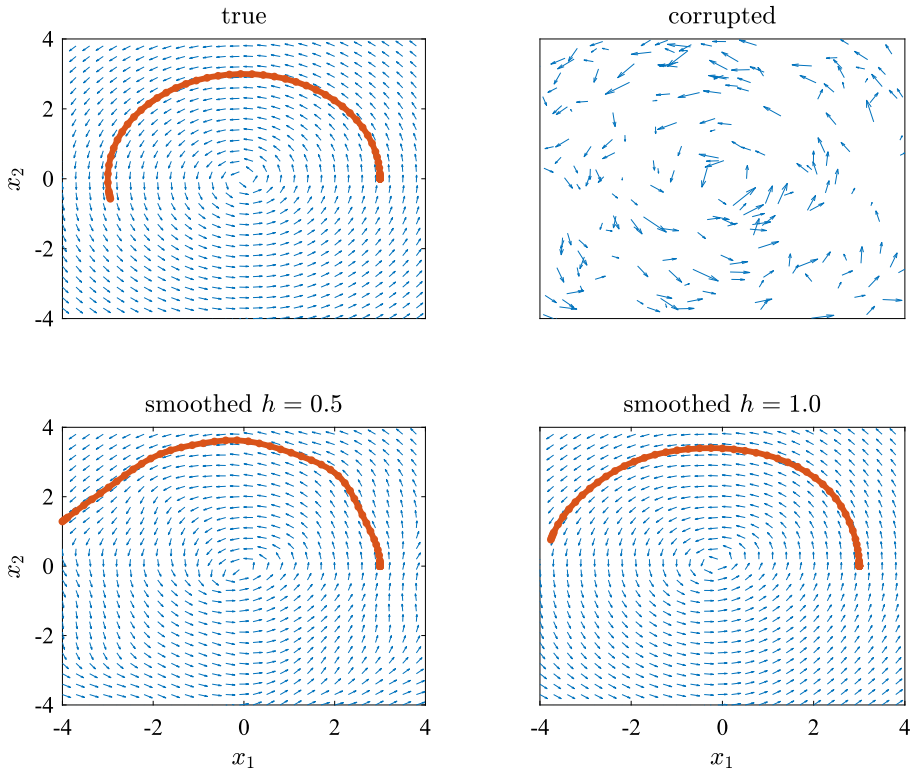

FIG 3. *The top row displays the circular vector field* $v_0(x_1, x_2) = (-x_2/\|x\|, x_1/\|x\|)$ *overlaid with the true integral curve* γ_0 *(left), and corrupted vector field sampled at 200 points uniformly from the* $[-4, 4]^2$ *grid with* $\frac{1}{2}N(0, I_2)$ *noise (right). The bottom row displays smoothed vector fields for two different bandwidths* $h = 0.5$ *and* $h = 1$*. All integral curves start at* $(3, 0)$ *and flow for* $T = 10$ *length of time.*

tions B was set to 200 and the number of Monte Carlo iterations to 200. The performance for different sample sizes n at desired 95% coverage is shown in Figure 4. We see that the nominal coverage is attained as sample size increases and the bootstrap method is not overly conservative until $n = 400$.

5.3. Local principal flow

Hastie and Stuetzle (1989) introduced the concept of principal curves, smooth curves that pass through the "middle" of a multivariate point cloud, to describe data exhibiting nonlinear variation. Many variations on the principal curve has since followed including the concept of local principal curves (sometimes people use the name ridge (Ozertem and Erdogmus, 2011)). Top-down methods for local principal curves start with the first principal component and successively bend it (Einbeck, Evers and Bailer-Jones, 2008) while bottom-up methods are

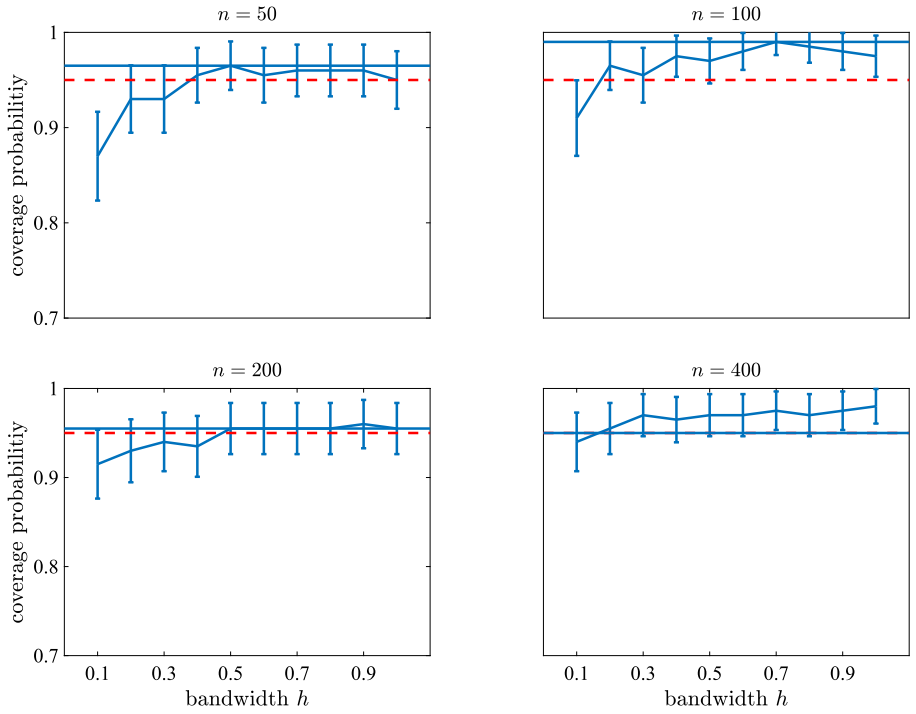

FIG 4. *Bootstrap performance for noisy principal eigenvector field. Dashed horizontal line is the desired 95% coverage probability. Solid horizontal line is the coverage probability for bootstrapping $\gamma(t,h)$ over t and h simultaneously. The coverage for $\gamma(t,h)$ at fixed h is also displayed with Monte Carlo error (± 1 standard error) bars.*

exemplified by the the local principal curve proposed in Einbeck, Tutz and Evers (2005) which we describe below.

The local principal curve of Einbeck, Tutz and Evers (2005) is constructed point by point using only local information at each iteration. Suppose we have a multivariate point cloud $\{X_1,\dots,X_n\} \subset \mathbb{R}^d$. Starting from an initial point, a local center of mass μ^x is calculated. Next a local covariance matrix centered at μ^x is formed: $\frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)(\mu^x - X_i)^T(\mu^x - X_i)$, for some kernel function K and bandwidth h. From this, the principal eigenvector is extracted. We move μ^x in the direction of said eigenvector by some step size. This is repeated until μ^x stops changing. The local principal curve is comprised of this series of μ^x.

In this illustration, we show how an adaptation of the local principal curve of Einbeck, Tutz and Evers (2005) can be formulated as a solution to an evolution equation. Rather than centering the local covariance matrix at a local center of mass, consider for each point $x \in \mathbb{R}^d$ a local covariance matrix directly centered there: $\hat{C}_n(x,h) = \frac{1}{n}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)(x - X_i)^T(x - X_i)$. Let $\hat{v}_n(x,h)$ be the associated principal eigenvector of $\hat{C}_n(x,h)$. The corresponding

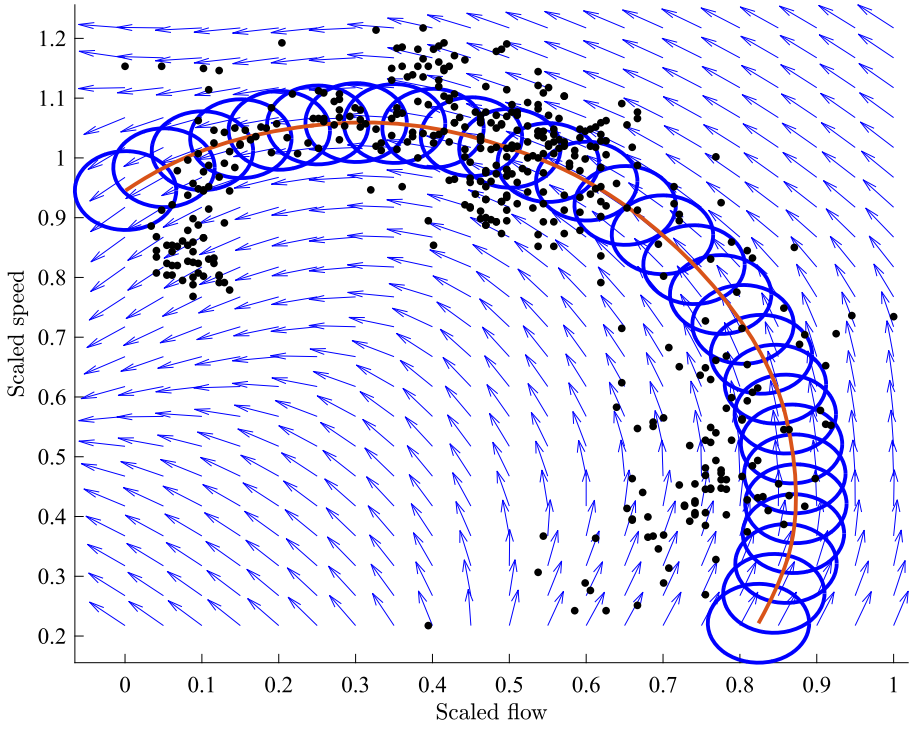

FIG 5. *The local principal flow illustrated on a speed flow dataset. Black dots are original data points (speed, flow) after scaling. Displayed is the principal eigenvector field of the smoothed ($h = 0.22$) local covariance field. The initial condition was chosen to be* $(0.35957, 1.0182)$ *and the integral curve was then propagated in both directions. Bootstrapped 95% confidence circles, uniform over time t, are indicated by blue circles.*

empirical integral curve $\hat{\gamma}_n$ can be interpreted as an infinitesimal local principal curve-like object, which we will call a *local principal flow*, so named for its similarity to the (global) principal flow introduced in Panaretos, Pham and Yao (2014). Define the scale-space theoretical counterpart of $\hat{C}_n(x,h)$ to be $C(x,h) \equiv EK\left(\frac{x-X}{h}\right)(x-X)^T(x-X)$. Let $\hat{C}_n^\circ(x,h)$ be the bootstrapped local covariance matrix, for either the multinomial or multiplier bootstrap defined in Section 3. Denote by γ and $\hat{\gamma}_n^\circ$ the associated local principal flows of C and $\hat{C}_n^\circ$, respectively.

We fit the local principal flow for a speed-flow diagram obtained from the R package LPCM (Einbeck and Dwyer, 2011). The data consists of 444 observations on speed and flow recorded from 9th of July 2007, 9am, to 10th of July 2007, 10pm, on a particular lane of a Californian freeway. The nonlinear variation of the data is apparent from Figure 5. Rather than performing a standard principal component analysis, we fit the local principal flow for K being the standard bivariate Gaussian density and bandwidth $h = 0.22$. The multinomial

bootstrap (with $B = 1000$) was used to find an approximate 95% confidence region for the local principal flow $\gamma(t, h)$, uniform in t. The confidence region appears as blue circles in Figure 5.

5.4. Heat flows

Now we consider an evolution equation on an infinite-dimensional state space, corresponding to a partial differential equation. The motivation for this example arises from novel tractography methods in diffusion tensor imaging that are probabilistic, rather than deterministic, in nature. We will examine specifically the tractography method described in Section D of Hageman et al. (2009), based on the anisotropic heat equation,

$$\frac{\partial \gamma(t)(x)}{\partial t} = \nabla \cdot [D \nabla \gamma(t)(x)], \tag{5.2}$$

where D is the diffusion tensor at $x \in \mathbb{R}^d$. In a typical interpretation, $\gamma(t)(x)$ represents the density of a diffusing material at location x and time t.

Let $\mathcal{P}_d$ denote the space of $d \times d$ symmetric positive-definite matrices. Suppose it is only possible to observe, with measurement error, the true diffusion coefficient D_0 at a finite number of positions. Specifically, for i.i.d. observations $D_i \in \mathcal{P}_d$ occurring at random locations X_i uniformly distributed in $\mathcal{U} \subset \mathbb{R}^d$, consider the diffusion coefficient estimator $\hat{D}(x, h) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) D_i$ where K is the standard multivariate Gaussian density and bandwidth $h \in H \subset (0, \infty)$. Define the scale-space theoretical counterpart to be

$$D(x, h) \equiv E\hat{D}(x, h) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) E(D_i | X_i).$$

Next, let the bootstrapped diffusion coefficient be $\hat{D}^\circ(x, h) = \frac{1}{nh^d} \sum_{i=1}^n W_i \times K\left(\frac{x - X_i}{h}\right) D_i$, where the weights W_i correspond either to the multinomial bootstrap or the multiplier bootstrap in Section 3.

Call γ a **heat flow** of the tensor field $Q_h = Q(x, h)$ where $Q : \mathcal{U} \times H \to \mathcal{P}_d$ if γ solves the evolution equation

$$\frac{\partial \gamma(t, h)(x)}{\partial t} = \nabla \cdot [Q_h \nabla \gamma(t, h)(x)], \quad \gamma(0, h) = x_0 \quad \forall h \in H,$$

for initial condition $x_0 \in C^1(\mathcal{U}, \mathbb{R})$. Fix the initial condition x_0, and let the heat flows associated to each of $D_h = D(x, h), \hat{D}_h = \hat{D}(x, h)$, and $\hat{D}_h^\circ = \hat{D}^\circ(x, h)$ be denoted, respectively, $\gamma(t, h), \hat{\gamma}_n(t, h)$ and $\hat{\gamma}_n^\circ(t, h)$. Assume throughout that the vector field D_h associated to the heat equation is locally Lipschitz so that there exists some $G = \{(t, h) \in \mathbb{R} \times H : t \in J_h \subset \mathbb{R}\}$ on which its heat flow is well-defined. The following result justifies performing inference for the resolution-h heat flow via bootstrapping.

Theorem 8. Let $\mathcal{U}$ and H be compact subsets of $\mathbb{R}^d$ and $(0,\infty)$, respectively. If $\mathrm{vec}(D(x,h))$ is twice continuously differentiable, then $\sqrt{n}(\hat{\gamma}_n - \gamma)$ and $\sqrt{n}(\hat{\gamma}_n^\circ - \hat{\gamma}_n)$ converge weakly over G to the same mean-zero $\mathcal{U}$-valued Gaussian process.

6. Future work

We conclude with a discussion on possible future research directions. First, the uniform-in-bandwidth results in this paper could lead to methods of bandwidth selection based on the idea of stability, e.g. in the spirit of Rinaldo and Nugent (2012); Chen et al. (2015); Ciollaro, Genovese and Wang (2016). In these works, a range of bandwidths is considered desirable if the features of interest persist across it.

Second, more refined statements in the current framework may be possible if we focus on specific classes of evolution equations. Gradient flow equations provide one such example. These equations, whose study forms an active research area, naturally appear in many real systems trying to decrease energy (or increase entropy at fixed energy) over time. Gradient flow equations take the form

$$\dot{\gamma} = -\nabla I(\gamma) \quad \text{and} \quad \gamma(0) = x_0,$$

where I is a functional from a metric space X to $\mathbb{R}$. Intuitively, the solution $\gamma(t)$ "flows downhill" in the direction $-\nabla I/\|\nabla I\|$ with velocity proportional to $\|\nabla I\|$.

It is easy to see that the gradient line of Section 5.1 falls under this framework since the vector field is the gradient of a probability density function. That the anisotropic heat equation in Section 5.4 can also be encompassed in the gradient flow framework is truly surprising, however. The results of Lisini (2009) make apparent that the heat flow associated to Equation (5.2) is also the gradient flow of the functional $I(u) = \int_{\mathbb{R}^d} F(u(x))\,dx$ where $F(z) = z\log z$ with respect to the 2-Wasserstein distance between probability measures on the space $\mathbb{R}^d$, endowed with the Riemannian distanced induced by D^{-1}, the inverse of the diffusion tensor.

Finally, future directions of research might extend the current framework to more exotic evolution equations. For instance, time-dependent vector fields are not treated here but are important tools in the visualization field (Theisel et al., 2004). Also, the increasing interest in manifold data leading to applications involving integral curves on manifolds (using the idea of charts (Lee, 2012)) is not yet treatable by the current work.

Appendix A: Linear differential equations in Banach spaces

The review in this section relies heavily on Krein (1971). Consider an equation of the form

$$\dot{u} = A(t)u + f(t), \quad u(0) = u_0 \in E \tag{A.1}$$

where $A(t)$ is a bounded operator in a Banach space E, $f(t)$ is a given function taking value in E, and $u(t)$ is an unknown function also taking value in E. The derivative $\dot{u}$ is understood to be the limit of the difference quotient with respect to the norm of E.

If $A(t)$ and $f(t)$ are continuous (or, more generally, measurable and integrable on every finite interval), then the solution to the homogeneous counterpart to (A.1),

$$\dot{u} = A(t)u, \quad u(0) = u_0 \tag{A.2}$$

exists for any $u_0 \in E$ and is given by $u(t) = U(t,0)u_0$ where $U(t,s)$ is known as the evolution operator of (A.2), and is given by

$$U(t,s) = \mathcal{I} + \int_0^t A(t_1)\,dt_1 + \sum_{n=2}^{\infty} \int_s^t \int_s^{t_n} \cdots \int_s^{t_2} A(t_n)\cdots A(t_1)\,dt_1 \ldots\, dt_n \tag{A.3}$$

where $\mathcal{I}$ is the identity operator. Going back to the non-homogeneous equation in (A.1), its solution is $u(t) = U(t,0)u_0 + \int_0^t U(t,\tau)f(\tau)\,d\tau$.

Example 1. Suppose $E = \mathbb{R}$ and $A(t) = 1$. Then $U(t,0) = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots$ i.e. the power expansion of $\exp(t)$.

Example 2. Suppose $E = \mathbb{R}$ and $A(t) = t$. Then $U(t,0) = 1 + \frac{t^2}{2\cdot 1!} + \frac{t^4}{2\cdot 2!} + \ldots$ i.e. the power expansion of $\exp(t^2/2)$.

These examples suggest that in the Euclidean setting, the solution to the linear differential equation in (A.1) simplifies substantially, which is indeed the case. Suppose $E = \mathbb{R}^d$, then u and f are vector-valued functions. Take the operator $A(t)$ to be a $d \times d$ matrix. First, suppose A commutes, i.e. $A(t_1)A(t_2) = A(t_2)A(t_1)$, the evolution operator in (A.3) simplifies to $U(t,s) = \exp\left\{\int_s^t A(\tau)\,d\tau\right\} \exp\left\{-\int_0^s A(\tau)\,d\tau\right\}$ where, for a matrix A, the matrix exponential $\exp tA$ is formally defined as the convergent power series $\exp tA = I + tA + \frac{t^2A^2}{2!} + \ldots$. The reader is referred to Moler and Van Loan (2003) on various ways to compute the exponential of a matrix.

In general, however, the matrix A will not commute. In that case the solution to (A.1) is given by

$$u(t) = V(t)u_0 + V(t)\int_0^t V^{-1}(\tau)f(\tau)\,d\tau \tag{A.4}$$

where $V(t)$ is the solution to

$$\dot{V} = A(t)V, \quad \text{and} \quad V(0) = I.$$

The Magnus expansion discussed in Appendix A can be used for $V(t)$ whereby we first write V as a matrix exponential $V(t) = \exp\Omega(t)$ where $\Omega(0) = \mathcal{O}$. The Magnus expansion is a series expansion for the matrix in the exponent

$\Omega(t) = \sum_{k=1}^{\infty} \Omega_k(t)$. We refer the reader to Blanes et al. (2009) for derivation details of the expansion. To illustrate, the first three terms of the series are

$$\Omega_1(t) = \int_0^t A(t_1)\, dt_1$$
$$\Omega_2(t) = \frac{1}{2}\int_0^t dt_1 \int_0^{t_1} dt_2 [A(t_1), A(t_2)]$$
$$\Omega_3(t) = \frac{1}{6}\int_0^t dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 ([A(t_1), [A(t_2), A(t_3)]] + [A(t_3), [A(t_2), A(t_1)]])$$

where $[A, B] = AB - BA$ is the matrix commutator of A and B. Thus approximation of the solution in (A.4) via the Magnus expansion is obtained by truncating the series expansion.

Appendix B: Proofs of formal statements

Proof of Theorem 1. Define the parameter space Θ to be a closed subset of $C^1(I, \mathcal{U})$ that contains an open neighborhood of γ_0. Additionally, suppose Θ is such that $\sup_{\theta\in\Theta}\|\theta\|_{C^1} < \infty$ and $\theta(0) = x_0$ for all θ in Θ. We endow Θ with the C^1 norm which makes it a Banach space. Also consider the Banach space $\mathbb{L} = C(I, \mathcal{X})$ equipped with the supremum norm.

Consider the operators Ψ and Ψ_n, treated as elements of $l^\infty(\Theta, \mathbb{L})$, given by $\Psi(\gamma) = \dot\gamma - v_0 \circ \gamma$ and $\Psi_n(\gamma) = \dot\gamma - \hat v_n \circ \gamma$. The integral curve $\gamma_0 \in C^1(I, \mathcal{U})$ of v_0 may now be formulated as a zero of Ψ. Viewing the extraction of the zero from Ψ and Ψ_n as a continuous mapping allows the application of a functional delta method on the process $r_n(\Psi_n - \Psi)$ to derive the limiting distribution of the empirical integral curve. Heuristically, as Ψ_n approaches Ψ in an appropriate sense, so too should their respective roots. We will apply the master Z-estimator result that is Corollary 13.7 of Kosorok (2008) which makes this idea rigorous.

We first check $\Psi_n, \Psi \in l^\infty(\Theta, \mathbb{L})$. We have the following

$$\begin{aligned}
\|\Psi\gamma\|_{\mathbb{L}} &= \|\dot\gamma - v_0 \circ \gamma\|_\infty \\
&= \left\|(\dot\gamma - v_0 \circ \gamma) - (\dot\gamma_0 - v_0 \circ \gamma_0)\right\|_\infty \\
&\le \left\|(\dot\gamma - \dot\gamma_0)\right\|_\infty + k\|\gamma - \gamma_0\|_\infty \\
&\le k\left(\left\|(\dot\gamma - \dot\gamma_0)\right\|_\infty + \|\gamma - \gamma_0\|_\infty\right) \\
&\le k\left(\|\dot\gamma\|_{C^1} + \|\gamma\|_{C^1}\right) < \infty
\end{aligned}$$

where we have, in the third line, used the fact that v_0 is locally Lipschitz and $\gamma, \gamma_0 \in \Theta$, and taken k to be greater than or equal to 1. We can show that $\Psi_n \in l^\infty(\Theta, \mathbb{L})$ in a similar way.

Next, the unique existence of γ_0 is guaranteed by the fact that v_0 is locally Lipschitz. Thus the identifiability condition in Corollary 13.7 of Kosorok (2008) is satisfied. To derive the Fréchet derivative of Ψ at γ_0, write $\Psi(\tilde\gamma) - \Psi(\gamma_0) = \frac{\partial}{\partial t}(\tilde\gamma - \gamma_0) - (Dv_0 \circ \gamma_0)(\tilde\gamma - \gamma_0) + R(\tilde\gamma, \gamma_0)$ where $\tilde\gamma$ is an element of Θ and the

remainder term is $R(\tilde{\gamma},\gamma_0) = (Dv_0\circ\gamma)(\tilde{\gamma}-\gamma_0) - [v_0\circ\tilde{\gamma} - v_0\circ\gamma_0]$. The notation $(Dv\circ\gamma_0)$ is understood to be the map from I to $L(\mathcal{X},\mathcal{X})$. By Assumption A1, we have $\left\|v_0(\tilde{\gamma}(t)) - v_0(\gamma_0(t)) - Dv_0(\gamma_0(t))(\tilde{\gamma}(t)-\gamma_0(t))\right\|_{\mathcal{X}} = o(\left\|\tilde{\gamma}(t)-\gamma_0(t)\right\|_{\mathcal{U}})$. From this it follows that $\left\|R(\tilde{\gamma},\gamma_0)\right\|_\infty / \|\tilde{\gamma}-\gamma_0\|_{C^1} \to 0$ as $\|\tilde{\gamma}-\gamma_0\|_\Theta \to 0$. Hence the Fréchet derivative, $\dot{\Psi}_{\gamma_0}:\Theta\to\mathbb{L}$, is $\dot{\Psi}_{\gamma_0}(h) = \dot{h} - (Dv_0\circ\gamma_0)h$.

We now show that $\dot{\Psi}_{\gamma_0}$ is continuously invertible. First, note that $\dot{\Psi}^{-1}_{\gamma_0}(f)(t)$ is the solution to the following linear differential equation in the Banach space $\mathcal{X}$

$$\dot{u}(t) = (Dv_0\circ\gamma_0)(t)u(t) + f(t), \quad u(0) = 0. \tag{B.1}$$

Since the linear operator $(Dv_0\circ\gamma_0)(t)$ acting in the Banach space $\mathcal{X}$ is the Fréchet derivative of v_0 at $\gamma_0(t)$, it is bounded and continuous. The general theory on linear differential equations reviewed in Appendix A includes (B.1) as a special case. It shows $\dot{\Psi}^{-1}_{\gamma_0}(f)(t) = \int_0^t U(t,\tau)f(\tau)\,d\tau$ where $U(t,\tau)$ is the evolution operator given by

$$\begin{aligned} U(t,\tau) = \mathcal{I} &+ \int_\tau^t (Dv_0\circ\gamma_0)(t_1)\,dt_1 \\ &+ \sum_{j=2}^\infty \int_\tau^t \int_\tau^{t_j} \cdots \int_\tau^{t_2} (Dv_0\circ\gamma_0)(t_j)\cdots(Dv_0\circ\gamma_0)(t_1)\,dt_1\ldots dt_j \end{aligned} \tag{B.2}$$

where $\mathcal{I}$ is the identity operator. That the inverse operator $\dot{\Psi}^{-1}_{\gamma_0}$ is continuous can be seen from this form.

Now we check the conditions regarding the estimating equation Ψ_n. First, we check that $\Psi_n \xrightarrow{p} \Psi$ uniformly in $l^\infty(\Theta,\mathbb{L})$. Asymptotic normality of $\hat{v}_n$ implies $\hat{v}_n \xrightarrow{p} v_0$ uniformly in $l^\infty(\mathcal{U},\mathcal{X})$. It then follows that the restriction $(\hat{v}_n\circ\gamma) \xrightarrow{p} (v_0\circ\gamma)$ uniformly in $l^\infty(I,\mathcal{X})$ and thus $\Psi_n \xrightarrow{p} \Psi$ in $l^\infty(\Theta,\mathbb{L})$. The asymptotic normality of $\hat{v}_n$ gives $r_n(\Psi_n - \Psi) \xrightarrow{d} X$ in $(l^\infty(\Theta,\mathbb{L}),\|\cdot\|_\infty)$ where $X(\gamma) = -(\mathbb{G}\circ\gamma)$. Now all of the conditions of Corollary 13.7 in Kosorok (2008) are satisfied and we have $r_n(\hat{\gamma}_n - \gamma_0) \xrightarrow{d} -\dot{\Psi}^{-1}_{\gamma_0}X(\gamma_0)$, i.e. the desired result in (2.2).

Finally, if $\mathbb{G}$ is a Gaussian process, the process $\mathbb{G}\circ\gamma_0$ is a Gaussian process. This is because all finite dimensional distributions of a Gaussian process are Gaussian. Since $\dot{\Psi}^{-1}_{\gamma_0}$ is continuous and continuous mappings preserve Gaussianity, ξ is itself a Gaussian process if $\mathbb{G}$ is a Gaussian process. □

Remark 1. The differential operator that appears in $\dot{\Psi}_{\gamma_0}$ is typically unbounded between two Banach spaces. For instance, consider the differential operator from $C^1[0,1]$ to $C[0,1]$ where both spaces are endowed with the uniform norm; if $f_n = x^n$, then $\|f_n\|_\infty = 1$ and $\left\|\dot{f}_n\right\|_\infty = n$. However, with our particular choice of Θ and $\mathbb{L}$ and the norms they are equipped with, the differential operator is bounded and thus continuous.

Proof of Theorem 2. This is an adaptation of the proof of Theorem 1 in Koltchinskii, Sakhanenko and Cai (2007) to the Banach space setting. Without loss

of generality, we take the interval $I \subset \mathbb{R}$ to be $I = [0, T]$ for some $T > 0$. Let us begin by representing the integral curves γ_0 and $\hat{\gamma}_n$ in their integral form: $\gamma_0(t) = x_0 + \int_0^t v_0(\gamma_0(s))\, ds$ and $\hat{\gamma}_n(t) = x_0 + \int_0^t \hat{v}_n(\hat{\gamma}_n(s))\, ds$. These integrals are to be interpreted as Bochner integrals (Mikusiński, 1978), whose construction for Banach-space-valued functions is analogous to that of a Lebesgue integral for real-valued functions.

Next, define $\hat{\eta}(t) = \int_0^t [\hat{v}_n(\gamma_0(s)) - v_0(\gamma_0(s))]\, ds$ and

$$\hat{R}(t) = \int_0^t \left[\hat{v}_n(\hat{\gamma}_n(s)) - \hat{v}_n(\gamma_0(s)) - Dv_0(\gamma_0(s))(\hat{\gamma}_n(s) - \gamma_0(s))\right]\, ds.$$

Consider the difference of the empirical integral curve and the population integral curve, decomposed as a sum of two terms: $\hat{\gamma}_n(t) - \gamma_0(t) = \hat{z}(t) + \hat{\delta}(t)$. The first term satisfies $\hat{z}(t) = \int_0^t \hat{\eta}(s)\, ds + \int_0^t Dv_0(\gamma_0(s))\hat{z}(s)\, ds$ and the second term satisfies $\hat{\delta}(t) = \int_0^t Dv_0(\gamma_0(s))\hat{\delta}(s)\, ds + \hat{R}(t)$.

We say $F \in C_0^1(I, \mathcal{X})$ if $F \in C^1(I, \mathcal{X})$ and $F(0) = 0$. Let $\mathcal{D}$ be an operator from $C_0^1(I, \mathcal{X})$ to $C(I, \mathcal{X})$, where for $F \in C_0^1(I, \mathcal{X})$, $u = \mathcal{D}F$ satisfies

$$du(t) = (Dv_0 \circ \gamma_0)(t)u(t)\, dt + dF(t), \quad u(0) = 0. \tag{B.3}$$

Writing $r_n(\hat{\gamma}_n(t) - \gamma_0(t)) = r_n z_n(t) + r_n \delta_n(t)$, we will show that $r_n z_n(t)$ converges weakly to the tight process $\xi(t) = \mathcal{D}\eta(t)$ and $\sup_{0 \le t \le T} \|\delta_n(t)\| = o_p(r_n^{-1})$. These two statements will be established below in Lemmas 1 to 3. Then applying Slutsky's, we get the desired result. □

Lemma 1. Under the conditions of Theorem 2, we have uniform consistency of the empirical integral curve $\sup_{0 \le s \le T} \|\hat{\gamma}_n(s) - \gamma_0(s)\|_{\mathcal{X}} \xrightarrow{p} 0$.

Proof. By definition we have

$$\begin{aligned}\hat{\gamma}_n(t) - \gamma_0(t) &= \int_0^t \left[\hat{v}_n(\hat{\gamma}_n(s)) - v_0(\gamma_0(s))\right]\, ds \\ &= \int_0^t (\hat{v}_n - v_0)(\hat{\gamma}_n(s))\, ds + \int_0^t \left[v_0(\hat{\gamma}_n(s)) - v_0(\gamma_0(s))\right]\, ds.\end{aligned}$$

The local Lipschitz property of v_0 gives us

$$\|\hat{\gamma}_n(t) - \gamma_0(t)\|_{\mathcal{X}} \le T \sup_{x \in \mathcal{U}} \|\hat{v}_n(x) - v_0(x)\|_{\mathcal{X}} + L \int_0^t \|\hat{\gamma}_n(s) - \gamma_0(s)\|_{\mathcal{X}}\, ds.$$

By the Grönwall-Bellman inequality, e.g. Lemma 1 of Chandra (1970), we have

$$\|\hat{\gamma}_n(t) - \gamma_0(t)\|_{\mathcal{X}} \le T \sup_{x \in \mathcal{U}} \|\hat{v}_n(x) - v_0(x)\|_{\mathcal{X}} \exp\left\{\int_0^t L\, ds\right\}.$$

Uniform consistency now follows immediately from Assumption A3. □

Lemma 2. Under the conditions of Theorem 2, $r_n z_n(t) \xrightarrow{d} \xi(t) = \mathcal{D}\eta(t)$ where $\mathcal{D}$ is as in (B.3).

Proof. We proceed by showing that $\mathcal{D}$ is a continuous operator which implies $r_n z_n(t) = \mathcal{D}\hat{\eta}_n(t) \xrightarrow{d} \xi(t) = \mathcal{D}\eta(t)$. Now, $\mathcal{D}$ is a linear operator between two normed spaces. Thus by e.g. Theorem 2.2.4 in Atkinson and Han (2009), $\mathcal{D}$ is continuous if and only if it is bounded, i.e. there exists some $M > 0$ such that for all $F \in C_0^1(I, \mathcal{X})$, $\|\mathcal{D}F\| \leq M\|F\|$. We can write $u = \mathcal{D}F$ in its integral representation: $u(t) = F(t) + \int_0^t Dv_0(\gamma_0(s))u(s)\,ds$. Then we have

$$\begin{aligned}
\big\|u(t)\big\|_{\mathcal{X}} &\leq \big\|F(t)\big\|_{\mathcal{X}} + \int_0^t \big\|Dv_0(\gamma_0(s))\big\|_{op}\big\|u(s)\big\|_{\mathcal{X}}\,ds \\
&\leq \sup_{t\in I}\big\|F(t)\big\|_{\mathcal{X}} + \int_0^t \big\|Dv_0(\gamma_0(s))\big\|_{op}\big\|u(s)\big\|_{\mathcal{X}}\,ds \\
&\leq \left\{\sup_{t\in I}\big\|F(t)\big\|_{\mathcal{X}}\right\} \exp\left(\int_0^t \big\|Dv_0(\gamma_0(s))\big\|_{op}\,ds\right)
\end{aligned}$$

where the last line follows from an application of the Grönwall-Bellman inequality, which can be applied since the map $t \mapsto \big\|Dv_0(\gamma_0(t))\big\|_{op}$ is continuous on I by Assumption A2. □

Lemma 3. Under the conditions of Theorem 2, $\sup_{0\leq t\leq T}\big\|\delta_n(t)\big\|_{\mathcal{X}} = o_p(r_n^{-1})$.

Proof. We will show that

$$\sup_{0\leq t\leq T}\big\|R_n(t)\big\|_{\mathcal{X}} = o_p\left(\int_0^T \big\|\hat{\gamma}_n(t) - \gamma_0(t)\big\|_{\mathcal{X}}\,dt\right) \tag{B.4}$$

and

$$\big\|\delta_n(t)\big\|_{\mathcal{X}} \leq C \sup_{0\leq t\leq T}\big\|R_n(t)\big\|_{\mathcal{X}} \tag{B.5}$$

If (B.4) and (B.5) hold, we get the desired result:

$$\begin{aligned}
\sup_{0\leq t\leq T}\big\|\delta_n(t)\big\|_{\mathcal{X}} &= o_p\left(\int_0^T \big\|\hat{\gamma}_n(t) - \gamma_0(t)\big\|_{\mathcal{X}}\,dt\right) \\
&= o_P\left(\int_0^T \big\|z_n(t) + \delta_n(t)\big\|_{\mathcal{X}}\,dt\right) \\
&= o_P\left(\int_0^T \big\|z_n(t)\big\|_{\mathcal{X}}\,dt\right) \\
&= o_P(O_P(r_n^{-1})) = o_P(r_n^{-1})
\end{aligned}$$

where the fourth line follows from the third since Lemma 2 implies $\int_0^T \big\|z_n(t)\big\|_{\mathcal{X}}\,dt = O_P(r_n^{-1})$.

Now, let us establish (B.4). We decompose R_n as follows

$$\begin{aligned}
R_n(t) = &\int_0^t \left[(\hat{v}_n - v_0)(\hat{\gamma}_n(s)) - (\hat{v}_n - v_0)(\gamma_0(s))\right] ds \\
&+ \int_0^t \left[v_0(\hat{\gamma}_n(s)) - v_0(\gamma_0(s)) - Dv_0(\gamma_0(s))(\hat{\gamma}_n(s) - \gamma_0(s))\right] ds.
\end{aligned}$$

Let us take a look at each of these terms in turn. First we have

$$\begin{aligned}
&\big\|(\hat{v}_n - v_0)(\hat{\gamma}_n(s)) - (\hat{v}_n - v_0)(\gamma_0(s))\big\|_{\mathcal{X}} \\
&= \left\| \int_0^1 D(\hat{v}_n - v_0)(a\hat{\gamma}_n(s) + (1-a)\gamma_0(s))(\hat{\gamma}_n(s) - \gamma_0(s))\, da \right\|_{\mathcal{X}} \\
&\le \int_0^1 \big\|D(\hat{v}_n - v_0)(a\hat{\gamma}_n(s) + (1-a)\gamma_0(s))(\hat{\gamma}_n(s) - \gamma_0(s))\big\|_{\mathcal{X}}\, da \\
&\le \int_0^1 \big\|D(\hat{v}_n - v_0)(a\hat{\gamma}_n(s) + (1-a)\gamma_0(s))\big\|_{op} \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}}\, da \\
&\le \sup_{0\le a\le 1} \big\|D(\hat{v}_n - v_0)(a\hat{\gamma}_n(s) + (1-a)\gamma_0(s))\big\|_{op} \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}} \\
&\le \sup_{x\in\mathcal{U}} \big\|D(\hat{v}_n - v_0)(x)\big\|_{op} \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}}.
\end{aligned}$$

Next, we look at the second term of R_n. We have

$$\begin{aligned}
&\big\|v_0(\hat{\gamma}_n(s)) - v_0(\gamma_0(s)) - Dv_0(\gamma_0(s))(\hat{\gamma}_n(s) - \gamma_0(s))\big\|_{\mathcal{X}} \\
&= \left\| \int_0^1 [Dv_0(a\hat{\gamma}_n(s) + (1-a)\gamma_0(s)) - Dv_0(\gamma_0(s))](\hat{\gamma}_n(s) - \gamma_0(s))\, da \right\|_{\mathcal{X}} \\
&\le \sup_{0\le a\le 1} \big\|Dv_0(a\hat{\gamma}_n(s) + (1-a)\gamma_0(s)) - Dv_0(\gamma_0(s))\big\|_{op} \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}} \\
&\le r\big(\big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}}\big) \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}}
\end{aligned}$$

where $r(\delta) = \sup_{x\in\mathcal{U}} \sup_{y\in\mathcal{U}, \|y\|\le\delta} \big\|Dv_0(x+y) - Dv_0(x)\big\|_{op}$. Note that by Assumption A2, $r(\delta)$ as $\delta \to 0$. Thus for all $t \in [0,T]$, we have

$$\begin{aligned}
\big\|R_n(t)\big\|_{\mathcal{X}} \le{}& \sup_{x\in\mathcal{U}} \big\|D\hat{v}_n(x) - Dv_0(x)\big\|_{op} \int_0^t \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}}\, ds \\
&+ r\left(\sup_{0\le s\le T} \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}} \right) \int_0^t \big\|\hat{\gamma}_n(s) - \gamma_0(s)\big\|_{\mathcal{X}}\, ds.
\end{aligned}$$

We showed in Lemma 1 the uniform consistency of the empirical integral curve. This, in addition to Assumption A3, shows (B.4) holds.

Finally, we establish (B.5). From Assumption A2, we get that the map $t \mapsto \big\|Dv_0(\gamma_0(t))\big\|_{op}$ is continuous on I. Then it follows that

$$\begin{aligned}
\big\|\delta_n(t)\big\| &\le \sup_{0\le t\le T} \big\|R_n(t)\big\| + \int_0^t \big\|Dv_0(\gamma_0(s))\big\|_{op} \big\|\delta_n(s)\big\|\, ds \\
&\le \sup_{0\le t\le T} \big\|R_n(t)\big\| \exp\left\{ \int_0^t \big\|Dv_0(\gamma_0(s))\big\|_{op}\, ds \right\}
\end{aligned}$$

by the Grönwall-Bellman inequality, see Lemma 1 of Chandra (1970). Thus we have (B.5). □

Proof of Theorem 3. First, for a Donsker class $\mathcal{F} = \{f_x : x \in \mathcal{U}\}$, we have $\sqrt{n}(\hat{v}_n - v_0) \xrightarrow{d} \mathbb{B}$ and $\sqrt{n}c(\hat{v}_n^\circ - \hat{v}_n) \xrightarrow[W]{p} \mathbb{B}$. We will employ Corollary 13.8 in Kosorok (2008). Let Ψ_n and Ψ be as in the proof of Theorem 1. Define $\Psi_n^\circ(\gamma) = \dot{\gamma} - \hat{v}_n^\circ \circ \gamma$. Then $\hat{\gamma}_n^\circ$ is a zero of Ψ_n°. Since we have bootstrap consistency of the vector field, we also have $\sqrt{n}c(\Psi_n^\circ - \Psi_n) = \sqrt{n}c(\hat{v}_n - \hat{v}_n^\circ) \xrightarrow[W]{d} -\mathbb{G}$. Thus $\sqrt{n}c(\Psi_n^\circ - \Psi_n)$ and $\sqrt{n}(\Psi_n - \Psi)$ have the same limiting distribution, and the desired result on bootstrap consistency for the integral curve follows. □

Proof of Theorem 4. We can apply the arguments in the proof of Theorem 1 with the following specifications to obtain the desired result. Let the Banach space Θ be a closed subset of $C^1(D, \mathcal{U})$ that contains an open neighborhood of γ. In addition suppose Θ is such that $\sup_{\theta \in \Theta} \|\theta\|_{C^1} < \infty$ and $\theta(0, h) = x_0$ for all $\theta \in \Theta, h \in H$. Let $\mathbb{L}$ be the Banach space $C(D, \mathcal{X})$. The operators $\Psi(\gamma)(t, h) = \dot{\gamma}(t, h) - v(\gamma(t, h), h)$ and $\Psi_n(\gamma)(t, h) = \dot{\gamma}(t, h) - \hat{v}_n(\gamma(t, h), h)$ are both to be considered elements of $l^\infty(\Theta, \mathbb{L})$. □

Proof of Theorem 5. The proof can proceed in the same manner as in the proof of Theorem 3. Namely, the Donsker property of $\mathcal{F}$ implies $\sqrt{n}(\hat{v}_n - v_0) \xrightarrow{d} \mathbb{B}$ and $\sqrt{n}c(\hat{v}_n^\circ - \hat{v}_n) \xrightarrow[W]{p} \mathbb{B}$. Again, Corollary 13.8 in Kosorok (2008) can be applied with Ψ_n and Ψ as in the proof of Theorem 4 and $\Psi_n^\circ(\gamma)(t, h) = \dot{\gamma}(t, h) - \hat{v}_n^\circ(\gamma(t, h), h)$ to get the desired result. □

Proof of Theorem 6. Define the vector fields $v(x, h) = \nabla f(x, h)$ and $\hat{v}_n(x, h) = \nabla \hat{f}(x, h)$. Considering the two-parameter process $\sqrt{n}(\nabla \hat{f}(x, h) - \nabla f(x, h))$ as an empirical process with $(x, h) \in \mathcal{U} \times H$, its weak convergence can be established using Theorem 3.1 of Chaudhuri and Marron (2000) when $d = 1$. The result for $d > 1$ can be established in a similar way. Bootstrap consistency of the gradient vector fields is also given by Theorem 3.1 of Chaudhuri and Marron (2000), i.e. $\sqrt{n}(\nabla \hat{f}^\circ(x, h) - \nabla \hat{f}(x, h))$ converges weakly to the same limiting distribution as $\sqrt{n}(\nabla \hat{f}(x, h) - \nabla f(x, h))$. Finally, since f is twice continuously differentiable, v is locally Lipschitz and Fréchet differentiable at $(\gamma(t, h), h)$ for all $(t, h) \in D$. Thus the conditions of Theorem 5 are satisfied and we have the desired result. □

Proof of Theorem 7. For $d = 1$, the conditions of Theorem 3.2 of Chaudhuri and Marron (2000) are satisfied, yielding the weak convergence of $\sqrt{n}(\hat{v}_n(x, h) - v(x, h))$ and its bootstrapped version over $(x, h) \in \mathcal{U} \times H$ to a Gaussian process. Similar arguments can be applied for $d > 1$. Thus the conditions of Theorem 5 are satisfied and we have the desired result. □

Proof of Theorem 8. When $d = 1$, Theorem 3.2 in Chaudhuri and Marron (2000) can be used to establish the weak convergence of the two-parameter stochastic process $\hat{D}(x, h)$ as well as for the bootstrapped process $\hat{D}^\circ(x, h)$. A similar argument could be applied to extend the result to $d(d + 1)/2$ when $d > 1$. Then vectorizing symmetric positive-definite matrices in $\mathbb{R}^{d \times d}$ would give

us $\sqrt{n}(\hat{D}(x,h) - D(x,h))$ converges weakly to a matrix-valued Gaussian process on $\mathcal{U} \times H$. The continuous mapping theorem next gives the weak convergence of $\sqrt{n}(\hat{v}_n(f,h) - v(f,h))$ and $\sqrt{n}(\hat{v}_n^\circ(f,h) - \hat{v}_n(f,h))$ over (f,h) to the same $C(\mathbb{R}^d, \mathbb{R})$-valued tight process. This allows us to apply Theorem 5 to conclude bootstrap consistency for the heat flow. □

Acknowledgment

This research was partly supported by a European Research Council Starting Grant to Victor M. Panaretos. We acknowledge Tomas Rubin for helpful comments on an earlier draft. Finally, we are grateful to the Associate Editor and two referees for their insights.

References

Arias-Castro, E., Mason, D. and Pelletier, B. (2016). On the Estimation of the Gradient Lines of a Density and the Consistency of the Mean-Shift Algorithm. *Journal of Machine Learning Research* **17** 1–28.

Atkinson, K. and Han, W. (2009). Linear Operators on Normed Spaces. In *Theoreticl Numerical Analysis: A Functional Analysis Framework* 51–113.

Basser, P. J., Mattiello, J. and LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical journal* **66** 259–67.

Bellman, R. (2008). *Stability Theory of Differential Equations*. Dover Publications, New York.

Bierens, H. J. (1987). Kernel estimators of regression functions. In *Advances in econometrics: Fifth world congress* 99–144.

Blanes, S., Casas, F., Oteo, J. a. and Ros, J. (2009). The Magnus expansion and some of its applications. *Physics Reports* **470** 151–238.

Bottou, L. (2010). *Large-Scale Machine Learning with Stochastic Gradient Descent* In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics Paris France, August 22-27, 2010 Keynote, Invited and Contributed Papers* 177–186. Physica-Verlag HD, Heidelberg.

Brunel, N. J.-B., Clairon, Q. and D'Alché-Buc, F. (2014). Parametric Estimation of Ordinary Differential Equations With Orthogonality Conditions. *Journal of the American Statistical Association* **109** 173–185.

Chandra, J. (1970). On a Generalization of the Gronwall-Bellman in Partially Ordered Banach Spaces. *Journal of Mathematical Analysis and Applications* **681** 668–681.

Chaudhuri, P. and Marron, J. S. (2000). Scale Space View of Curve Estimation. *The Annals of Statistics* **28** 408–428.

Chen, Y.-C. C., Genovese, C. R. and Wasserman, L. (2015). Asymptotic theory for density ridges. *Annals of Statistics* **43** 1896–1928.

Chen, Y.-C., Genovese, C. R. and Wasserman, L. (2016). A Comprehensive Approach to Mode Clustering. *Electronic Journal of Statistics* **10** 210–241.

Chen, Y. C., Genovese, C. R. and Wasserman, L. (2017a). Density Level Sets: Asymptotics, Inference, and Visualization. *Journal of the American Statistical Association* 1–13.

Chen, Y. C., Genovese, C. R. and Wasserman, L. (2017b). Statistical inference using the morse-smale complex. *Electronic Journal of Statistics* **11** 1390–1433.

Chen, Y.-C., Genovese, C. R., Ho, S. and Wasserman, L. (2015). Optimal Ridge Detection using Coverage Risk. In *Advances in Neural Information Processing Systems 28* 316–324.

Ciollaro, M., Genovese, C. R. and Wang, D. (2016). Nonparametric Clustering of Functional Data Using Pseudo-Densities.

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

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7** 1–26.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Society for Industrial and Applied Mathematics.

Einbeck, J. and Dwyer, J. (2011). Using principal curves to analyse traffic patterns on freeways. *Transportmetrica* **7** 229–246.

Einbeck, J., Evers, L. and Bailer-Jones, C. (2008). Representing complex data using localized principal components with application to astronomical data. In *Lecture Notes in Computational Science and Engineering* **58** 178–201.

Einbeck, J., Tutz, G. and Evers, L. (2005). Local principal curves. *Statistics and Computing* **15** 301–313.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2009). On the path density of a gradient field. *Annals of Statistics* **37** 3236–3271.

Hageman, N. S., Toga, A. W., Narr, K. L. and Shattuck, D. W. (2009). A diffusion tensor imaging tractography algorithm based on Navier-Stokes fluid mechanics. *IEEE Transactions on Medical Imaging* **28** 348–60.

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

Hastie, T. and Stuetzle, W. (1989). Principal Curves. *Journal of the American Statistical Association* **84** 502–516.

Hill, B. J., Kendall, W. S. and Thönnes, E. (2012). Fibre-generated point processes and fields of orientations. *Annals of Applied Statistics* **6** 994–1020.

Huckemann, S., Hotz, T. and Munk, A. (2008). Global models for the orientation field of fingerprints: An approach based on quadratic differentials. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30** 1507–1519.

Koltchinskii, V., Sakhanenko, L. and Cai, S. (2007). Integral curves of noisy vector fields and statistical problems in diffusion tensor imaging: nonparametric kernel estimation and hypotheses testing. *Annals of Statistics* **35** 1576–1607.

Kosorok, M. R. (2008). Introduction to Empirical Processes and Semiparametric Inference. *Springer* **August** 1–491.

Krein, S. G. (1971). *Linear differential equations in Banach space*. American Mathematical Society, Providence, R.I.

Lee, K. K. (1992). *Lectures on Dynamical Systems, Structural Stability and Their Applications*. World Scientific, Singapore.

Lee, J. M. (2012). *Introduction to Smooth Manifolds*, 2 ed. Springer New York, New York.

Lisini, S. (2009). Nonlinear diffusion equations with variable coefficients as gradient flows in Wasserstein spaces. *ESAIM: Control, Optimisation and Calculus of Variations* **15** 712–740.

Mikusiński, J. (1978). *The Bochner integral*. Birkhauser Verlag.

Moler, C. and Van Loan, C. (2003). Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review* **45** 3–49.

Mukherjee, P., Berman, J. I., Chung, S. W., Hess, C. P. and Henry, R. G. (2008). Diffusion tensor MR imaging and fiber tractography: theoretic underpinnings. *American Journal of Neuroradiology* **29** 632–41.

Nelson, E. (1969). *Topics in Dynamics I: Flows*. Princeton University Press, Princeton.

Ozertem, U. and Erdogmus, D. (2011). Locally Defined Principal Curves and Surfaces. *Journal of Machine Learning Research* **12** 1249–1286.

Panaretos, V. M., Pham, T. and Yao, Z. (2014). Principal flows. *Journal of the American Statistical Association* **109** 424–436.

Praestgaard, J. and Wellner, J. A. (1993). Exchangeably Weighted Bootstraps of the General Empirical Process. *The Annals of Probability* **21** 2053–2086.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter Estimation for Differential Equations: A Generalized Smoothing Approach. *Differential Equations* **69** 741–796. MR1245301

Rinaldo, A. and Nugent, R. (2012). Stability of Density-Based Clustering. *Journal of Machine Learning Reasearch* **13** 905–948. MR2368570

Rinaldo, A. and Wasserman, L. (2010). Generalized density clustering. *Annals of Statistics* **38** 2678–2722. MR2930628

Theisel, H., Weinkauf, T., Hege, H. C. and Seidel, H. P. (2004). Stream line and path line oriented topology for 2D time-dependent vector fields. *IEEE Visualization 2004 - Proceedings, VIS 2004* 321–328. MR2722453

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes. Springer series in statistics*. Springer.

Vanhems, A. (2006). Nonparametric study of solutions of differential equations. *Econometric Theory* **22** 127–157. MR1385671

Walker, J. A. (1980). *Dynamical Systems and Evolution Equations*. Springer, Boston. MR2213495

Zachmanoglou, E. C. and Thoe, D. W. (1986). *Introduction to partial differential equations with applications*. Dover Publications, New York. MR0561511