Normal-bundle Bootstrap

Ruda Zhang† and Roger Ghanem‡

Abstract. Probabilistic models of data sets often exhibit salient geometric structure. Such a phenomenon is summed up in the manifold distribution hypothesis, and can be exploited in probabilistic learning. Here we present normal-bundle bootstrap (NBB), a method that generates new data which preserve the geometric structure of a given data set. Inspired by algorithms for manifold learning and concepts in differential geometry, our method decomposes the underlying probability measure into a marginalized measure on a learned data manifold and conditional measures on the normal spaces. The algorithm estimates the data manifold as a density ridge, and constructs new data by bootstrapping projection vectors and adding them to the ridge. We apply our method to the inference of density ridge and related statistics, and data augmentation to reduce overfitting.

Key words. probabilistic learning, data manifold, dynamical systems, resampling, data augmentation

AMS subject classifications. 37M22, 53-08, 53A07, 62F40, 62G09

1. Introduction. When data sets are modeled as multivariate probability distributions, such distributions often have salient geometric structure. In regression, the joint probability distribution of explanatory and response variables is centered around the response surface. In representation learning and deep learning, a common assumption is the manifold distribution hypothesis, that natural high-dimensional data concentrate close to a nonlinear low-dimensional manifold [2, 12]. In topological data analysis, including manifold learning, the goal is to capture such structures in data and exploit them in further analysis [23].

The goal of this paper is to present a method that generates new data, which preserve the geometric structure of a probability distribution modeling the given data set. As a variant of the bootstrap resampling method, it is useful for the inference of statistical estimators. Our method is also useful for data augmentation, where one wants to increase training data diversity to reduce overfitting, without collecting new data.

Our method is inspired by constructions in differential geometry and algorithms for nonlinear dimensionality reduction. Principal component analysis of a data set decomposes the Euclidean space of variables into orthogonal subspaces, in decreasing order of maximal data variance. If we consider the first few principal components to represent the geometry of the underlying distribution, and the remaining components to represent the normal space to the principal component space, we decompose the distribution into one on the principal component space and noises in the normal spaces at each point of the principal component space. Normal bundle of a manifold embedded in a Euclidean space generalizes such linear decomposition, such that every point in a neighborhood of the manifold can be uniquely represented as the
sum of its projection on the manifold and the projection vector. There are a few concepts that generalize principal components to nonlinear summaries of data. Principal curve [13] and, more generally, principal manifold is a smooth submanifold where each point is the expectation of the distribution in its normal space. More recently, [17] proposed a variant called density ridge, where each point is the mode of the distribution in a neighborhood in its normal space. Density ridge is locally defined and is estimated by subspace-constrained mean shift (SCMS), a gradient descent algorithm. Compared with principal curve algorithms, the SCMS algorithm is much faster, applicable to any manifold dimension, robust to outliers, and the ridge is fully learned from data.

Normal-bundle bootstrap (NBB) picks a point on the estimated density ridge and adds to it the projection vector of a random point, whose projection is in a neighborhood of the picked point on the ridge. With this procedure, the distribution on the ridge is preserved, while distributions in the normal spaces are locally randomized. Thus, the generated data will have greater diversity and remain consistent with the original distribution, including its geometric structure. Our method should work well for data sets in any Euclidean or Hilbert space, as long as the underlying distribution is concentrated around a low-dimensional submanifold, and the sample size is sufficient for the manifold dimension. Figure 1a-c illustrates density ridge, its normal bundle, and the normal-bundle bootstrap algorithm.

1.1. Related literature. Within bootstrap methods, normal-bundle bootstrap is mostly close to residual bootstrap in regression analysis, but our method is in the context of dimension reduction. Residual bootstrap fits a regression model on the data, and adds random residual in the response variables to each point on the fitted model, assuming the errors are identically distributed. Such residuals in our context are the projection vectors. Because the normal spaces on a manifold are not all parallel in general, we cannot bootstrap all the projection vectors. Instead, we only assume that the distributions in the normal spaces are continuously varying over the density ridge, and bootstrap nearby projection vectors. Also in regression analysis, wild bootstrap [24] allows for heteroscedastic errors, and bootstraps by flipping the sign of each residual at random, assuming error distributions are symmetric. Such assumption does not apply in our context, because each point on the density ridge is the mode of the distribution on a normal disk, which can be asymmetric and biased in general.

In probabilistic learning on manifolds, [20] proposed a Markov chain Monte Carlo (MCMC) sampler to generate new data sets, which preserve the concentration of probability measure estimated from the original data set [21] and have applications in uncertainty quantification [26]. This paper handles the same problem, but explicitly estimates the manifold by the density ridge, and generates new data by bootstrapping, which avoids the computational cost of MCMC sampling.

There is a large literature at the broad intersection of differential geometry and statistics. For parametric statistics on special manifolds with analytic expression, which includes directional statistics, see [10]. For nonparametric statistical theory on manifolds and its applications, especially for shape and object data, see [4] and [19]. Statistical problems on submanifolds defined by implicit functions are studied recently in [7].

Several MCMC methods have been proposed to sample from probability distributions on Riemannian submanifolds: [5] proposed a general constrained framework of Hamiltonian Monte
Figure 1. Density ridge, normal bundle, and NBB. (a) For a 2d Gaussian PDF (blue contours), its 1d density ridge (bold line) is its 1st principal component line, where the normal spaces (thin lines) are parallel to the 2nd principal component. Probability density on the normal spaces (right margin) declines faster than that on the ridge (top margin). (b) In general, density ridges are nonlinear, and its normal bundle decomposes the original distribution into one on the ridge and one on each normal space. (c) The NBB algorithm moves data points (solid blue) to the ridge (red) and for each point on the ridge, picks neighboring points on the ridge (shaded segment) and adds the projection vectors (dashed line) to construct new data points (hollow blue). Using a smooth frame can keep the constructed points in the normal space. (d) Commutative diagrams of normal bundle: $B$, basin of attraction; $R$, density ridge; $\pi$, projection; $U$, a neighborhood in density ridge.

Carlo (HMC) methods for manifolds defined by implicit constraints; [6] proposed a similar HMC method, but for manifolds with explicit forms of tangent spaces and geodesics.

The machine learning and deep learning communities also have various methods for estimating and sampling from probability densities with salient geometric structures. Manifold Parzen windows (MParzen) algorithm [22] is a kernel density estimation method which captures the data manifold structure. The estimated density function is easy to sample from, and we compare it with our method. Denoising auto-encoder [3] is a feed-forward neural network that implicitly estimates the data-generating distribution, and can sample from the learned model by running a Markov chain that adds noise and samples from the learned denoised distribution iteratively. Normalizing flow [18] is a deep neural network that represents a parametric family of probabilistic models, which is the outcome of a simple distribution mapped through a sequence of simple, invertible, differentiable transformations. It can be used for density estimation, sampling, simulation, and parameter estimation.

2. Mathematics: geometric decomposition of Euclidean spaces and probability measures. Consider a probability measure $\mu$ on the Euclidean space $\mathbb{R}^n$, which has a probability density function (PDF) $p$. Given a data set $X$ which is a random sample of size $N$ from $\mu$, we want to generate new data that are distinct from $X$, but consistent with $\mu$. In particular, we
want to solve this problem more efficiently by exploiting the geometric structure of $p$, which may be represented by a submanifold $R$ of $\mathbb{R}^n$. The mathematical foundation of our method is to decompose $\mathbb{R}^n$ into a collection $\{F_r\}_{r \in R}$ of submanifolds indexed by points in $R$, where each submanifold $F_r$ intersects $R$ at $r$ orthogonally. In this way, $\mu$ also gets decomposed into probability measures on submanifolds $R$ and each $F_r$.

**Definition 2.1.** Ridge of dimension $d \in \{0, \ldots, n\}$ for a twice differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$, denoted as $\text{Ridge}(f, d)$, is the set of points where the $c = n - d$ smallest eigenvalues of the Hessian are negative, and the span of their eigenspaces are orthogonal to the gradient: $\text{Ridge}(f, d) := \{ x \in \mathbb{R}^n : \lambda_c < 0, \text{Lag} = 0 \}$. Here, Hessian $H = \nabla \nabla f$ has an eigen-decomposition $H = V \Lambda V^T$, where $\Lambda = \text{diag}(\lambda)$ and $\lambda = (\lambda_i)_{i=1}^n$ is in increasing order. Let $V = (V_c ; V_d)$ where $V_c$ and $V_d$ are column matrices of $c$ and $d$ eigenvectors respectively. Denote projection matrices $U = V_d V_d^T$, $L = V_c V_c^T = I - U$, and gradient $g = \nabla f$.

**Assumption 2.2.** Let $D = \{ x \in \mathbb{R}^n : p(x) > 0 \}$, assume that: (1) $p|D \in C^2(D, \mathbb{R}_{>0})$; (2) for some $d \in \{1, \ldots, n-1\}$, $\text{Ridge}(p, d) \subset D$ is an embedded $d$-dimensional submanifold of $\mathbb{R}^n$.

**Density ridge** [17] is a ridge of a probability density function. With Assumption 2.2, (1) guarantees that $\text{Ridge}(p, d)$ is well-defined for every $d \in \{0, \ldots, n\}$; per the manifold distribution hypothesis we also require (2), and with the specific $d$ we define $R = \text{Ridge}(p, d)$ and codimension $c = n - d$. We note that this manifold assumption on density ridge is not very restrictive. In fact, it is analogous to a modal regression problem that assumes the conditional modes not to bifurcate. The remaining part of this section lays out the related mathematical concepts in differential geometry, measure theory, and dynamical system.

**2.1. Differential geometry.** We call the Euclidean space of dimension $n$ the Euclidean $n$-space; similarly, if a manifold has dimension $d$, we call it a $d$-manifold. An embedded submanifold $(M, T, A)$ of $\mathbb{R}^n$ is a subset $M \subset \mathbb{R}^n$ endowed with the subspace topology $T$ and the subspace smooth structure $A$, such that the inclusion map $\iota : M \mapsto \mathbb{R}^n$ is smooth and its differential has full rank. A Riemannian submanifold $(M, g)$ of $\mathbb{R}^n$ is an embedded submanifold $M$ endowed with the induced Riemannian metric $g = \iota^* \bar{g}$, where $\bar{g}$ is the Euclidean metric (the standard Riemannian metric on $\mathbb{R}^n$) and $\iota^*$ is the pullback operator by $\iota$. In the following, $M$ denotes a Riemannian $d$-submanifold of $\mathbb{R}^n$. At a point $p \in M$, tangent space $T_p M$ is the $d$-dimensional vector space consisting of all the vectors tangent to $M$ at $p$, and normal space $N_p M$ is the $c$-dimensional orthogonal complement to $T_p M$. The normal bundle $N M$ is the disjoint union of all the normal spaces: $N M = \sqcup_{p \in M} N_p M$. It is often identified with the product manifold $M \times \mathbb{R}^c$ so that its elements can be written as $(p, v)$, where $p \in M$, $v \in N_p M \cong \mathbb{R}^c$. The natural projection of $N M$ is the map $\pi_1 : N M \mapsto M$ such that $\pi_1(p, v) = p$.

We focus on neighborhoods of $M$ in $\mathbb{R}^n$ that are diffeomorphic images of open subsets of $N M$ under the addition map $E(p, v) = p + v$, so we can identify the two without ambiguity. For example, a tubular neighborhood $B$ is such a neighborhood that is diffeomorphic to a collection of normal disks of continuous radii: $B = E(\mathcal{D})$, where $\mathcal{D} = \{ (p, v) \in N M : |v| < \delta(p) \}$ and $\delta \in C^0(M, \mathbb{R}_+)$. The existence of tubular neighborhoods is guaranteed by the tubular neighborhood theorem [16, Thm 6.24]. Note that $E$ is bijective on $\mathcal{D}$, so its restriction $E|_{\mathcal{D}} : \mathcal{D} \mapsto B$ has an inverse: $\Phi = (E|_{\mathcal{D}})^{-1}$. A retraction from a topological space onto a subspace is a surjective continuous map that restricts to the identity map on the codomain.
A smooth submersion is a smooth map whose differentials are surjective everywhere. On a tubular neighborhood, we can define a retraction that is also a smooth submersion as such: \( r = \pi_1 \circ \Phi, \) \( r : B \to \mathcal{M}. \) It is identical to the projection onto \( \mathcal{M}, \) that is, \( r = P_M|_{B}, \) where \( P_M(x) = \arg\min_{p \in \mathcal{M}} \|p - x\|. \) Thus, we will call \( r \) the canonical projection of \( B, \) and denote it as \( \pi, \) which should not be confused with \( \pi_1. \)

Fiber bundle is a way to decompose a manifold into a manifold-indexed collection of homeomorphic manifolds of a lower dimension. Besides the normal bundle \( \mathcal{N}, \) we have now obtained another fiber bundle \( (B, \pi, \Phi) \) over \( \mathcal{M}, \) where \( B \) is the total space, \( \pi \) is the canonical projection, \( \Phi \) is the trivialization, and \( \mathcal{M} \) is the base space. The fiber \( \mathcal{F}_p \) over a point \( p \in \mathcal{M} \) is the preimage \( \mathcal{F}_p = \pi^{-1}(p), \mathcal{F}_p \subseteq B. \) In the case of tubular neighborhoods, the fibers are open disks. For simplicity, we will denote a fiber bundle by its total space, e.g. denote \( (B, \pi, \Phi) \) as \( B. \) Since \( \mathcal{D} = \Phi(B) \subset \mathcal{N}, \mathcal{M}, \) when there is no ambiguity, we will call \( B \) a normal bundle of \( \mathcal{M}. \) The normal bundle \( (B, \pi, \Phi) \) decomposes the neighborhood \( B \) into a collection \( \{\mathcal{F}_p\}_{p \in \mathcal{M}} \) of fibers indexed by the submanifold, so that every point in the neighborhood can be written uniquely as the sum of a point on the submanifold and a normal vector. In the special case of an \( \varepsilon \)-tubular neighborhood \( B_\varepsilon, \) this is a direct sum decomposition: \( B_\varepsilon = \mathcal{M} \oplus \mathcal{F}, \) where model fiber \( \mathcal{F} \) is an open disk of radius \( \varepsilon \) and dimension \( c. \)

2.2. Measure and density. Probability measures and probability density functions can also be extended to Riemannian manifolds. A measure \( \mu \) is a non-negative function on a sigma-algebra of an underlying set \( X, \) which is distributive with countable union of mutually disjoint sets. A natural choice of sigma-algebra for a topological space \( (X, \mathcal{T}) \) is its Borel sigma algebra, the sigma-algebra generated by its topology \( \mathcal{T}; \) this applies to all manifolds. A probability measure is just a normalized measure: \( \mu(X) = 1. \) We use superscript to indicate the underlying set of a measure if it is not \( \mathbb{R}^n. \) For example, \( \mu^\mathcal{M} \) denotes a probability measure on \( \mathcal{M}. \)

The Riemannian density \( dV_g \) on \( (\mathcal{M}, g) \) is a density uniquely determined by \( g. \) This density is not a probability density function, but a concept defined for smooth manifolds; the notation \( dV_g \) is intended to resemble a volume element. If \( \mathcal{M} \) is compact, its volume \( \text{Vol}(\mathcal{M}) \) is the integral of its Riemannian density: \( \text{Vol}(\mathcal{M}) = \int_\mathcal{M} dV_g; \) and its Hausdorff measure \( \mathcal{H}^d \) is the integral of its Riemannian density over measurable sets: \( \mathcal{H}^d(A) = \int_A dV_g, A \subset \mathcal{M}. \) We obtain a probability measure on \( \mathcal{M} \) by normalizing its Hausdorff measure: \( \mu_0^\mathcal{M} = \mathcal{H}^d/\text{Vol}(\mathcal{M}). \) Any function \( f \in C^0(\mathcal{M}, \mathbb{R}_{\geq 0}), \int_\mathcal{M} f dV_g = 1, \) is a probability density function with respect to \( \mu_0^\mathcal{M}, \) in the sense that it defines a probability measure \( \mu^\mathcal{M} = f\mu_0^\mathcal{M}. \) We denote such a probability density function as \( p^\mathcal{M}. \) Note that \( \mu_0^\mathcal{M} \) is used here as a reference probability measure, which can be considered as the uniform distribution on \( \mathcal{M}; \) in fact, it is the uniform distribution in the usual sense if \( \mathcal{M} \) has a positive Lebesgue measure.

On a normal bundle \( (B, \pi, \Phi) \) over \( \mathcal{M}, \) any probability measure \( \mu^B \) induces a probability measure \( \mu^\mathcal{M}_U \) on \( \mathcal{M} \) by marginalization: \( \mu^\mathcal{M}_U(U) = \mu^B(\pi^{-1}(U)), U \subset \mathcal{M}. \) Moreover, if \( \mu^B \) can be written as \( \mu^B = \rho\mu^\mathcal{M}_B, \) it induces a probability measure \( \mu^\mathcal{F} \) on each fiber \( \mathcal{F} \) by conditioning: \( \mu^\mathcal{F} = p^\mathcal{F}\mu^\mathcal{M}_{\pi^\mathcal{F}} \), where \( p^\mathcal{F} = \rho(\int_\mathcal{F} p^\mathcal{F} dV_g)^{-1}|_\mathcal{F}. \)

2.3. Dynamical system. A continuous-time dynamical system, or a flow, \( \phi : \mathbb{R} \times X \to X \) is a continuous action of the real group \( \mathbb{R} \) on a topological space \( X; \forall t, t' \in \mathbb{R}, \forall x \in X, \phi(0, x) = x \) and \( \phi(t', \phi(t, x)) = \phi(t + t', x). \) If the action is only on the semi-group \( \mathbb{R}_{\geq 0}, \) we call
it a semi-flow. The trajectory $\phi_x$ through a point $x \in X$ is the parameterized curve $\phi_x : \mathbb{R} \rightarrow X$, $\phi_x(t) = \phi(t, x)$. The time-$t$ map $\phi^t$, $t \in \mathbb{R}$, is the map $\phi^t : X \rightarrow X$, $\phi^t(x) = \phi(t, x)$. The time-$\infty$ map $\phi^{\infty}$ is the map $\phi^{\infty} : S \rightarrow S$, $\phi^{\infty}(x) = \lim_{t \rightarrow \infty} \phi^t(x)$, and $S \subset X$ is where the limit exists. A vector field $v(x)$ on a smooth manifold is a continuous map that takes each point to a tangent vector at that point. A flow generated by a vector field, if exists, is a differentiable flow such that $\forall x \in X$, $\frac{\partial \phi}{\partial t}(0, x) = v(x)$.

**Proposition 2.3** (flow). Let the subspace-constrained gradient field $v : D \rightarrow \mathbb{R}^n$, $v(x) = L(x)g(x)$. If $p(x)$ has bounded super-level sets $B_c = \{ x \in \mathbb{R}^n : p(x) \geq c \}$ for all $c > 0$, then $v(x)$ generates a semi-flow $\phi : \mathbb{R}_{\geq 0} \times D \rightarrow D$. If $p(x)$ has a compact support $D$, let $v(x) = 0$, $\forall x \in \partial D$, then $v(x)$ generates a flow $\phi : \mathbb{R} \times \overline{D} \rightarrow \overline{D}$. Moreover, if $v(x)$ is locally Lipschitz or $C^k$, $k \geq 1$, then $\phi$ is locally Lipschitz or $C^k$, respectively.

**Proof.** Because $p \in C^2(\mathbb{R}^n, \mathbb{R}_{\geq 0})$, we have $H = \nabla \nabla p \in C^0(\mathbb{R}^n, S(n))$, where $S(n) = \{ A \in \mathbb{R}^{n \times n} : A = A^T \}$. So the subspace $\text{Span}(V_c)$ spanned by the eigenvectors of the bottom-$c$ eigenvalues of $H$ is continuously varying: $\text{Span}(V_c) \in C^0(\mathbb{R}^{n \times n}, G_{c,n})$, where the Grassmann manifold $G_{c,n}$ consists of $c$-subspaces of $\mathbb{R}^n$. This means the projection matrix $L = V_cV_c^T$ is also continuously varying. Since $g = \nabla p \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, we have $v(x) = L(x)g(x)$ is continuous, and therefore it is a vector field on $\mathbb{R}^n$. Let $\partial B_c$ be the boundary of $B_c$. For each $x \in \partial B_c$, if $g(x) \neq 0$, then by the regular level set theorem [14, Thm 3.2], there is a neighborhood $U(x) \subset \mathbb{R}^n$ such that $\partial B_c \cap U(x)$ is a $C^2$ hypersurface in $\mathbb{R}^n$. Additionally, $g(x) \in N_x\partial B_c$ points in the inward normal direction. Therefore, the projection of $g(x)$ onto any subspace would still points inwards or vanish, which applies to $v(x) = L(x)g(x)$. If $g(x) = 0$, apparently $v(x) = 0$. So for all $x \in \partial B_c$, $v(x)$ points into $B_c$ or vanish. Because $B_c$ is a closed set and assumed to be bounded, it is compact. Thus, the vector field $v(x)$ is forward complete, that is, it generates a unique semi-flow $\phi : \mathbb{R}_{\geq 0} \times B_c \rightarrow B_c$. As $c \rightarrow 0$, $B_c$ expands to $\mathbb{R}^n$, so $v(x)$ generates a semi-flow on $\mathbb{R}^n$. If $p(x)$ is compactly supported, then so is $v(x)$, therefore $v(x)$ is complete and generates a unique global flow $\phi : \mathbb{R} \times \overline{D} \rightarrow \overline{D}$.

**Proposition 2.4** (convergence). If $p(x)$ is analytic and has bounded super-level sets, then every forward trajectory converges to a fixed point: $\forall x \in \mathbb{R}^n$, $\exists x^* \in v^{-1}(0)$, $\lim_{t \rightarrow +\infty} \phi_x(t) = x^*$. 

**Proof.** When $v(x) \neq 0$, we have $v = Lg = V_cV_c^Tg \neq 0$, which means $V_c^Tg \neq 0$ and therefore $v(g) = gV_cV_c^Tg \geq 0$. Because $v(g) = gV_cV_c^Tg \geq 0$, we have $(v, g) = 0$ implies $v = 0$. Let $\theta(u, w) = (u, w)/(\|u\|\|w\|)$, then $\{ x : \theta(g, v) = \pi/2 \} \subset v^{-1}(0)$. Let $\delta \in [0, \pi/2]$ and $U(\delta) = \{ x : \theta(g, v) \geq \pi/2 - \delta \} \subset v^{-1}(0)$. Let $\tilde{v}(x) = 0$ if $x \in U(\delta)$ and $\tilde{v}(x) = v(x)$ otherwise. Let $\phi \in \mathcal{F}$ be the semi-flow generated by $\tilde{v}$. Then $\forall \delta > 0$, $\forall x \in D \setminus U(\delta)$, $\forall t \in \mathbb{R}_{\geq 0}$: $\theta(g(\xi), \tilde{v}(\xi)) \leq \pi/2 - \delta$, where $\xi = \phi_x(t)$. By Lojasiewicz theorem with an angle condition (see [15, 1]), either $\lim_{t \rightarrow +\infty} \|\phi_x(t)\| = \infty$ or $\exists x^* \in \mathbb{R}^n$, $\lim_{t \rightarrow +\infty} \phi_x(t) = x^*$. Because $p(\phi_x(t))$ is non-decreasing and $p(x)$ has compact super-level sets, $\phi_x(t)$ must converge to a point $x^*$. And because $\tilde{v}(x^*) = 0$, we have $x^* \in U(\delta)$. Let $x^* = \lim_{t \rightarrow +\infty} x^*$, then $x^* \in U(0) = v^{-1}(0)$.

Due to the convergence property of $\phi$, we can focus on its fixed points. For $\phi$, the set of asymptotically stable fixed points is $\mathcal{R}$, which can be easily checked by definition. In fact, $\mathcal{R}$ is the attractor of $\phi$, and nearby trajectories approach along normal directions [11, Lemma 8]. The basin of attraction $B$ of $\mathcal{R}$ is the union of images of all trajectories that tend towards it: $B = \{ x \in \mathbb{R}^n : \phi^{\infty}(x) \in \mathcal{R} \}$. By [11, Lemma 3], $B$ contains an $\varepsilon$-tubular neighborhood $B_\varepsilon$ of $\mathcal{R}$,
where $\mathcal R$ is exponentially attractive. Here we show that, under a stronger manifold assumption, $B$ is a set of probability one.

Proposition 2.5 (basin). If $p(x)$ is analytic and has a compact support $\overline D$, and $A_u = \{ x \in D : v = 0, \lambda_c > 0 \}$ and $A_c = \{ x \in D : v = 0, \lambda_c = 0 \}$ are, respectively, embedded d- and $(d-1)$-submanifolds of $\mathbb R^n$, then $B$ is a subset of full Lebesgue measure and therefore has probability one: $\lambda(D \setminus B) = 0$, $\mu(B) = 1$, where $\lambda$ is the Lebesgue measure on $\mathbb R^n$.

Proof. $v^{-1}(0) = \mathcal R \cup A_u \cup A_c$, and $\forall x \in v^{-1}(0)$, $W(x) = \{ \xi \in D : \lim_{\tau \to +\infty} \phi_\tau(t) = x \}$ has dimension at most $c$. For $x \in A_u$, $x$ has an unstable manifold of dimension at least one, so $D_u = \{ x \in D : \lim_{\tau \to +\infty} \phi_\tau(t) \in A_u \}$ has Lebesgue measure zero. Because $A_c$ has dimension $d-1$, and $d-1 + c = n-1 < n$, so $D_c = \{ x \in D : \lim_{\tau \to +\infty} \phi_\tau(t) \in A_c \}$ also has Lebesgue measure zero. Because $D = B \cup D_u \cup D_c$, so $\lambda(D \setminus B) = \lambda(D_u \cup D_c) = \lambda(D_u) + \lambda(D_c) = 0$. Because $\mu(D) = 1$ and $\mu(D \setminus B) \leq \max(p)\lambda(D \setminus B) = 0$, so $\mu(B) \geq \mu(D) - \mu(D \setminus B) = 1$ and therefore $\mu(B) = 1$.

Now we have (yet another) fiber bundle $(B, \pi, \Phi)$ over $\mathcal R$, where canonical projection $\pi(x) = \phi^\infty(x)$ and trivialization $\Phi(x) = (\pi(x), x - \pi(x))$. $\pi$ is a retraction that approximates the projection $P^\mathcal R(x) = \arg\min_{p \in \mathcal R} \| p - x \|$ to the second order [25, Lem 2.8]. If $\phi$ is smooth within $B$, then $\pi$ is a smooth submersion. Because the fiber $\mathcal F_r = \pi^{-1}(r)$ over each point $r \in \mathcal R$ is a level set of $\pi$, by the submersion level set theorem [16, Cor 5.13], it is a properly embedded $c$-submanifold. In terms of the dynamical system, each $\mathcal F_r$ is a stable manifold, because every forward trajectory starting on $\mathcal F_r$ stays within $\mathcal F_r$ and converges to $r$. Because $\mathcal F_r$ intersects $\mathcal R$ at $r$ orthogonally, we will also call $(B, \pi, \Phi)$ a normal bundle over $\mathcal R$, when there is no ambiguity. See Figure 1d for commutative diagrams of this bundle and its restriction to a subset of the ridge. As with the general case discussed earlier, any probability measure $\mu^B$ on $B$ induces a probability measure $\mu^\mathcal R$ on $\mathcal R$ by marginalization: $\mu^\mathcal R = \mu^B \circ \pi^{-1}$. But the dynamical system offers a more explicit perspective on this marginalization process: $\phi$ continuously transforms $\mu^B$ towards $\mathcal R$ such that at $t > 0$, $\mu^B_t = \mu^B \circ (\phi^t)^{-1}$, and the induced measure is the asymptotic measure, $\mu^\mathcal R = \lim_{t \to \infty} \mu^B_t$.

3. Algorithm. We formally describe normal-bundle bootstrap in Algorithm 3.1, and analyze its properties. Let $K_h$ be a density kernel with bandwidth $h$, density estimate $\hat p_h(x) = N^{-1} \sum_{i=1}^N K_h(x-x_i)$, and leave-one-out density estimate $\hat p_{h,-i}(x) = (N-1)^{-1} \sum_{j \neq i} K_h(x-x_j)$. Let $\alpha \in (1, +\infty)$ be an oversmoothing factor and $k \in \{0, \ldots, N\}$ be the number of nearest neighbors. A standing assumption of the algorithm is that $d$ is small, so that $N$ does not have to be too large for good estimation. On the other hand, $n$ can be reasonably large under typical computational constraints.

In this algorithm, smooth frame construction (line 3) and coordinate representation (line 4) can be removed to save computation, but with less desirable results. In this case data construction (line 6) directly uses projection vectors $x_i - \hat r_i$ or normal vectors $\hat n_i = L_i(x_i - \hat r_i)$, where $l = K(i, j)$. Algorithms SCMS and SmoothFrame are given in supplementary materials.

3.1. Qualitative properties of the dynamical system. In subsection 2.3 we have shown that, under suitable conditions, subspace-constrained gradient field $v$ generates a flow $\phi$ whose attractor is $\mathcal R$, and the basin of attraction $B$ is a fiber bundle with canonical projection $\pi = \phi^\infty$. The dynamical system $\phi$ is determined by $p$. Because the dynamical system is stable [11, Thm
Algorithm 3.1 NormalBundleBootstrap($X, d, \alpha, k$)

1: $h \leftarrow \alpha \arg \max_h \sum_{i=1}^{N} \log \hat{p}_{h,-i}(x_i)$ ▶ kernel bandwidth selection
2: $(\hat{r}_i, V_c,i) \leftarrow \text{SCMS}(x_i; \log \hat{p}_h, d)$, for $i \in N$ ▶ ridge estimation
3: $E \leftarrow \text{SmoothFrame}(\hat{r}, V_c, n - d)$ ▶ align bottom-$c$ eigenvectors
4: $[\hat{n}]_i \leftarrow E_i^T(x_i - \hat{r}_i)$, for $i \in N$ ▶ coordinates of normal vectors
5: $K \leftarrow \text{KNN}(\hat{r}, k)$ ▶ $k$-nearest neighbors on ridge
6: $\hat{x}_{ij} \leftarrow \hat{r}_i + E_i[\hat{n}] K(i,j)$, for $i \in N, j \in k$ ▶ construct new data

---

Figure 2. Subspace-constrained gradient flow as projection to estimated density ridge. Data (blue points); true (gray curve) and estimated (red curve) density ridge; trajectories (orange curves), pointing towards estimated ridge. (a) True ridge is the unit circle, a manifold without boundary; the estimated ridge is also without boundary. (b) True ridge is a parabola segment, a manifold with boundary; the estimated ridge is unbounded.

$p$ can be replaced by an estimate $\hat{p}$ to obtain an attractor $\hat{R} = \text{Ridge}(\hat{p}, d)$ that approximates $\mathcal{R}$. Here we use a density estimate $\hat{p}_h$ with Gaussian kernel $K_h(x) \propto \exp(-x^2/(2h^2))$. Denote the generated flow as $\phi_N$ and the estimated ridge as $\hat{R}_N = \text{Ridge}(\hat{p}_h, d)$.

It is preferable to define $\phi_N$ by $\log \hat{p}_h$ instead of $\hat{p}_h$. Note that $\text{Ridge}(\log \hat{p}_h, d) = \hat{R}_N$. If $\phi_N$ is defined by $\log \hat{p}_h$, then $\mathcal{B}$ is larger and independent of the size of normal space distribution [11, Thm 7], and trajectories are more orthogonal to $\hat{R}_N$ (see Figure 2). Moreover, $\hat{R}_N$ is exponentially stable within $\mathcal{B}$, as $v$ is approximately linear in normal spaces [11, Lemma 8].

The attractor $\hat{R}_N$ may be bounded or unbounded. If $\mathcal{R}$ is a compact submanifold without boundary, as is often assumed in previous studies, $\hat{R}_N$ can be compact and without boundary. If $\mathcal{R}$ has a boundary, $\hat{R}_N$ would be unbounded, see Figure 2. This is also true if $\mathcal{R}$ is noncompact, as is the Gaussian example in Figure 1a. In such cases, although finite data is always bounded, the attractor will be unbounded.
3.2. Statistical properties. As in subsections 2.2 and 2.3, the normal bundle \((B, \pi, \Phi)\) over the density ridge \(\mathcal{R}\) decomposes the original probability measure \(\mu\) into a “marginalized measure” \(\mu_1^N\) on the ridge and a “conditional measure” \(\mu_{F^r}\) on each fiber, where \(r \in \mathcal{R}\). If we know \(\mu_1^N\) and each \(\mu_{F^r}\), we can sample \(\mu\) as follows: first sample \(r \sim \mu_1^N\), and then sample \(\mu_{F^r}\).

Although such measures are unknown, we can still estimate them from available data, and use them for inference and data augmentation. Here we show that normal-bundle bootstrap constructs new data points that are consistent with the conditional measures on the normal spaces, and have nice finite-sample validity.

**Assumption 3.1** ([11], Sec 2.2). In a neighborhood \(B\) of ridge \(\mathcal{R}\), (A0) \(p(x)\) is three times differentiable; (A1) \(p(x)\) is sharply curved in normal spaces: \(\lambda_c < -\beta\) and \(\lambda_c < \lambda_{c+1} - \beta\), where \(\beta > 0\); (A2) trajectories \(\phi_x(t)\) are not too wiggly and tangential gradients \(U(x)g(x)\) are not too large: \(\|U(x)g(x)\| \max_{i,j,k} \left| \frac{\partial H_{ij}}{\partial x_k}(x) \right| < \frac{\beta^2}{2n^{3/2}}\).

**Theorem 3.2** (consistency). Let Assumption 3.1 hold for the measure \(\mu\) in the basin of attraction \(B\), and the conditional measure \(\mu_{F^r}\) varies slowly over the ridge \(\mathcal{R}\), then for each estimated ridge point \(\hat{r} = \pi_N(x) = \phi_N^\infty(x)\), as sample size \(N \to \infty\), the distributions of the constructed data points \(\tilde{x}_j, j \leq k\), converge to the distribution restricted to the fiber of the estimated ridge point: \(\tilde{x}_j|\hat{r} \sim d x|_{F^r}\).

**Proof.** The normal bundle \((B, \pi_N)\) over the estimated ridge \(\hat{\mathcal{R}}_N\) decomposes the original measure \(\mu\) into the marginalized measure \(\mu_1^N\) and the conditional measures \(\mu_{F^r}\), \(\hat{r} \in \hat{\mathcal{R}}_N\). Because the data is distributed as the original measure, \(x \sim \mu\), each estimated ridge point is then distributed as the marginal measure, and the normal vector at each estimated ridge point is distributed as the conditional measure at that ridge point: \(\hat{r} \sim \mu_1^N\) and \(n|\hat{r} \sim \mu_{F^r}\).

Since the Gaussian kernel is smooth, the density estimate \(\hat{p}_h(x)\) satisfies condition (A0). By [11, Thm 5], as sample size \(N\) goes to infinity, the estimated ridge \(\hat{\mathcal{R}}_N\) within the basin of attraction \(B\) converges to the true ridge: \(\lim_{N \to \infty} \text{Haus}(\mathcal{R}, \hat{\mathcal{R}}_N) = 0\), where the Hausdorff distance between two sets is defined as \(\text{Haus}(A, B) = \max\{\sup_{x \in A} d(x, B), \sup_{x \in B} d(x, A)\}\). Because the conditional measures \(\mu_{F^r}\) over the true ridge \(\mathcal{R}\) vary slowly, and the estimated ridge approximates the true ridge, the conditional measures \(\mu_{F^r}\) over the estimated ridge \(\hat{\mathcal{R}}_N\) also vary slowly. For an estimated ridge point \(\hat{r}\), the normal vectors at its \(k\)-nearest neighbors \(\hat{r}_j, j \leq k\), are thus distributed similarly to the normal vector at this point: \(n_j|\hat{r}_j \sim \mu_{F^r}\), \(\mu_{F^r} \approx \mu_{F^r}\). As sample size \(N\) goes to infinity, the distances to its \(k\)-nearest neighbors vanishes: \(\lim_{N \to \infty} d(\hat{r}, \hat{r}_j) = 0\). Therefore, the distributions of neighboring normal vectors converge to the distribution of the normal vector at the estimated ridge point: \(\lim_{N \to \infty} \mu_{F^r} = \mu_{F^r}\). Note that this limit is understood in the sense of a metric on measure spaces, such as the Wasserstein metrics. The constructed data points add neighboring normal vectors to the estimated ridge point, \(\tilde{x}_j|\hat{r} = \hat{r} + n_j\); as a result, their distributions converge to the original measure restricted to the fiber of the estimated ridge point: \(\tilde{x}_j|\hat{r} \sim d x|_{F^r}\).

We have shown that the normal-bundle bootstrapped data have desirable large-sample asymptotic behavior, but their finite-sample behavior is also very good. In fact, as soon as the estimated ridge becomes close enough to the true ridge such that the conditional measures \(\mu_{F^r}\) over the estimated ridge vary slowly, the conditional measures on neighboring fibers become...
similar to each other: $\mu^{\mathcal{R}} \approx \mu^{\mathcal{F}}$. This would suffice to make the constructed data distribute similarly to the original measure restricted to a fiber: $\hat{x}_j | \hat{\mathcal{F}} \sim x | \mathcal{F}$. Even if the estimated ridge has a finite bias to the true ridge, see e.g. Figure 2a, it would not affect the conclusion. Suppose the true ridge is the unit circle and the conditional measures $\mu^{\mathcal{R}}$ over the true ridge are identical, if the estimated ridge is a circle of a smaller radius, then the conditional measures $\mu^{\mathcal{R}}$ over the estimated ridge are also identical, but with a constant bias to $\mu^{\mathcal{R}}$. Despite such a bias, the constructed data will have the same distribution as the restricted measure: $\hat{x}_j | \hat{\mathcal{F}} \sim x | \mathcal{F}$.

We will illustrate the finite-sample advantage of NBB in section 4.

### 3.3. Computational properties.

SCMS [17] is an iterative algorithm that updates point locations by $x_{t+1} = x_t + s(x_t)$, where $s(x) = L(x) m(x)$ is the subspace-constrained mean-shift vector and $m(x)$ is the mean-shift vector. If density estimate $\hat{p}_h$ uses a Gaussian kernel with bandwidth $h$, then $m(x) = h^2 \hat{g}_h(x) / \hat{p}_h(x)$, where $\hat{g}_h(x) = \nabla \hat{p}_h(x)$ is the plug-in estimate of density gradient. A naive implementation of SCMS would have a computational complexity of $O(N^2 n^3)$ per iteration, where the $O(N^2)$ part comes from computing for each update point $x_t$ using all data points, and the $O(n^3)$ part comes from eigen-decomposition of the Hessian. Although estimates of density, gradient, and Hessian all need to be computed for each update point, the most costly operation is the eigen-decomposition.

However, a better implementation can reduce the computational complexity to $O(kdn^2)$ per iteration for one update point. Here we use the $k$-nearest data points, assuming that the more distant points have negligible contribution to the estimated terms. And we use partial eigen-decomposition to obtain the top $d$ eigen-pairs in $O(dn^2)$ time.

Another direction to accelerate computation is by reducing the number of iterations. Recall that the attractor $\hat{\mathcal{R}}_N$ is exponentially stable, therefore $\{x_t\}_{t \in \mathbb{N}}$ is linearly convergent. We can use Newton’s method for root finding to achieve quadratic convergence. For $x$ in a neighborhood $B$ of $\hat{\mathcal{R}}_N$, let subspace $S = \text{Span}(V_c)$, affine space $A = x + S$, and let $C$ be the component of $A \cap B$ containing $x$. Then ridge point $r = C \cap \hat{\mathcal{R}}_N$ is the unique zero of $v|_C$ and it is regular. Recall that $v = L g$, $L = V_c V_c^T$, Newton’s method for $v|_C = 0$ updates by $x_{t+1} = x_t + L_t \delta_t$, where $\delta_t$ solves $L_0 H_1 L_0 \delta_t = -L_0 g_t$ or $L_t H_1 L_t \delta_t = -L_t g_t$. Both converge quadratically near $\hat{\mathcal{R}}_N$, while the former only requires (partial) eigen-decomposition at the first step, and the latter has a larger convergence region [25, Lem 2.12].

### 4. Experiments.

In this section we showcase the application of normal-bundle bootstrap in inference and data augmentation, using two simple examples.

#### 4.1. Inference: confidence set of density ridge.

Normal-bundle bootstrap constructs new data points that approximate the distributions on normal spaces of the estimated density ridge, and thus can be used for inference of population parameters of these distributions. For example, it can provide confidence sets of the true density ridge via repeated mode estimation in each normal space, and provide confidence sets of principal manifolds [13] via repeated mean estimation in each normal space.

For a confidence set $\hat{C}_N$ of $\mathcal{R}$, it is asymptotically valid as a uniform confidence set at level $1 - \alpha$ if $\liminf_{N \to \infty} P(\mathcal{R} \subset \hat{C}_N) \geq 1 - \alpha$; similarly, it is valid as a pointwise confidence set if $\liminf_{N \to \infty} \mathbb{E}_0^{\mathcal{R}}(\mathcal{R} \subset \hat{C}_N) \geq 1 - \alpha$. Pointwise confidence sets are less conservative and can be more useful. We define an NBB pointwise confidence set $\hat{C}^{\text{NBB}}_N = \hat{\mathcal{R}}_N \oplus D_\alpha =$
\{ \hat{\tau} + \hat{n} : \hat{\tau} \in \hat{\mathcal{R}}_N, \hat{n} \in D_\alpha(\hat{\tau}) \}, \text{ where disk } D_\alpha(\hat{\tau}) = \hat{m} \oplus \varepsilon_\alpha = \{ \hat{n} \in N_{\hat{\tau}} \hat{\mathcal{R}}_N : d(\hat{n}, \hat{m}) < \varepsilon_\alpha \}.

For \( \hat{\tau}, \hat{m}_i \) is the mode estimated from the constructed points \( \hat{x}_i \). Radius \( \varepsilon_\alpha \) is determined by \( P(d(m, \hat{m}) < \varepsilon_\alpha) = 1 - \alpha \), where \( m \) is the mode of \( p|_{\mathcal{F}_\tau} \) and corresponds to \( \mathcal{R} \cap \mathcal{F}_\tau \); its estimator \( \hat{\varepsilon}_\alpha \) is the \( \alpha \)-upper quantile of \( \{ d(\hat{n}_i, \hat{m}) \}_{i=1}^B \), where \( \hat{m}^* \) denotes a bootstrap estimate using a bootstrap resample of the constructed points. Note that an NBB pointwise confidence set for a principal manifold can be defined simply by replacing \( m \) and \( \hat{m} \) with mean and sample mean.

Alternatively, confidence sets for \( \mathcal{R} \) can also be obtained by bootstrap. \cite{9} showed that a bootstrap uniform confidence set \( \hat{C}_N^{\mathcal{B}} \) converges in Hausdorff distance at a rate of \( O(N^{-1/2}) \) to the smoothed density ridge \( \mathcal{R}_h = \text{Ridge}(p_h, d) \), where smoothed density \( p_h = p \ast K_h \) and \( \ast \) denotes convolution. Here, \( \hat{C}_N^{\mathcal{B}} = \mathcal{R}_h \oplus \varepsilon_\alpha = \{ x \in \mathbb{R}^d : d(x, \mathcal{R}_h) < \varepsilon_\alpha \} \) is the \( \varepsilon_\alpha \)-uniform tubular neighborhood of \( \mathcal{R}_h \), the estimated ridge using kernel bandwidth \( h \). Radius \( \varepsilon_\alpha \) is determined by \( P(d_H(\hat{\mathcal{R}}_h, \mathcal{R}_h) < \varepsilon_\alpha) = 1 - \alpha \), where \( d_H(\hat{\mathcal{R}}_h, \mathcal{R}_h) = \sup_{x \in \hat{\mathcal{R}}_h} d(x, \mathcal{R}_h) \); its estimator \( \hat{\varepsilon}_\alpha \) is the \( \alpha \)-upper quantile of \( \{ d_H(\hat{\mathcal{R}}_h^*, \hat{\mathcal{R}}_h) \}_{i=1}^B \). A bootstrap pointwise confidence set of \( \mathcal{R}_h \) can be similarly defined where \( \varepsilon_\alpha \) is determined by \( P(d(\mathbf{r}, \hat{\mathcal{R}}_h) < \varepsilon_\alpha) = 1 - \alpha \) and estimator \( \hat{\varepsilon}_\alpha \) is the \( \alpha \)-upper quantile of \( \{ d(\hat{\mathbf{r}}_{i,h}, \hat{\mathcal{R}}_h) \}_{i=1}^B \). But if \( N \) is small and therefore \( h \) is large, \( \mathcal{R}_h \) can have large bias from \( \mathcal{R} \), so the bootstrap confidence sets can have poor coverage of \( \mathcal{R} \).

Here we compare the pointwise confidence sets of density ridge by NBB and bootstrap. As an experiment, data are sampled uniformly on the unit circle, and a Gaussian noise is added in the radial direction: \( \mathbf{x} = r \mathbf{e}^{i\theta}, \ \theta \sim U[0, 2\pi], \ r \sim N(1, 0.2^2) \). The 1d density ridge of \( \mathbf{x} \) is numerically identical with the unit circle. Figure 3(a-b) illustrates \( \hat{C}_N^{\mathcal{B}} \) and \( \hat{C}^{\mathcal{B}} \) on a random sample, and Figure 3(c-d) compares their finite-sample validity and average compute time over independent samples. \( \hat{C}_N^{\mathcal{B}} \) is valid throughout the range of sample sizes computed, while the validity of \( \hat{C}^{\mathcal{B}} \) slowly improves. Moreover, \( \hat{C}^{\mathcal{B}} \) is computationally costlier than \( \hat{C}_N^{\mathcal{B}} \), due to repeated ridge estimation. Although repeated mode estimation is also costly, it is faster than ridge estimation of the same problem size, and the constructed points in each normal space is only a fraction of the original sample. Specifically, the computational complexity of \( \hat{C}^{\mathcal{B}} \) is \( O(n^3NB) \) per iteration, from bootstrap repetitions of ridge estimation; that of \( \hat{C}_N^{\mathcal{B}} \) is \( O(nkB) \), where \( O(nK) \) comes from estimating gradient using \( k \) constructed data points, and \( O(NB) \) comes from computing for all normal spaces and all bootstrap repetitions. Note that other population parameters like mean and quantiles can be estimated much faster than the mode, so the related inference using NBB will be much faster than in this example, such as confidence sets of principal manifolds.

4.2. Data augmentation: regression by deep neural network. For machine learning tasks, the data constructed by normal-bundle bootstrap can be used to augment the original data to avoid overfitting. The idea behind this is that when the amount of training data is insufficient for a model not to overfit, but enough for a good estimate of the density ridge, we can include the NBB constructed data to increase the amount of training data. Because for each estimated ridge point, the NBB constructed data is balanced around the true ridge in the sense that their estimated mode is near the true ridge point, so the augmented training data can resist overfitting to the noises.

Here we consider a regression problem with one input parameter and a functional output. Let \( S^1 \subset \mathbb{R}^2 \) be the unit circle, \( \theta \in \mathbb{R} \) be a rotation angle (with unit \( \pi \)), \( \tau : S^1 \rightarrow S^1 \) be the map
between initial and final configurations of the circle, and $f$ be the relationship between $\theta$ and $\tau$ such that $f(\theta) = \tau$. The task is to learn $f$ from data. We discretize the circle into a set of $l$ random points with initial angles $\{\gamma_j \pi\}_{j=1}^l \subset [0, 2\pi)$. Under the true model, when $\theta = \theta_i$ their coordinates can be written as $(x_{ij}, y_{ij}) = (\cos(\pi(\theta_i + \gamma_j)), \sin(\pi(\theta_i + \gamma_j)))$. Assume that all variables are subject to measurement error such that we can only observe $\tilde{\theta} \sim N(\theta, 0.2^2)$ and $\tilde{x}_j, \tilde{y}_j \sim N(\theta, 0.2^2), j \in \{1, \ldots, l\}$. We obtain training data $(\tilde{\theta}_i, (\tilde{x}_{ij}, \tilde{y}_{ij})_{j=1}^l)_{i=1}^N$, and obtain another set of data for validation. Specifically, we have $l = 8$ and $N = 32$, so the training data is a $32 \times 17$ matrix.

For the neural network, we use a sequential model with four densely connected hidden layers, which have 256, 128, 64, and 32 units respectively and use the ReLU activation function;
the output layer has 16 units. We train the network to minimize mean squared error. For data augmentation, we set $k = 16$ in NBB, and combine the constructed data with training data. Figure 4 illustrates the original and augmented training data, and compares the training and validation errors with and without data augmentation. We can see that without augmentation the network starts to overfit around epoch 100, while with augmentation the network trains faster, continues to improve over time, and has a lower error.

5. Discussion. In this section we discuss the determination of hyper-parameters for NBB: kernel bandwidth $h$, ridge dimension $d$, and number of neighbors $k$.

Kernel bandwidth $h$ should be selected for optimal estimation of the density ridge. A good estimate should resemble the shape of the true ridge while bias can be well tolerated, because with a smooth frame, NBB can correct for bias away from the estimated ridge. Silverman’s rule-of-thumb bandwidth tends to oversmooth the ridge, because the true density is supposed to have a salient geometric structure rather than been an isotropic Gaussian. Maximum likelihood bandwidth tends to be too small, such that the estimated ridge often has isolated points. We use an oversmoothing parameter $\alpha$, usually between 2 and 4, and good estimates can be often obtained across a wide range of $\alpha$ values. [8] gave a method to select $h$ that minimizes coverage risk estimates.

Ridge dimension $d$ is often apparent in specific problems. In low-dimensional problems with $n \leq 3$, the structure can often be examined visually. In regression, $d$ is the number of explanatory variables. In identifying implicit relations in a system, such as by symbolic regression or sparse regression, $d = n - c$ is the system’s degree of freedom, where $c$ is the number of constraint equations. If data is generated from a manifold, possible subject to ambient noise, $d$ is the manifold dimension. If no external information is available to determine
d, we can use eigengaps of the Hessian $H = \nabla \nabla \log \hat{p}_h$; find $c \in \{1, \ldots, n - 1\}$ with the largest
\[
\min \{ \lambda c_{c+1} - \lambda c : x \in X \}.
\]

Number of neighbors $k$ determines the amount of new data constructed by NBB, and we would prefer it to be as large as possible. For an estimated ridge point $\hat{r}_i$, $k_i$ should not exceed the largest local smooth frame containing the point. And the faster the distributions on normal spaces vary over the ridge, the smaller $k$ should be. If a global smooth frame can be constructed and the noises are identical across the ridge, we can set $k = n$. Typically, we let $k = \varepsilon n$, with $\varepsilon \in (0, 1/2]$. One criteria is that given $\hat{r}_i$, the normal vectors $[\hat{n}]_{K(i,j)}$ should be uni-modal. So if mode estimation on $[\hat{n}]_{K(i,j)}$ gives multiple points, $k_i$ should be decreased.

6. Conclusion. We introduced normal-bundle bootstrap, a method to resample data sets with salient geometric structure. The constructed new data are consistent with the distributions on normal spaces, and we demonstrated its uses in inference and data augmentation.

Acknowledgments. The authors thank Ernest Fokoue of Rochester Institute of Technology for valuable discussions.

Appendix A. Algorithms. Here are some algorithms used in Algorithm 3.1 for normal-bundle bootstrap. KNN for $k$-nearest neighbors is a common algorithm and therefore not listed.

SCMS is an implementation of subspace-constrained mean shift [17] for ridge estimation, where we use the logarithm of a Gaussian kernel density estimate. Note that density estimate $\hat{p}_h$ in the algorithm input is replaced with $(X, h)$ since we are assuming a Gaussian kernel. Note that this naive implementation can be accelerated using local data and Newton-like methods.

**Algorithm A.1** SCMS($y; X, h, d, \theta_0 = 0.05$)

1: repeat
2: $z_i \leftarrow (x_i - y)/h$, for $i \in N$
3: $c_i \leftarrow \exp(-\text{sum}(z_i^2)/2)$, for $i \in N$
4: $p_i \leftarrow c_i / \sum_{i=1}^{N} c_i$, for $i \in N$
5: $r_{pz,i} \leftarrow \sqrt{p_i z_i}$, for $i \in N$
6: $s_{pz} \leftarrow \sum_{i=1}^{N} p_i z_i$
7: $r_{pz}^T s_{pz} - s_{pz} s_{pz} = V A V$  \(\triangleright\) eigen-decomposition
8: $m_c \leftarrow N^{-1} \sum_{i=1}^{N} c_i$
9: $m_{cz} \leftarrow N^{-1} \sum_{i=1}^{N} c_i z_i$
10: $m \leftarrow h m_{cz}/m_c$  \(\triangleright\) mean-shift vector
11: $s \leftarrow (I - V d V^T) m$  \(\triangleright\) SCMS vector
12: $\theta \leftarrow m^T s / \sqrt{\text{sum}(m^2) \cdot \text{sum}(s^2)}$
13: $y \leftarrow y + s$
14: until $\theta > \theta_0$  \(\triangleright\) convergence criteria
15: return $(y, V, \lambda)$

**SmoothFrame** constructs smooth frames of the normal bundle of an estimated density ridge, where procedure **Align** adapts the moving frame algorithm [27] for the normal bundle. This
Algorithm A.2 SmoothFrame($R, V_c, c, j = 1$)

1. $(K, D) \leftarrow \text{KNN}(R, N - 1)$ \hspace{1cm} \triangleright \text{index and distance matrices of nearest neighbors}
2. $k \leftarrow \text{repeat}(1, N)$ \hspace{1cm} \triangleright \text{$K, D$-indices of nearest unaligned neighbor}
3. $E[j] \leftarrow V_c[j]$ \hspace{1cm} \triangleright \text{initial reference orthonormal $c$-frame}
4. $b \leftarrow 1$ \hspace{1cm} \triangleright \text{a-index of the last aligned point}
5. \textbf{while} $b < N$ \textbf{do}
6. \hspace{1cm} $a[b] \leftarrow j$ \hspace{1cm} \triangleright \text{indices of points in order of alignment}
7. \hspace{1cm} replace($K, j, \text{NULL}$) \hspace{1cm} \triangleright \text{remove indices of aligned points}
8. \hspace{1cm} \textbf{for all} $i$ \textbf{in} $a$ \textbf{do}
9. \hspace{1cm} \hspace{1cm} $k[i] \leftarrow k[i] + 1$ \hspace{1cm} \triangleright \text{maintain the property of } k
10. \hspace{1cm} \textbf{end for}
11. \hspace{1cm} $i \leftarrow a[\text{which.min}(D[a, k[a]])]$ \hspace{1cm} \triangleright \text{index of aligned point closest to the unaligned}
12. \hspace{1cm} $j \leftarrow K[i, k[i]]$ \hspace{1cm} \triangleright \text{index of the next point to align}
13. \hspace{1cm} Align($j, i$) \hspace{1cm} \triangleright \text{align } E[j] \text{ to } E[i]
14. \hspace{1cm} $b \leftarrow b + 1$
15. \textbf{end while}
16. \hspace{1cm} return $E$

19. \textbf{procedure} Align($j, i$)
20. \hspace{1cm} $E[j] \leftarrow V_c[j]$ \hspace{1cm} \triangleright \text{initial orthonormal $c$-frame}
21. \hspace{1cm} $\Theta \leftarrow E[j]^T E[i]$ \hspace{1cm} \triangleright \text{cosine matrix to reference frame } E[i]
22. \hspace{1cm} $\Theta = A \Sigma B^T$ \hspace{1cm} \triangleright \text{singular value decomposition}
23. \hspace{1cm} $Q \leftarrow AB^T$ \hspace{1cm} \triangleright \text{rotation matrix}
24. \hspace{1cm} $E[j] \leftarrow E[j]Q$ \hspace{1cm} \triangleright \text{aligned orthonormal $c$-frame}
25. \textbf{end procedure}

The algorithm recursively aligns the nearest unaligned point, which is “optimized” for stability but not for speed. It might be faster if using one reference frame for a neighborhood, such that the neighborhoods cover the data set. Moreover, when $c$ is large, only the top among the bottom-$c$ eigenvectors are significant to correct for biases introduced in ridge estimation, so a smooth subframe of the normal bundle suffice, which saves computation and storage. For the remaining normal directions, assuming negligible bias to the true ridge and radial symmetry (in addition to unimodality) of noise distribution, one may bootstrap the norm of the residual noise and multiply it with a random residual direction.

This algorithm, as written, assumes that a smooth global frame exists for the normal bundle of the estimated density ridge, or equivalently, that the normal bundle is trivial. The normal bundle of a density ridge does not need to be trivial, or not even orientable. Consider the uniform distribution on a Mobius band in the Euclidean 3-space, under a small additive Gaussian noise, the 2d density ridge includes the band, so the estimated density ridge approximates the band, which is non-orientable. Therefore, an (estimated) density ridge does not need to admit a smooth global frame for its normal bundle. In case the normal bundle
is not trivial, several smooth frames need to be constructed to cover the ridge. In terms of computation, one needs to run this algorithm on several subsets of the estimated ridge, such that for every point on the ridge, there is a frame that contains enough neighbors to the point.

On the other hand, for a constraint manifold, i.e. regular level set $\mathcal{M} = F^{-1}(0)$, its normal bundle is trivial (see [16, 10-18]), admits a smooth global frame (see [16, 10.20]), and it is orientable (see [16, 15-8]); in particular, the Jacobian $J^T(x)$ is a smooth/$C^{k-1}$ global frame for $\mathcal{N}\mathcal{M}$. By QR decomposition where $R$ has all positive diagonal entries, $Q(x)$ is smooth/$C^{k-1}$ orthonormal global frame for $\mathcal{N}\mathcal{M}$. Because non-orientable submanifolds of Euclidean spaces (e.g. the Mobius band) do not have global frames, they cannot be constraint manifolds.

**Appendix B. List of Symbols.** Here we provide the system of symbols we used in this article.

**Manifold:**
- $\mathbb{R}^n$, Euclidean n-space;
- $(\mathcal{M}, g)$, Riemannian submanifold of dimension $d$ with induced Riemannian metric;
- $\mu^0_\mathcal{M}(A) = \int_A dV_g / \int_\mathcal{M} dV_g$, normalized Hausdorff measure, a reference probability measure on the submanifold;
- $\mu^\mathcal{M}$, $\mu^\mathcal{M} = p^\mathcal{M} \mu^0_\mathcal{M}$, probability density/measure on the submanifold;
- $\mathcal{T}_p\mathcal{M}$, $\mathcal{N}_p\mathcal{M}$, tangent/normal space at a point on the submanifold;
- $\mathcal{N}\mathcal{M} = \bigcup_{p \in \mathcal{M}} \mathcal{N}_p\mathcal{M}$, normal bundle of the submanifold;

**Fiber bundle:**
- $(\mathcal{B}, \pi, \Phi)$, fiber bundle, a tuple of total space, projection, and trivialization;
- $\mathcal{M} = \pi(\mathcal{B})$, base space of the bundle, a manifold;
- $\mathcal{F}_r = \pi^{-1}(r)$, fiber over a point on the base space;
- $\mathcal{B}|_S = \pi^{-1}(S)$, restriction of a fiber bundle to a subset of its base space;
- $\Phi(x) = (\pi(x), x - \pi(x))$, trivialization of the normal bundle;
- $\mathcal{D} = \Phi(\mathcal{B}) \subset \mathcal{N}\mathcal{M}$, trivialized normal bundle;
- $\mu^\mathcal{M} = \mu \circ \pi^{-1}$, measure induced by projection on the base space;
- $\mu^\mathcal{F} = p^\mathcal{F} \mu^0_\mathcal{M}$, $p^\mathcal{F} = \frac{p}{\int_{\mathcal{F}} p \, dV_g}|_{\mathcal{F}}$, measure induced on each fiber, and its density function;

**Dynamical system:**
- $g(x) = \nabla p(x)$, $H(x) = \nabla \nabla p(x)$, gradient/Hessian of density function;
- $V$, $\Lambda = \text{diag}(\lambda)$, matrices of eigenvectors/eigenvalues of the Hessian;
- $U = V_d V_d^T$, $L = I - U$, orthonormal frames of the top-$d$/bottom-$c$ eigenvectors of the Hessian;
- $\mathcal{R} = \{x \in \mathbb{R}^n : \lambda_c(x) < 0, L(x) g(x) = 0\}$, density ridge of dimension $d$;
- $\mathcal{V}(x) = L(x) g(x)$, subspace-constrained gradient field;
- $\phi(t, x)$, $\phi^t(x)$, semi-flow generated by $\mathcal{V}$, and its time-$t$ map;
- $(\mathcal{B}, \phi^\infty)$, normal bundle of the density ridge (basin of attraction as total space, and time-infinite map as projection);
- $\mathcal{U}$, neighborhood on density ridge;

**Algorithm:**
- $X$, data set of $N$ points;
- $\hat{p}_h(x)$, estimated density function with kernel bandwidth $h$;
\[ m(x) = h^2 \hat{\eta}(x)/\hat{\eta}_h(x), \text{ mean-shift vector based on Gaussian kernel;} \]
\[ s(x) = L(x)m(x), \text{ subspace-constrained mean-shift vector;} \]
\[ \alpha, \text{ smoothing factor;} \]
\[ k, \text{ number of nearest neighbors;} \]
\[ x_i, \hat{r}_i = \pi(x_i), \hat{u}_i = x_i - \hat{r}_i, \hat{x}_{ij} = \hat{r}_i + \hat{u}_{ij}, \text{ data point, ridge point, normal vector, and constructed data point;} \]

REFERENCES

[1] P.-A. Absil, R. Mahony, and B. Andrews, *Convergence of the iterates of descent methods for analytic cost functions*, SIAM Journal on Optimization, 16 (2005), pp. 531–547, https://doi.org/10.1137/040605266.

[2] Y. Bengio, A. Courville, and P. Vincent, *Representation learning: A review and new perspectives*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 1798–1828, https://doi.org/10.1109/TPAMI.2013.50.

[3] Y. Bengio, L. Yao, G. Alain, and P. Vincent, *Generalized denoising auto-encoders as generative models*, in Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013, pp. 899–907, http://papers.nips.cc/paper/5023-generalized-denoising-auto-encoders-as-generative-models.

[4] A. Bhattacharya and R. Bhattacharya, *Nonparametric Inference on Manifolds: With Applications to Shape Spaces*, Cambridge University Press, 2012, https://doi.org/10.1017/CBO9781139094764.

[5] M. Brubaker, M. Salzmann, and R. Urtasun, *A family of MCMC methods on implicitly defined manifolds*, in Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, vol. 22, 2012, pp. 161–172, http://proceedings.mlr.press/v22/brubaker12.html.

[6] S. Byrne and M. Girolami, *Geodesic monte carlo on embedded manifolds*, in Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence, 2003, pp. 37–46, http://papers.nips.cc/paper/992-geodesic-monte-carlo-on-embedded-manifolds.

[7] Y.-C. Chen, *Solution manifold and its statistical applications*, arXiv, 2020, https://arxiv.org/abs/2002.05297.

[8] Y.-C. Chen, C. R. Genovese, S. Ho, and L. Wasserman, *Optimal ridge detection using coverage risk*, in Advances in Neural Information Processing Systems 25, 2012, pp. 316–324, http://papers.nips.cc/paper/3990-optimal-ridge-detection-using-coverage-risk.

[9] Y.-C. Chen, C. R. Genovese, and L. Wasserman, *Asymptotic theory for density ridges*, Annals of Statistics, 43 (2015), pp. 1806–1928, https://doi.org/10.1214/15-AOS1329.

[10] Y. Chikuse, *Statistics on Special Manifolds*, vol. 174, Springer-Verlag, New York, 2003, https://doi.org/10.1007/978-0-387-21540-2.

[11] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman, *Nonparametric ridge estimation*, Annals of Statistics, 42 (2014), pp. 1511–1545, https://doi.org/10.1214/14-AOS1218.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016, https://mitpress.mit.edu/books/deep-learning.

[13] T. Hastie and W. Stuetzle, *Principal curves*, Journal of the American Statistical Association, 84 (1989), pp. 502–516, https://doi.org/10.1080/01621459.1989.10478797.

[14] M. W. Hirsch, *Differential Topology*, vol. 33, Springer, New York, NY, 1976, https://doi.org/10.1007/978-1-4684-9449-5.

[15] C. Lageman, *Konvergenz reell-analytischer gradientenähnlicher systeme*, diplomarbeit, Universität Würzburg, Würzburg, Germany, 2002.

[16] J. M. Lee, *Introduction to Smooth Manifolds*, vol. 218, Springer, New York, 2012, https://doi.org/10.1007/978-1-4419-9982-5.

[17] U. Ozertem and D. Erdogmus, *Locally defined principal curves and surfaces*, Journal of Machine Learning Research, 12 (2011), pp. 1249–1286, http://www.jmlr.org/papers/v12/ozertem11a.html.

[18] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, *Normalizing flows for probabilistic modeling and inference*, 2019, https://arxiv.org/abs/1912.02762.

[19] V. Patrangenaru and L. Ellingson, *Nonparametric Statistics on Manifolds and Their Applications to
[20] C. Soize and R. Ghanem, *Data-driven probability concentration and sampling on manifold*, Journal of Computational Physics, 321 (2016), pp. 242–258, https://doi.org/10.1016/j.jcp.2016.05.044.

[21] C. Soize and R. Ghanem, *Probabilistic learning on manifolds*. arXiv, 2020, https://arxiv.org/abs/2002.12653.

[22] P. Vincent and Y. Bengio, *Manifold parzen windows*, in Proceedings of the 15th International Conference on Neural Information Processing Systems, MIT Press, 2002, pp. 849–856, http://papers.nips.cc/paper/2203-manifold-parzen-windows.

[23] L. Wasserman, *Topological data analysis*, Annual Review of Statistics and Its Application, 5 (2018), pp. 501–532, https://doi.org/10.1146/annurev-statistics-031017-100045.

[24] C.-F. J. Wu, *Jackknife, bootstrap and other resampling methods in regression analysis*, Annals of Statistics, 14 (1986), pp. 1261–1295, https://doi.org/10.1214/aos/1176350142.

[25] R. Zhang, *Newton retraction as approximate geodesics on submanifolds*, 2020, https://arxiv.org/abs/2006.14751.

[26] R. Zhang, P. Wingo, R. Duran, K. Rose, J. Bauer, and R. Ghanem, *Environmental economics and uncertainty: Review and a machine learning outlook*, Oxford Research Encyclopedia of Environmental Science, (2020), https://doi.org/10.1093/acrefore/9780199389414.013.572.

[27] W. C. Rheinboldt, *On the computation of multi-dimensional solution manifolds of parametrized equations*, Numerische Mathematik, 53 (1988), pp. 165–181, https://doi.org/10.1007/BF01395883.