Normal-bundle Bootstrap

Ruda Zhang*
rzhang@samsi.info
rzhang27@ncsu.edu

Roger Ghanem
ghanem@usc.edu

Jul 29, 2020

Talk outline

Overview: probabilistic distributions with geometric structure

Related literature

Bootstrap variants for regression

Normal-bundle bootstrap: algorithm, statistical theory

Experiments:

1. Inference: confidence set of density ridge
2. Data augmentation: regression by deep neural net

More on ridge estimation

Discussion

Ongoing research

Overview

When data sets are modelled as multivariate probability distributions, **such distributions often have salient geometric structure**.

Examples:

- Regression;
- Topological data analysis, incl. manifold learning.
- Deep learning;

Manifold distribution hypothesis:

> Natural high-dimensional data concentrate close to a nonlinear low-dimensional manifold.

Goal: generate new data which preserve the geometric structure of a probability distribution modelling a given data set.

This Paper: **Normal-bundle bootstrap**, a variant of the bootstrap resampling method.

Applications:

1. Inference of statistical estimators.
2. Data augmentation: increase training data diversity to reduce overfitting, without collecting new data.

Inspirations:

- Algorithms for nonlinear dimensionality reduction: subspace-constrained mean shift (SCMS) for density ridge estimation;
- Constructions in differential geometry: fiber bundle;

Related literature

Probabilistic learning on manifolds:

- Soize and Ghanem, *Data-driven probability concentration and sampling on manifold*, Journal of Computational Physics, 2016
- Soize and Ghanem, *Probabilistic learning on manifolds*, arXiv, 2020
- **Zhang**, Wingo, Duran, Rose, Bauer, and Ghanem, *Environmental economics and uncertainty: Review and a machine learning outlook*, Oxford Research Encyclopedia of Environmental Science, 2020

Ridge estimation:

- Hastie and Stuetzle, *Principal curves*, Journal of the American Statistical Association, 1989
- Ozertem and Erdogmus, *Locally defined principal curves and surfaces*, Journal of Machine Learning Research, 2011
- Genovese, Perone-Pacifico, Verdinelli, and Wasserman, *Nonparametric ridge estimation*, Annals of Statistics, 2014
- Chen, Genovese, and Wasserman, *Asymptotic theory for density ridges*, Annals of Statistics, 2015
- Chen, Genovese, Ho, and Wasserman, *Optimal ridge detection using coverage risk*, in Advances in Neural Information Processing Systems, 2015

Statistics on manifolds:

- Chikuse, *Statistics on Special Manifolds*, Springer, 2003
- Bhattacharya and Bhattacharya, *Nonparametric Inference on Manifolds: With Applications to Shape Spaces*, Cambridge University Press, 2012
- Patrangenaru and Ellingson, *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*, CRC Press, 2015
- Chen, *Solution manifold and its statistical applications*. arXiv, 2020

Bootstrap variants for regression

Data generating mechanism:

$$\mathbf{y} = f(x) + \boldsymbol{\varepsilon}(x), \quad \mu_\varepsilon = 0$$

Fitted (regression) model:

$$\hat{y} = \hat{f}(x)$$

Residuals:

$$u_i = y_i - \hat{y}_i, \quad i \in \{1, \ldots, N\}$$

Residual bootstrap (global):

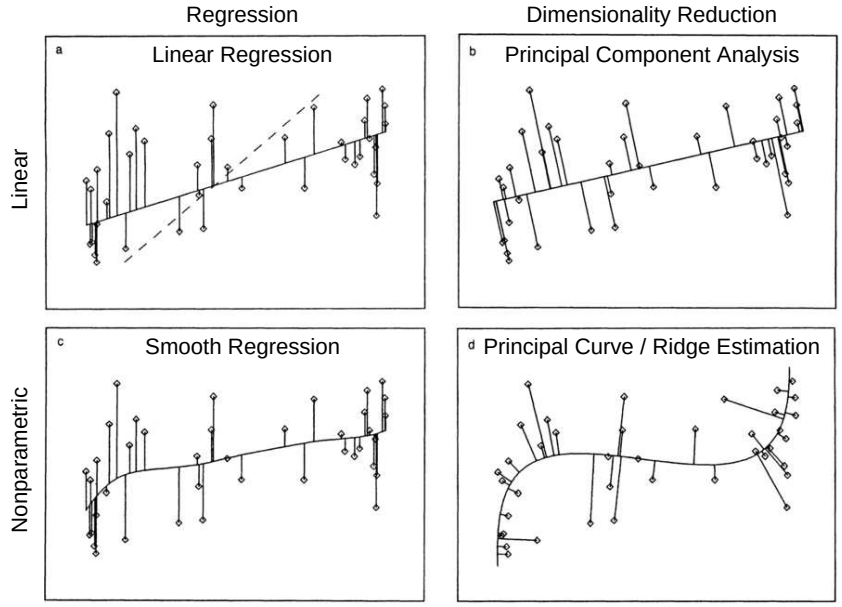
$$\tilde{y}_i = \hat{y}_i + u_{\mathbf{j}}, \quad \mathbf{j} \sim U\{1, \ldots, N\}$$

Wild bootstrap: *@WuCF1986*

$$\tilde{y}_i = \hat{y}_i + \mathbf{w} u_i, \quad \mathbf{w} \sim U\{-1, 1\}$$

(more generally, $\mu_w = 0, \sigma_w = 1$)

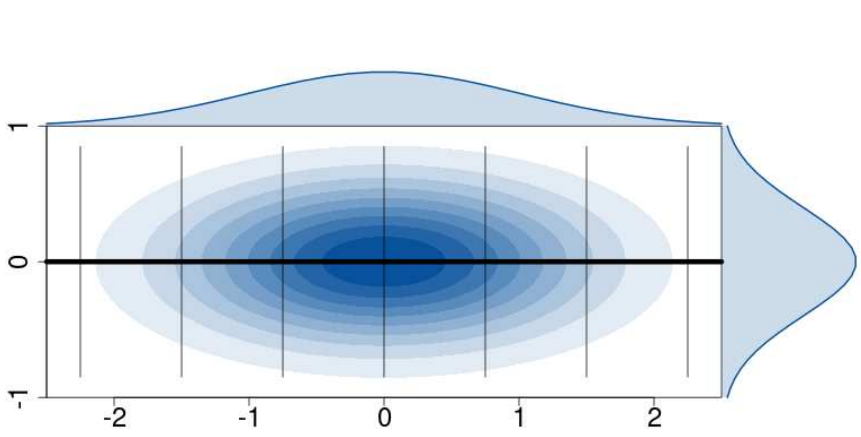
Replacing the model from regression to dimensionality reduction, we can obtain other bootstrap variants.

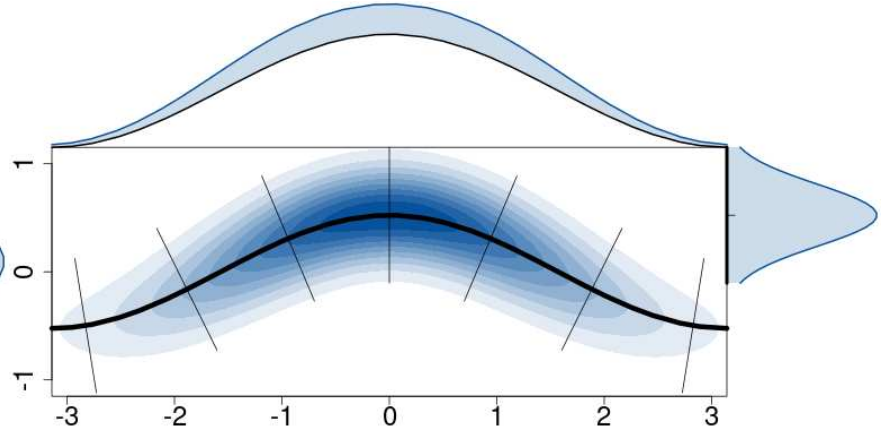


Dimensionality reduction vs. regression: linear and non-parametric methods. Adapted from [*@Hastie1989*, Fig 1].

Chien-Fu Jeff Wu, Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis, Annals of Statistics, 1986
Hastie and Stuetzle, *Principal curves*, Journal of the American Statistical Association, 1989
arxiv.org/abs/2007.13869

Density ridge and its normal bundle





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).

In general, density ridges are nonlinear, and its **normal bundle** (the collection of normal spaces) decomposes the original distribution into one on the ridge and one on each normal space.

Normal-bundle bootstrap

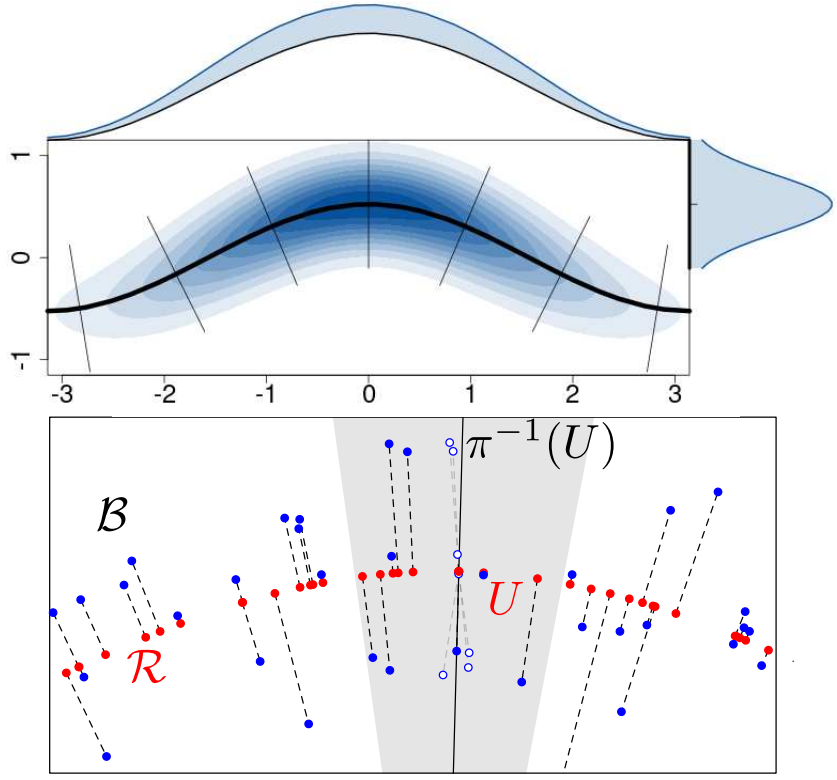
NBB algorithm: (marked steps are optional)

1. kernel bandwidth selection
2. ridge estimation
3. align bottom-c eigenvectors*
4. coordinates of normal vectors*
5. k-nearest neighbors on ridge
6. construct new data

$\mathrm{NBB}(X, d, \alpha, k)$:

1. $h \leftarrow \alpha \arg\max_h \sum_{i=1}^{N} \log \hat{p}_{h,-i}(x_i)$
2. $(\hat{r}_i, V_{c,i}) \leftarrow \mathrm{SCMS}(x_i; \log \hat{p}_h, d)$, for $i \in N$
3. $E \leftarrow \mathrm{SmoothFrame}(\hat{r}, V_c, n-d)$
4. $[\hat{n}]_i \leftarrow E_i^{\mathrm{T}}(x_i - \hat{r}_i)$, for $i \in N$
5. $K \leftarrow \mathrm{KNN}(\hat{r}, k)$
6. $\tilde{x}_{ij} \leftarrow \hat{r}_i + E_i[\hat{n}]_{K(i,j)}$, for $i \in N, j \in k$

The algorithm moves data points (blue) to the ridge (red) and, for each ridge point, picks neighboring ridge points (shade) and adds the projection vectors (dashed line) to construct new data points (hollow). Using a smooth frame can keep the constructed points in the normal space.



B, basin of attraction; R, density ridge; π, projection; U, a neighborhood in density ridge.

Example

Parabola with heteroscedastic error.

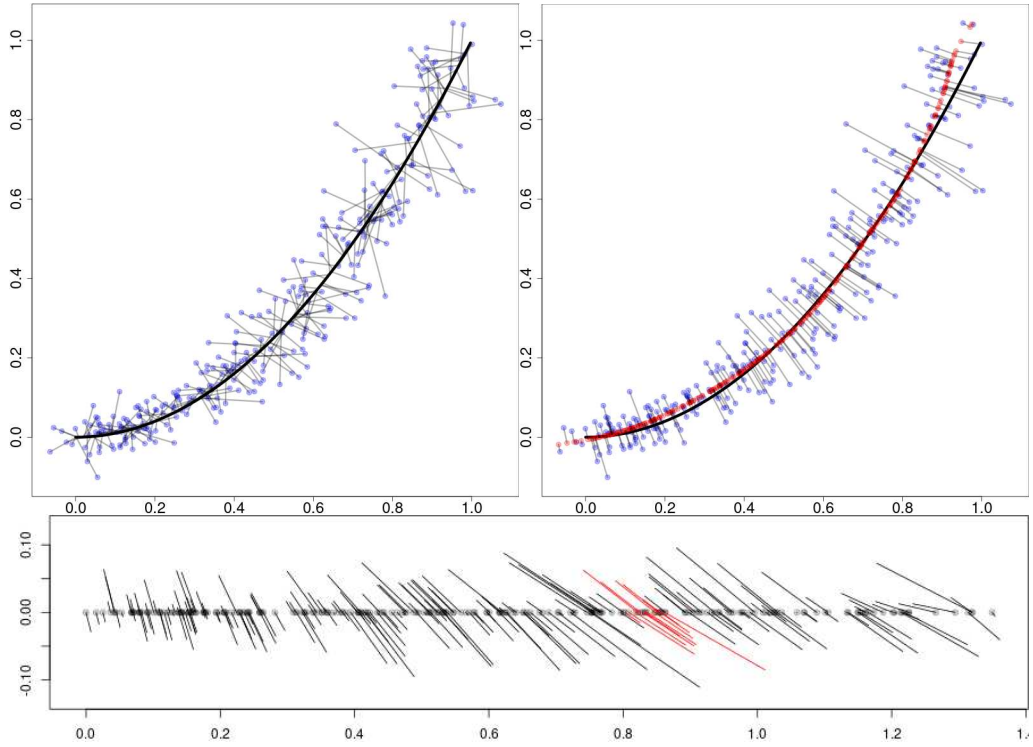
Data geneating mechanism:

$$\begin{aligned} \mathbf{x} &\sim U(0,1) \\ \mathbf{y} &= \mathbf{x}^2 \\ (\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) &\sim (\mathbf{x}, \mathbf{y}) + \sigma(\mathbf{x}) N(0, I) \\ \sigma(x) &= (1+x)/20 \end{aligned}$$

(left) data generation: noiseless model (black), data (blue).

(right) ridge estimation: estimated ridge (red).

(bottom) "residual" plot: projection vectors (black), neighborhood of a random ridge point (red). Neighborhood size k = N / 16. (Relative size: 6.25%)



Statistical theory

Assumption: In a neighborhood B of ridge $\mathcal{R}$,

- $p(x)$ is three times differentiable;
- $p(x)$ is sharply curved in normal spaces: $\lambda_c < -\beta$ and $\lambda_c < \lambda_{c+1} - \beta$, where $\beta > 0$;
- 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 (consistency): Let the assumptions hold for the measure μ in the basin of attraction $\mathcal{B}$, and the conditional measure $\mu^{\mathcal{F}_r}$ varies slowly over the ridge $\mathcal{R}$, then for each estimated ridge point $\hat{\mathbf{r}} = \pi_N(\mathbf{x}) = \phi_N^\infty(\mathbf{x})$, as sample size $N \to \infty$, the distributions of the constructed data points $\tilde{\mathbf{x}}_j$, $j \le k$, converge to the distribution restricted to the fiber of the estimated ridge point:

$$\tilde{\mathbf{x}}_j | \hat{\mathbf{r}} \xrightarrow{d} \mathbf{x}|_{\mathcal{F}_{\hat{r}}}$$

Finite-sample behavior is also very good:

As soon as the estimated ridge becomes close enough to the true ridge such that the conditional measures $\mu^{\mathcal{F}_{\hat{r}}}$ over the estimated ridge vary slowly, the conditional measures on neighboring fibers become similar to each other: $\mu^{\mathcal{F}_{\hat{r}_j}} \approx \mu^{\mathcal{F}_{\hat{r}}}$. This would suffice to make the constructed data distribute similarly to the original measure restricted to a fiber:

$$\tilde{\mathbf{x}}_j | \hat{\mathbf{r}} \dot{\sim} \mathbf{x}|_{\mathcal{F}_{\hat{r}}}$$

Even if the estimated ridge has a finite bias to the true ridge, it would not affect the conclusion.

Inference: confidence set of density ridge

NBB confidence set: (confidence level $1-\alpha$)

$$\hat{C}_N^{\text{NBB}} = \hat{\mathcal{R}}_N \oplus D_\alpha = \{\hat{r} + \hat{n} : \hat{r} \in \hat{\mathcal{R}}_N, \hat{n} \in D_\alpha(\hat{r})\}$$

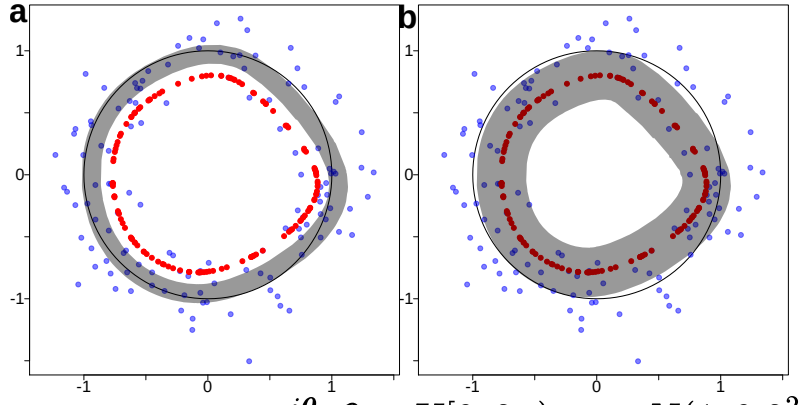
$$D_\alpha(\hat{r}) = \hat{m} \oplus \varepsilon_\alpha = \{\hat{n} \in N_{\hat{r}}\hat{\mathcal{R}}_N : d(\hat{n}, \hat{m}) < \varepsilon_\alpha\}$$

For $\hat{r}_i$, $\hat{m}_i$ is the mode estimated from the constructed points $\tilde{x}_{ij}$, and $\hat{\varepsilon}_{\alpha,i}$ is the α-upper quantile of $\{d(\hat{m}^*_{i,b}, \hat{m}_i)\}_{b=1}^B$.

Bootstrap confidence set: *@ChenYC2015*

$$\hat{C}_N^{\text{B}} = \hat{\mathcal{R}}_h \oplus \varepsilon_\alpha = \{x \in \mathbb{R}^n : d(x, \hat{\mathcal{R}}_h) < \varepsilon_\alpha\}$$

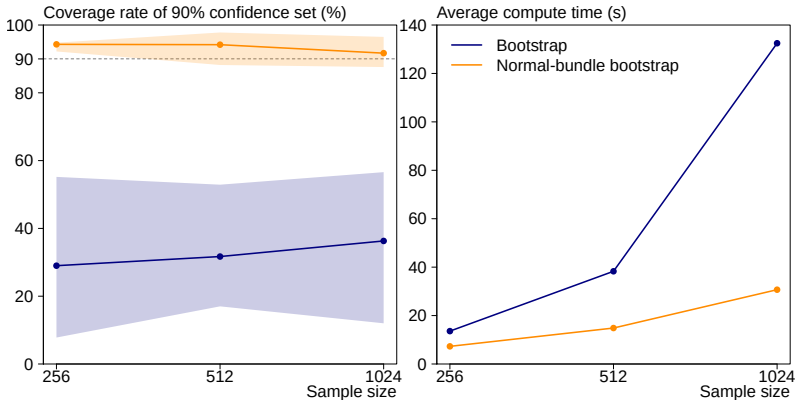
Here, $\hat{\varepsilon}_\alpha$ is the α-upper quantile of $\{d(\hat{r}^*_{i,b}, \hat{\mathcal{R}}_h)\}_{i=1\ldots N}^{b=1\ldots B}$.



Experiment: $\mathbf{x} = \mathbf{r}e^{i\boldsymbol{\theta}}$, $\boldsymbol{\theta} \sim U[0, 2\pi)$, $\mathbf{r} \sim N(1, 0.2^2)$. Data (blue), estimated ridge (red), true ridge $r \approx 1$ (black), 90% confidence sets of true ridge (gray) by NBB (a) vs. bootstrap (b). N = 128.

Chen, Genovese, and Wasserman, *Asymptotic theory for density ridges*, Annals of Statistics, 2015
arxiv.org/abs/2007.13869

Inference: metrics on NBB vs. boostrap



Metrics of NBB (orange) and bootstrap (blue) over an ensemble of samples: (c) coverage rate, mean (solid line) and 90% prediction interval (shade); (d) average computation time.

$\hat{C}^{\mathrm{NBB}}$ is valid throughout the range of sample sizes computed, while the validity of $\hat{C}^{\mathrm{B}}$ slowly improves. Moreover, $\hat{C}^{\mathrm{B}}$ is computationally costlier than $\hat{C}^{\mathrm{NBB}}$, due to repeated ridge estimation. Although repeated mode estimation is also costly, it is faster than ridge estimation of the same problem size.

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. *Hastie1989*

Hastie and Stuetzle, *Principal curves*, Journal of the American Statistical Association, 1989
arxiv.org/abs/2007.13869

Data augmentation: regression by deep neural net

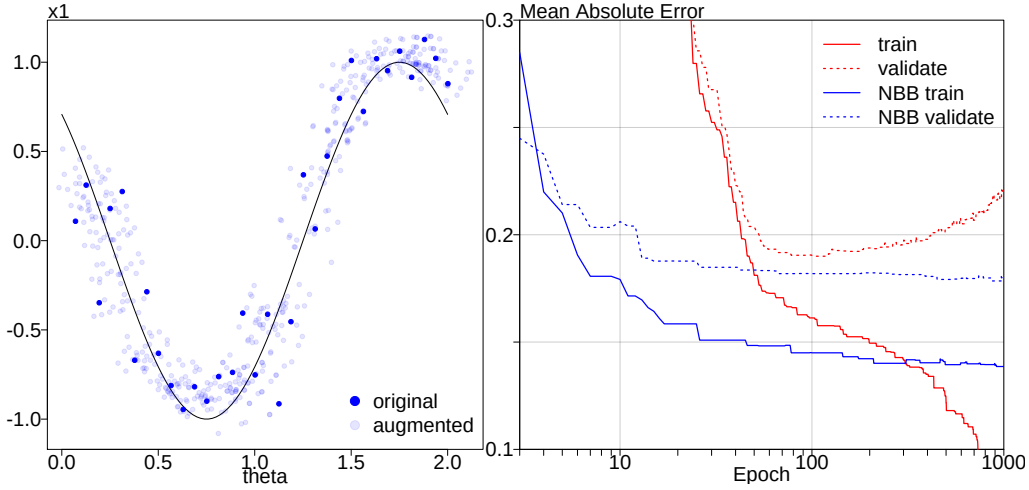
True model: $(x_j, y_j)(\theta) = (\cos(\pi(\theta + \gamma_j)), \sin(\pi(\theta + \gamma_j)))$, where $\theta \in [0, 2)$ and $\{\gamma_j\}_{j=1}^{l} \subset [0, 2)$.

Observation: $\left(\tilde{\boldsymbol{\theta}}, (\tilde{\mathbf{x}}_j, \tilde{\mathbf{y}}_j)_{j=1}^{l}\right) \Big| \theta \sim \left(\theta, (x_j(\theta), y_j(\theta))_{j=1}^{l}\right) + N(0, 0.2^2 I)$.

Training dataL $(\tilde{\theta}_i, (\tilde{x}_{ij}, \tilde{y}_{ij})_{j=1}^{l})_{i=1}^{N}$. Let $l = 8$ and $N = 32$, so the training data is a 32×17 matrix.

The neural network is a sequential model with 4 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. Set $k = 16$ in NBB.

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.



Data augmentation. (left) original and augmented data, showing (θ, x_1) only. Noiseless true model in black line. (right) training and validation error with and without NBB.

Density ridge

Definition: Ridge of dimension $d \in \{0, \dots, n\}$ 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:

$$\text{Ridge}(f, d) := \{x \in \mathbb{R}^n : \lambda_c < 0, Lg = 0\}$$

Here, Hessian $H = \nabla\nabla f$ has an eigen-decomposition $H = V \Lambda V^{\mathrm{T}}$, 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 $U = V_d V_d^{\mathrm{T}}$, $L = V_c V_c^{\mathrm{T}} = I - U$, and gradient $g = \nabla f$.

Assumption: 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 $d \in \{1, \dots, n-1\}$, $\text{Ridge}(p, d) \subset D$ is an embedded d-dimensional submanifold of $\mathbb{R}^n$.

Estimation by subspace-constrained mean shift (SCMS): *@Ozertem2011*

- subspace-constrained gradient flow:

$$v = Lg$$

- SCMS update:

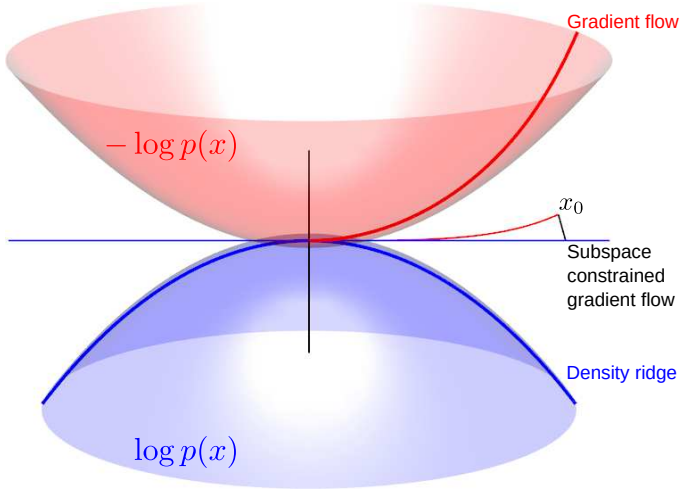
$$x_{t+1} = x_t + \kappa_t L \hat{g}_h(x_t)$$
$$\kappa_t = h^2 / \hat{p}_h(x_t)$$

Asymptotic theory:

- Geometric and topological consistency and convergence rates of estimated ridge to true ridge and hidden manifold. *@Genovese2014*
- Bootstrap gives consistent confidence sets of smoothed ridge. *@ChenYC2015*

Ozertem and Erdogmus, *Locally defined principal curves and surfaces*, Journal of Machine Learning Research, 2011.
Genovese, Perone-Pacifico, Verdinelli, and Wasserman, *Nonparametric ridge estimation*, Annals of Statistics, 2014.
Chen, Genovese, and Wasserman, *Asymptotic theory for density ridges*, Annals of Statistics, 2015.
arxiv.org/abs/2007.13869

Subspace-constrained gradient flow

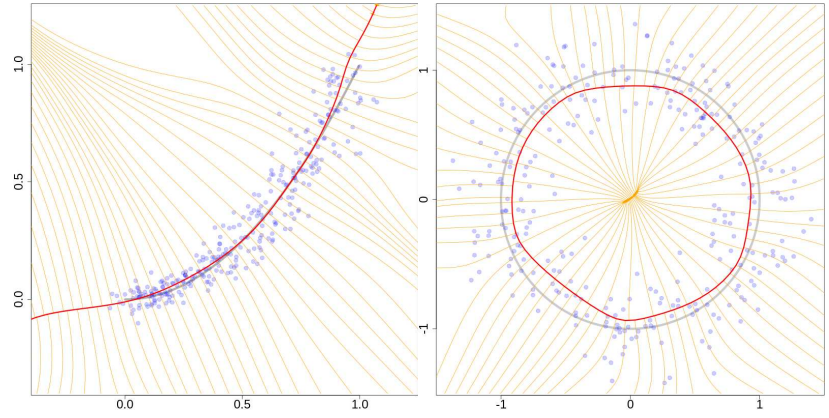


Gradient field:

$$g = \nabla \log \hat{p}_h$$

Subspace-constrained gradient field:

$$v = Lg$$
$$L = V_c V_c^{\mathrm{T}}$$



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.

Qualitative properties of dynamical system

With a stronger smoothness and manifold assumption, under the subspace-constrained gradient flow, each data point x_i converges to an estimated ridge point r_i with probability one:

Proposition (flow): 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 $\phi : \mathbb{R}_{\geq 0} \times D \mapsto D$. If $p(x)$ has a compact support $\overline{D}$, let $v(x) = 0$, $\forall x \in \partial D$, then $v(x)$ generates a flow $\phi : \mathbb{R} \times \overline{D} \mapsto \overline{D}$. Moreover, if $v(x)$ is locally Lipschitz or C^k, $k \geq 1$, then ϕ is locally Lipschitz or C^k, respectively.

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

Proposition (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 $\mathcal{B}$ is a subset of full Lebesgue measure and therefore has probability one: $\lambda(D \setminus \mathcal{B}) = 0$, $\mu(\mathcal{B}) = 1$, where λ is the Lebesgue measure on $\mathbb{R}^n$.

Discussion

Kernel bandwidth selection:

Maximum likelihood bandwidth tends to be too small, such that the estimated ridge often has isolated points. We use an oversmoothing parameter α, usually between 2 and 4, and good estimates can be often obtained across a wide range of α values. *@ChenYC2015b* gave a method to select h that minimizes coverage risk estimates.

Acceleration of SCMS:

A naive implementation of SCMS would be linearly convergent, and 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.

To reduce computation per iteration to $O(kNdn^2)$:

- local KDE: use the k-nearest data points for density estimation;
- partial eigen-decomposition: compute only the top-d eigen-pairs;

To reduce number of iterations, use Newton's method for root finding to obtain quadratic convergence:

$$x_{t+1} = x_t + L_t \delta_t$$

where δ_t solves

$$L_0 H_t L_0 \delta_t = -L_0 g_t$$

or

$$L_t H_t L_t \delta_t = -L_t g_t$$

The former only requires (partial) eigen-decomposition at the first step, and the latter has a larger convergence region. *@ZhangRD2020nr*

Chen, Genovese, Ho, and Wasserman, *Optimal ridge detection using coverage risk*, in Advances in Neural Information Processing Systems, 2015.
R. Zhang, Newton retraction as approximate geodesics on submanifolds, arXiv, 2020. arxiv.org/abs/2007.13869

Ongoing research

Probabilistic learning on manifolds (theory):

- **Normal-bundle bootstrap**.
- Kernel density estimation and sampling on submanifolds.
- Manifold-base joint probabilistic models.

Applications:

- Surrogate modeling, with application to oil spill models.

Numerical procedures:

- **Newton retraction as approximate geodesics on submanifolds**.
- Acceleration of SCMS.

Acknowledgment

The authors thank many people for support and valuable discussions:

- David Banks (Duke)
- Mansoor Haider (NCSU)
- Emily Griffith (NCSU)
- Ryan Murray (NCSU)
- Greg Forest (UNC Chapel Hill)
- Michael E Taylor (UNC Chapel Hill)
- Jeremy Louis Marzuola (UNC Chapel Hill)
- Ernest Fokoue (RIT)

This work was supported, in part, by the National Science Foundation grant DMS-1638521.