Bayesian manifold learning

David Dunson

Department of Statistical Science, Mathematics & ECE
Duke University

dunson@duke.edu

July 9, 2017

Joint work with Yun Yang & Didong Li

Overview

Motivation

- (Of course) very common to collect high-dimensional data
- Let p denote the ambient dimension of the data & n the sample size
- If $p \gg n$, we need to exploit lower-dimensional structure in the data
- Common to suppose data do not live everywhere in p-dimensional space
- May be concentrated near a *subspace* $\mathcal{M}$ having dimension d with $d \ll p$.

Subspace assumptions

- Suppose $X_i = (X_{i1}, \ldots, X_{ip})^T \in \mathcal{M} \subset \mathbb{R}^p$, for $i = 1, \ldots, n$, with $d \ll p$.
- $\mathcal{M}$ = *unknown* support of the data having intrinsic dimension d
- Most dimensionality reduction methods assume $\mathcal{M}$ is linear
- By learning the *mapping* $\Phi : \mathbb{R}^p \to \mathcal{M}$, we can replace the p-dimensional coordinates with d-dimensional coordinates
- Improve statistical efficiency & useful for interpretability

Common linear dimensionality reduction approaches

- Independent Component Analysis (ICA)
- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Factor Analysis (FA)

Nonlinear algorithms

- Sammon's Mapping
- Principal Curves and Manifolds
- Diffusion Maps
- Locally-Linear Embedding
- Hessian Locally-Linear Embedding
- Modified Locally-Linear Embedding
- Multicale Analysis of Plane Arrangements
- Geometric Multi-Resolution Analysis (GMRA)

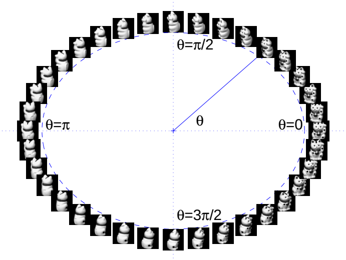
Our motivations

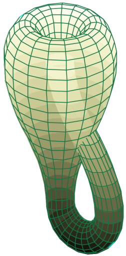
- UQ: we would like to incorporate uncertainty in dimensionality reduction & propagate this uncertainty
- Most approaches multistage - (i) estimate lower-dimensional coordinates; (ii) plug-in a second stage analysis
- Better dictionaries: we would like to flexibly represent a richer class of subspaces using fewer pieces
- Maybe $\mathcal{M}$ has locally varying curvature & is not a manifold

Two topics

- Bayesian manifold regression: we first consider the problem of manifold regression from a Bayes nonparametric perspective
- We show theoretically that (under some conditions) one can bypass manifold learning & rely on off-the-shelf Gaussian process
- Spherelets: we then propose a new dictionary for subspace learning using pieces of spheres
- A simple algorithm is shown to have state-of-the-art performance
- A Bayesian implementation for nonparametric subspace learning is also implemented

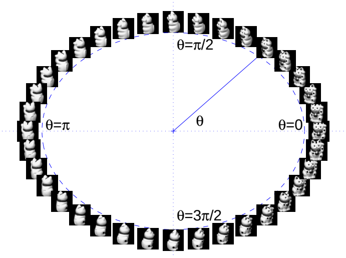
Regression on low dimensional manifold

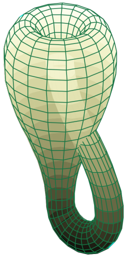




- Assumption: the covariates $X = (X_1, \ldots, X_p)^T$ lie on a d-dimensional manifold $\mathcal{M}$ in the ambient space $\mathbb{R}^p$

Regression on low dimensional manifold

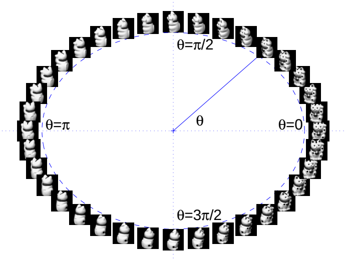


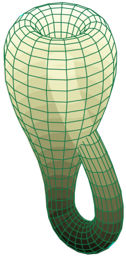


- Assumption: the covariates $X = (X_1, \ldots, X_p)^T$ lie on a d-dimensional manifold $\mathcal{M}$ in the ambient space $\mathbb{R}^p$
- For $\Sigma = C^\alpha(\mathcal{M})$, space of all α smooth functions on $\mathcal{M}$,

$$\text{minimax rate} \asymp n^{-\frac{\alpha}{2\alpha+d}}$$

Regression on low dimensional manifold





- Assumption: the covariates $X = (X_1, \ldots, X_p)^T$ lie on a d-dimensional manifold $\mathcal{M}$ in the ambient space $\mathbb{R}^p$
- For $\Sigma = C^\alpha(\mathcal{M})$, space of all α smooth functions on $\mathcal{M}$,

$$\text{minimax rate} \asymp n^{-\frac{\alpha}{2\alpha+d}}$$

- Ad hoc approach:
  1. project X into an estimated low dimensional space
  2. do nonparametric regression with projected coordinates

Regression on low dimensional manifold

- Drawbacks of the two stage approach: need to estimate high-dimensional nuisance parameters related to $\mathcal{M}$
- Question: possible to bypass the need of estimating $\mathcal{M}$, but can still exploit the low-dimensional manifold structure when exists?

Regression on low dimensional manifold

- Drawbacks of the two stage approach: need to estimate high-dimensional nuisance parameters related to $\mathcal{M}$
- Question: possible to bypass the need of estimating $\mathcal{M}$, but can still exploit the low-dimensional manifold structure when exists?
- Ye & Zhou (2008): least-square regularized method; Bickel & Li (2007): local polynomial regression
- Drawback: good performance relies on optimally choosing tuning parameters

Regression on low dimensional manifold

- Drawbacks of the two stage approach: need to estimate high-dimensional nuisance parameters related to $\mathcal{M}$
- Question: possible to bypass the need of estimating $\mathcal{M}$, but can still exploit the low-dimensional manifold structure when exists?
- Ye & Zhou (2008): least-square regularized method; Bickel & Li (2007): local polynomial regression
- Drawback: good performance relies on optimally choosing tuning parameters
- Our contribution: a tuning-free Bayesian nonparametric model based on Gaussian process prior,
  - near minimax optimal rate up to $\log n$ terms
  - adaptive to the unknown smoothness and manifold structure

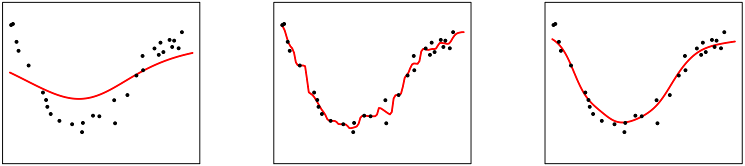
Gaussian process prior

- A Gaussian process $GP(m, K)$ is specified by:
  - Mean function
$$m(x) = E[f(x)]$$
  - Covariance function
$$K(x,y) = E\left[f(x) - Ef(x)\right]\left[f(y) - Ef(y)\right]$$

Gaussian process prior

- A Gaussian process $GP(m, K)$ is specified by:
  - Mean function
$$m(x) = E[f(x)]$$
  - Covariance function
$$K(x, y) = E\left[f(x) - Ef(x)\right]\left[f(y) - Ef(y)\right]$$
- Usually use zero mean function in the prior
- Popular choices for stationary covariance function K: square exponential kernel $K_a(x, y) = \exp\left\{-a^2||x - y||^2\right\}$, Matérn covariance kernel, etc.

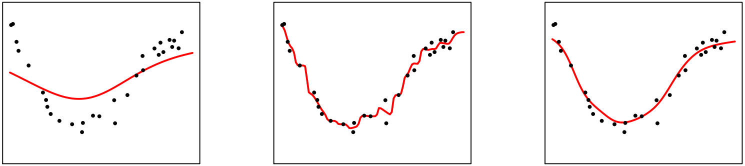
GP prior with random inverse bandwidth



- Adaptivity: let data choose the best inverse bandwidth parameter A
- Hierarchical prior structure:

$$f|A \sim GP(0, K_A), \quad A \sim F$$

GP prior with random inverse bandwidth



- Adaptivity: let data choose the best inverse bandwidth parameter A
- Hierarchical prior structure:

$$f|A \sim GP(0, K_A), \quad A \sim F$$

- Assume the p-variate truth f_0 has Hölder smoothness α
- van der Vaart & van Zanten (2009): If $A^p \sim Ga(a_0, b_0)$, then for M sufficiently large, posterior distribution satisfies

$$\Pi(||f - f_0||_2 \geq Mn^{-\alpha/(2\alpha+p)}(\log n)^{\beta} | D^n) \to 0 \text{ in } P_{f_0},\ n \to \infty$$

Bayesian manifold regression

- GP prior for the regression function:

$$f|A \sim GP(0, K_A), \quad A^d \sim Ga(a_0, b_0),$$

  with $K_a(x,y) = \exp\{-a^2||x-y||^2\}$ and $||\cdot||$ is the usual Euclidean norm in $\mathbb{R}^p$.
- Many ways to estimate the intrinsic dimension d: Likelihood based method (Levina & Bickel, 2004), multiscale SVD (Little et al. 2009)

Posterior convergence rate for Bayesian manifold regression

- $\mathcal{M}$ is a d-dimensional compact C^γ submanifold of $\mathbb{R}^p$
- Truth f_0 has smoothness $\alpha \leq \min\{2, \gamma - 1\}$

Theorem

For some sufficiently large $M > 0$, we have

$$\Pi(||f - f_0||_2 \geq M\epsilon_n \mid D^n) \to 0 \text{ in } P_{f_0},\ n \to \infty,$$
$$\text{with } \epsilon_n \asymp n^{-\frac{\alpha}{2\alpha+d}} (\log n)^{d+1}.$$

Switching gears - learning the subspace

- In the above approach, the subspace $\mathcal{M}$ is a nuisance parameter
- We show that you can bypass estimation of $\mathcal{M}$ in certain cases
- However, often there is interest in inference on the lower-dimensional structure in the data
- In addition, $\mathcal{M}$ may not be such a regular manifold
- $\mathcal{M}$ may have varying curvature & may be a stratified space

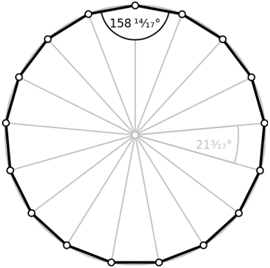
Pros and Cons of Current ‘Manifold Learning’ algs

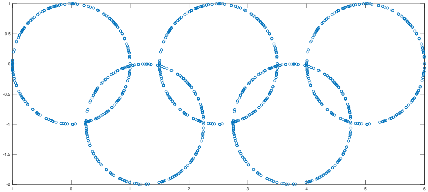
Pros

- Computational efficiency
- Work well for many “nice” manifolds

Cons

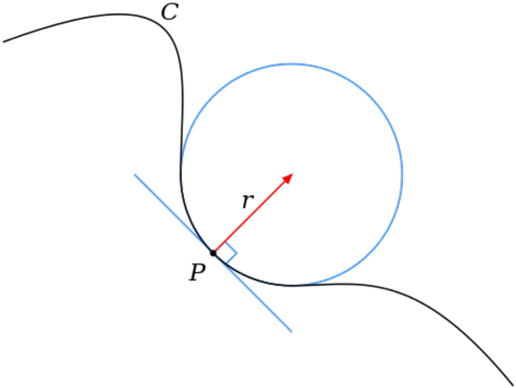
- Tend to find too many pieces (small scale) when the manifold has large curvature
- Can fail if $\mathcal{M}$ is not a manifold





New dictionary

- First order $\longrightarrow$ second order: $x^\top H x + f^\top x + c = 0$.
  Number of unknown parameters = $\frac{p(p+1)}{2} + p + 1 = O(p^2)$.
- $f(x) = f(a) + f'(a)(x-a) + R_1(x)$,
  $|R_1(x)| \leq M\frac{(x-a)^2}{2!}$, $|f''(x)| \leq M$.
- Curvature



Definition

A complete, simply connected constant sectional curvature Riemannian manifold is called a space form.

Theorem

Let M^d be a space form with curvature c, then

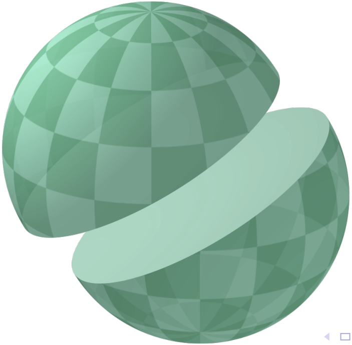
$$M^d \cong \begin{cases} S^d(\frac{1}{\sqrt{c}}) & c > 0 \\ \mathbb{R}^d & c = 0 \\ H^d(c) & c < 0 \end{cases},$$

where $S^d(\frac{1}{\sqrt{c}})$ is d dimensional sphere with radius $\frac{1}{\sqrt{c}}$ and $H^d(c)$ is d dimensional hyperbolic space with curvature c.

Spheres

Why spheres?

- Compactness.
- H^d has p symmetric axis, $N(H^d) = pN(S^d)$.
- Hyperplane=sphere with infinite radius.
- Projection Φ is easy to compute.
- Cell complex structure: $S^d = S^{d-1} \cup e_1^d \cup e_2^d$



Spherelets

Definition

Spherical error $\epsilon : \mathcal{F}^p \to \mathbb{R}_{\geq}$.

$$\epsilon(X) = \inf_{c,r} \frac{1}{n} \sum_{i=1}^{n} \inf_{x \in S(c,r)} \|X_i - x\|^2$$

- Riemannian divergence: $d_R(X, Y) = \epsilon(X \cup Y)$
- Euclidean divergence: $d_E(X, Y) = \inf_{i,j} \|x_i - y_j\|$
- Spherical divergence:

$$d_\lambda : \mathcal{F}^p \times \mathcal{F}^p \to \mathbb{R}_{\geq} : (X, Y) \mapsto d_R(X, Y) + \infty \mathbf{1}_{d_E(X,Y) > \lambda}$$

Algorithm

Input: X, ϵ, λ
Output: label, centers, radii, MSE

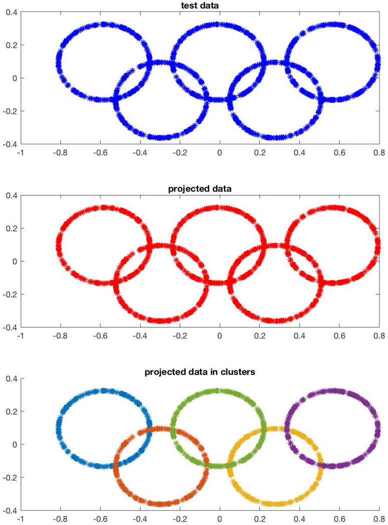
normalize X;
label=split(Xtrain,ϵ, λ);
label=merge(Xtrain, label, ϵ, λ);
find centers and radii;
calculate MSE;

$\lambda \Longleftrightarrow$ Euclidean$\Longleftrightarrow$ Topology (Path connectedness)
$\epsilon \Longleftrightarrow$Riemannian $\Longleftrightarrow$ Geometry (sphere)

Cross validation

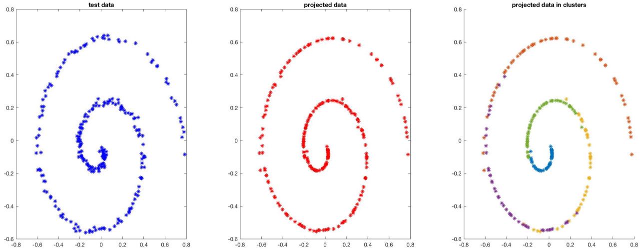
Olympic rings

$n = 1000$, $\epsilon = 10^{-5}$,$\lambda = 0.1$, MSE=1.7063×10^{-07}



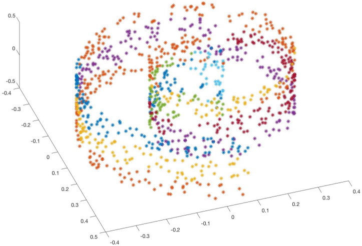
Noised spiral

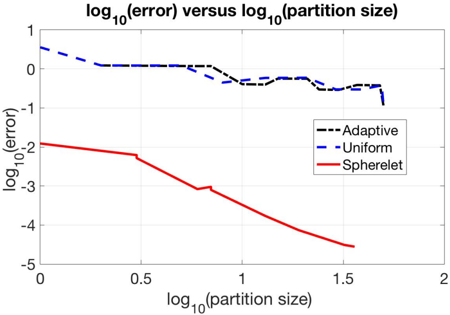
$n = 500$, $\epsilon = 10^{-4}$, $\lambda = 0.1$, $MSE = 1.4 \times 10^{-4}$.



Swissroll

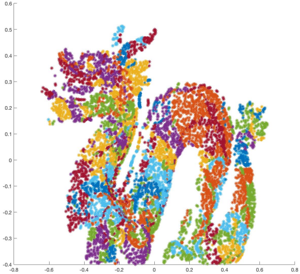
$n = 1000$

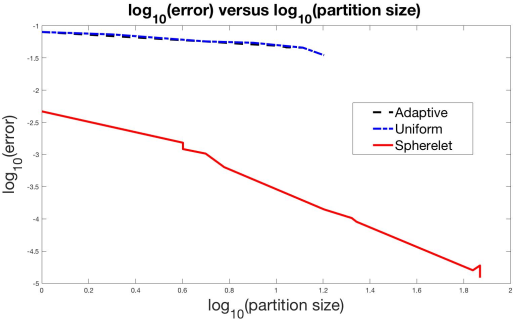




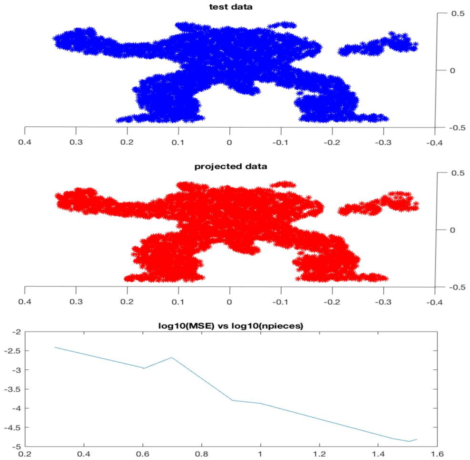
Dragon

$n = 1000$





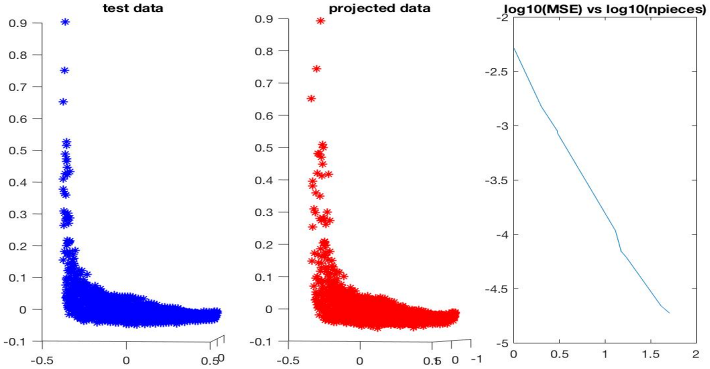
Armadillo



test data
0.5
0
-0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
projected data
0.5
0
-0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
log10(MSE) vs log10(npieces)
-2
-2.5
-3
-3.5
-4
-4.5
-5
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6

Atmosphere Boundary

Above the earth surface, where are two layers in the Troposphere: planetary boundary layer (L0) and free atmosphere (L1). The concentration of certain pollutants drop suddenly around this boundary, which provides an approach to estimate the altitude of this boundary. The data set contains the coordinates of the boundary surface



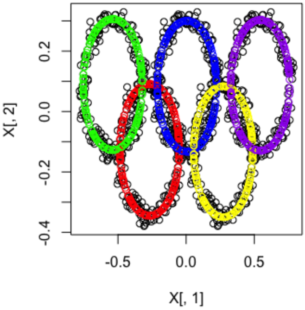
Bayesian nonparametric approach

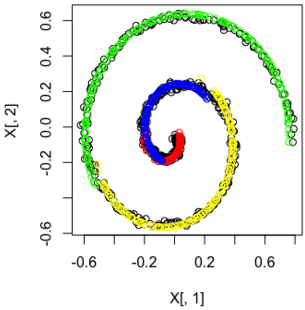
- We can also take a likelihood-based approach
- *Mixture of spherelets* model
- ith data point is generated from the hth sphere with probability π_h
- Data in component h are drawn by a von Mises-Fisher distribution with component-specific location & concentration
- Gaussian noise added to allow data to not fall exactly on a particular sphere

Computation - Mixture of spherelets model

- For a finite mixture model, an EM algorithm or MCMC algorithm can be easily implement for computation
- We initially take a fully Bayesian approach, placing default priors on the unknown parameters, and running MCMC
- A simple data augmentation Gibbs sampler can be defined - starting the chain at the output of our initial algorithm
- We use the over-fitted mixtures approach of Rousseau & Mengerson (2011) to allow uncertainty in the number of mixture components/clusters

Olympic Rings and Spiral-Bayesian version





Discussion

- The spherelets idea is *very* new & we are currently working on theoretical support
- One idea is to define the complexity of $\mathcal{M}$ using a spherelet covering number
- Allow manifolds & stratified spaces with locally varying curvature - more realistic than most notions in the literature
- The linear approximation covering number will be vastly larger than the spherelet cover
- Looking to obtain bounds on approximation error showing better performance for spherelets
- Also many interesting applied/methods directions - eg., data do not have to be real-valued vectors

Key References

- D. Li and D. Dunson, Efficient Manifold and Subspace Approximations with Spherelets, *https://arxiv.org/abs/1706.08263*, 2017.
- W. Liao and M. Maggioni , Adaptive Geometric Multiscale Approximations for Intrinsically Low-dimensional Data, *arXiv:1611.011*, 2016.
- G. Chen and M. Maggioni. Multiscale geometric and spectral analysis of plane arrangements. In *Conference on Computer Vision and Pattern Recognition*, 2011.
- W. Liao, M. Maggioni and S. Vigogna, Learning Adaptive Multiscale Approximations to Data and Functions near Low-Dimensional Sets, *IEEE* 2016.
- Y. Yang, D. Dunson, Bayesian manifold regression, *Annals of Statistics* 44(2):876-905.