Local linear regression on manifolds and its geometric implicaitons

Ming-Yen Cheng

Department of Mathematics, Hong Kong Baptist University

Joint work with Hau-Tieng Wu (Duke University)

Motivating Examples

- In the cryo Electron Microscopy problem (Frank, 2006), the images are located on the 3-dimensional manifold $SO(3)$.
- Radar signals can be modeled as being sampled from the Grassmannian manifold (Chikuse, 2003).
- The general manifold model for image and signal analysis is considered in Peyré (2009).
- Carlsson, et al. (2008) argued that natural images lie on a Klein bottle.

Local linear regression on unknown manifolds

- Y: scalar response variable
- X: p-dimensional predictor
- The distribution of X is assumed to be concentrated on a d-dimensional compact, smooth Riemannian manifold M embedded in $\mathbb{R}^p$ via the embedding $\iota : \mathrm{M} \hookrightarrow \mathbb{R}^p$.
- Consider the following regression model

$$Y = m(X) + \sigma(X)\,\epsilon, \tag{1}$$

where m and σ are functions defined on M, and ϵ is a random error independent of X with $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = 1$.

Local Linear Regression on Unknown Manifolds (MALLER)

Let $\{(X_l, Y_l)\}_{l=1}^{n}$ denote a random sample observed from the regression model (1) with $\mathcal{X} = \{X_l\}_{l=1}^{n}$ being sampled from the distribution of X.

Our nonparametric method to estimate the regression function m consists of the following four steps.

- Step 1: obtaining the intrinsic dimension d
- Step 2: reducing effects of the condition number
- Step 3: embedded tangent plane estimation
- Step 4: local linear regression on the tangent plane estimate

Step 1: Obtaining the intrinsic dimension

- Assume that we are given the intrinsic dimension d of the manifold M.
- If d is unknown a priori and needs to be estimated based on the data $\mathcal{X}$, estimate it by the maximum likelihood estimator proposed by Levina and Bickel (2005).
- Given that the sample size n is large enough, we assume the dimension estimate is correct and hence will not distinguish it from the true value of d from now on.

Step 2: reducing effects of the condition number

- $\mathcal{X} = \{X_1, \ldots, X_n\}$
- $\mathcal{N}^{\mathbb{R}^p}_{x,\delta} = \left\{X_j \in \mathcal{X} : \|X_j - x\|_{\mathbb{R}^p} < \sqrt{\delta}\right\}$: the set of Euclidean $\sqrt{\delta}$-neighbors of x
- $d(\cdot, \cdot)$: the geodesic distance
- $\mathcal{N}^{\mathrm{M}}_{x,\delta} = \left\{X_j \in \mathcal{X} : d(X_j, x) < \sqrt{\delta}\right\}$: the set of geodesic $\sqrt{\delta}$-neighbors of x from $\mathcal{X}$
- Apply the *self-tuning spectral clustering algorithm* (Zelnik-Manor and Perona, 2004) to the set $\mathcal{N}^{\mathbb{R}^p}_{x,\delta} \cup \{x\}$, and use the set

$$\mathcal{N}^{\mathrm{true}}_{x,\delta} := \left\{X_j \in \mathcal{N}^{\mathbb{R}^p}_{x,\delta} : X_j \text{ is in the same cluster as } x\right\}$$

as an estimate of $\mathcal{N}^{\mathrm{M}}_{x,\delta}$.

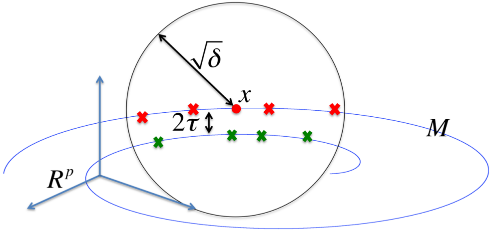


Figure : τ: reach, $1/\tau$: the condition number of M. The set of Euclidean $\sqrt{\delta}$-neighbors of x, $\mathcal{N}^{\mathbb{R}^p}_{x,\delta}$, consists of both the red and green crosses. The set of geodesic $\sqrt{\delta}$-neighbors of x, $\mathcal{N}^{\mathrm{M}}_{x,\delta}$, consists of the red crosses but not the green crosses.

Step 3: embedded tangent plane estimation

Tangent plane:

- $T_x\mathrm{M}$: the tangent plane of the manifold at $x \in \mathrm{M}$
- ι_*: the total differential of ι, that is, $\iota_* : T_x\mathrm{M} \to T_{\iota(x)}\mathbb{R}^p$
- $\iota_* T_x\mathrm{M}$: the embedded tangent plane into $\mathbb{R}^p$

Local PCA:

- Σ_x: the sample covariance matrix of $\mathcal{N}^{\text{true}}_{x,h_{\text{PCA}}}$.
- $\{U_k(x)\}_{k=1}^d$: the first d eigenvectors of Σ_x.
- Let B_x be the $p \times d$ matrix $B_x = \left[\ U_1(x) \quad \ldots \quad U_d(x)\ \right]$.

Projecting the design points onto a tangent plane estimate:

- For $l = 1, \ldots, n$, let $x_l = (x_{l,1}, \ldots, x_{l,d})^T = B_x^T(X_l - x)$: the projection of $X_l - x$ onto the affine space spanned by the orthonormal basis $\{U_k(x)\}_{k=1}^d$, which is an approximation to the embedded tangent plane $\iota_* T_x\mathrm{M}$.

Step 4: local linear regression on tangent plane estimate

- $K : [0, \infty] \to \mathbb{R}$: nonzero kernel function so that $K|_{[0,1]} \in C^1([0,1])$ and $K|_{(1,\infty]} = 0$
- $h > 0$: a bandwidth
- Let

$$\hat{\boldsymbol{\beta}}_x = \underset{\boldsymbol{\beta} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{l=1}^{n} \Big(Y_l - \beta_0 - \sum_{k=1}^{d} \beta_k x_{l,k} \Big)^2 \mathrm{I}_{\mathcal{N}_{x,h}^{\text{true}}}(X_l) K_h(X_l, x),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^T$, $K_h(X_l, x) := \frac{1}{h^{d/2}} K\left(\frac{\|X_l - x\|_{\mathbb{R}^p}}{\sqrt{h}}\right)$ and I is the indicator function.

- Let $\mathbb{X}_x = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix}^T$, and
$\mathbb{W}_x = \operatorname{diag}\left(K_h(X_1, x) \mathrm{I}_{\mathcal{N}_{x,h}^{\text{true}}}(X_1), \dots, K_h(X_n, x) \mathrm{I}_{\mathcal{N}_{x,h}^{\text{true}}}(X_n) \right)$.
- The functional $\hat{\beta}_x$ can be written as

$$\hat{\beta}_x = (\mathbb{X}_x^T \mathbb{W}_x \mathbb{X}_x)^{-1} \mathbb{X}_x^T \mathbb{W}_x Y,$$

where $Y = (Y_1, \dots, Y_n)^T$, if $(\mathbb{X}_x^T \mathbb{W}_x \mathbb{X}_x)^{-1}$ exists.
- The estimator of $m(x)$ we propose is given by

$$\hat{m}(x, h) := v_1^T \hat{\beta}_x = v_1^T (\mathbb{X}_x^T \mathbb{W}_x \mathbb{X}_x)^{-1} \mathbb{X}_x^T \mathbb{W}_x Y, \qquad (2)$$

where $v_k \in \mathbb{R}^{d+1}$ is the $(d+1) \times 1$ unit vector with the k-th entry being 1.

If the interest is to estimate the embedded gradient of m at x, the following estimator is considered:

$$\iota_* \widehat{\texttt{grad} m}(x) := \sum_{i=1}^{d} \widehat{\nabla_{\partial_i(x)} m}(x, h) U_i(x). \tag{3}$$

where $\texttt{grad}$ denotes the gradient,

$$\widehat{\nabla_{\partial_i(x)} m}(x, h) := v_{i+1}^T \hat{\boldsymbol{\beta}}_x, \tag{4}$$

and $\{\partial_i(x)\}_{i=1}^d$ is the orthonormal basis of $T_x\mathrm{M}$ closest to the estimated orthonormal basis $\{U_k(x)\}_{k=1}^d$.

Theoretical results

Notation

- Take the metric g to be the one such that, for $u, v \in T_x\mathrm{M}$,

$$g_x(u, v) := \langle \iota_* u, \iota_* v \rangle.$$

- The exponential map at $x \in \mathrm{M}$ is denoted as $\exp_x$.
- The volume form on M induced from g is denoted as $\mathrm{d}V$.
- Define the set of points close to the boundary $\partial\mathrm{M}$ with distance less than $\delta \geq 0$, where δ is small enough, as

$$M_\delta(x) = \left\{ y \in \mathrm{M} : \min_{y \in \partial \mathrm{M}} d(x, y) \leq \delta \right\},$$

where $d(x, y)$ is the geodesic distance between x and y.

- Denote by ∇ the Levi-Civita connection, Δ the Laplace-Beltrami operator, and Hess the second order covariant derivative operator on (M, g).

Probability density function of the random vector $X : \Omega \to \iota(\mathrm{M})$:

- X: a measurable function with respect to the probability space $(\Omega, \mathcal{F}, P)$
- $\tilde{\mathcal{B}}$: the Borel sigma algebra of $\iota(\mathrm{M})$.
- $\tilde{P}_X$: the probability measure of X defined on $\tilde{\mathcal{B}}$, induced from P.
- Assume that $\tilde{P}_X$ is absolutely continuous w.r.t. the volume measure $\mathrm{d}V$ so that $\mathrm{d}\tilde{P}_X(x) = f(\iota^{-1}(x))\iota_* \mathrm{d}V(x)$ for some $f \in C^2(\mathrm{M})$. That is, for an integrable function $\zeta : \iota(\mathrm{M}) \to \mathbb{R}$,

$$
\begin{aligned}
\mathbb{E}\zeta(X) &= \int_\Omega \zeta(X(\omega))\mathrm{d}P(\omega) = \int_{\iota(\mathrm{M})} \zeta(x)\mathrm{d}\tilde{P}_X(x) \\
&= \int_{\iota(\mathrm{M})} \zeta(x) f(\iota^{-1}(x))\iota_* \mathrm{d}V(x) = \int_{\mathrm{M}} \zeta(\iota(y)) f(y)\mathrm{d}V(y).
\end{aligned}
$$

In this sense we interpret f as the p.d.f. of X on M.

Assumptions:

(A1) $h \to 0$ and $nh^{d/2} \to \infty$ as $n \to \infty$.

(A2) f belongs to $C^2(\mathrm{M})$ and satisfies

$$0 < \inf_{x \in \mathrm{M}} f(x) \leq \sup_{x \in \mathrm{M}} f(x) < \infty. \tag{5}$$

(A3) For every given $h > 0$ and every point $x \in \mathrm{M}_{\sqrt{h}}$, the set $B^{\mathrm{M}}_{\sqrt{h}}(x) \cap \mathrm{M}$ contains a non-empty interior set.

(A4) Assume that $h_{\mathrm{PCA}}^{1/2} < \min(2\tau, \mathrm{inj}(\mathrm{M}))$ and $h^{1/2} < \min(2\tau, \mathrm{inj}(\mathrm{M}))$, where $\mathrm{inj}(\mathrm{M})$ is the injectivity radius of M and $1/\tau$ is the condition number of M.

Denote $\mu_{i,j} := \int_{B_1^{\mathbb{R}^d}(0)} K^i(\|u\|_{\mathbb{R}^d})\|u\|_{\mathbb{R}^d}^j \mathrm{d}u$ and we normalize K so that $\mu_{1,0} = 1$.

Theorem 1. Suppose $h_{\mathrm{PCA}} \asymp n^{-2/(d+1)}$ and $h \geq h_{\mathrm{PCA}}$. When $x \in \mathrm{M} \backslash \mathrm{M}_{\sqrt{h}}$, the conditional mean square error (MSE) for the estimator $\hat{m}(x, h)$ is

$$\mathrm{MSE}\{\hat{m}(x,h)|\mathcal{X}\} = h^2 \frac{\mu_{1,2}^2}{4d^2}(\Delta m(x))^2 + \frac{1}{nh^{d/2}} \frac{\mu_{2,0}\sigma^2(x)}{f(x)} + O_p(h^{5/2}) + O_p\Big(\frac{1}{n^{1/2}h^{d/4-2}} + \frac{1}{nh^{d/2-1}} + \frac{1}{n^{3/2}h^{3d/4}}\Big).$$

Thus, the minimal asymptotic conditional MSE is achieved when $h \asymp n^{-2/(d+4)}$.

For $x \in \mathsf{M}_{\sqrt{h}}$ and $h > 0$, define

$$\nu_{i,x} := \begin{bmatrix} \nu_{i,x,11} & \nu_{i,x,12} \\ \nu_{i,x,12}^T & \nu_{i,x,22} \end{bmatrix} := \begin{bmatrix} \int_{\frac{1}{\sqrt{h}}\mathcal{D}(x)} K^i(\|u\|)\mathrm{d}u & \int_{\frac{1}{\sqrt{h}}\mathcal{D}(x)} K^i(\|u\|)u^T\mathrm{d}u \\ \int_{\frac{1}{\sqrt{h}}\mathcal{D}(x)} K^i(\|u\|)u\mathrm{d}u & \int_{\frac{1}{\sqrt{h}}\mathcal{D}(x)} K^i(\|u\|)uu^T\mathrm{d}u \end{bmatrix},$$

$$\mathcal{D}(x) := \exp_x^{-1}(B^{\mathsf{M}}_{\sqrt{h}}(x) \cap \mathsf{M}) \subset T_x\mathsf{M},$$

$$C := \begin{bmatrix} 1 & 0 \\ 0 & h^{\frac{1}{2}} I_d \end{bmatrix}.$$

Here, I_k denotes the $k \times k$ identity matrix for any $k \in \mathbb{N}$.

Theorem 2. Suppose $x \in \mathrm{M}_{\sqrt{h}}$, $h_{\text{PCA}} \asymp n^{-2/(d+1)}$ and $h \geq h_{\text{PCA}}$. The conditional MSE of the estimator $\hat{m}(x,h)$ is

$$\mathrm{MSE}\{\hat{m}(x,h)|\mathcal{X}\} = \frac{h^2}{4}\frac{[\mathrm{tr}\big(\mathrm{Hess}m(x)\nu_{1,x,22}\big)]^2}{\nu_{1,x,11}^2} + \frac{v_1^T\nu_{1,x}^{-1}\nu_{2,x}\nu_{1,x}^{-1}v_1}{nh^{d/2}}\frac{\sigma^2(x)}{f(x)}$$
$$+O_p\Big(h^{3/2}h_{\text{PCA}}^{3/4} + h^{5/2}\Big) + O_p\Big(\frac{1}{n^{1/2}h^{d/4-2}} + \frac{1}{nh^{d/2-1/2}} + \frac{1}{n^{3/2}h^{3d/4}}\Big)$$

Corollary 1. Suppose $\partial\mathrm{M}$ is smooth, $x \in \mathrm{M}_{\sqrt{h}}$, $h_{\text{PCA}} \asymp n^{-2/(d+1)}$ and $h \geq h_{\text{PCA}}$. Then the asymptotic conditional bias of $\hat{m}(x,h)$ is a linear combination of the second order covariant derivative of m:

$$\mathbb{E}\{\hat{m}(x,h) - m(x)|\mathcal{X}\} = \frac{h}{2}\sum_{k=1}^{d} c_k(x)\nabla^2_{\partial_k,\partial_k}m(x)$$
$$+O_p(h^{1/2}h_{\text{PCA}}^{3/4} + h^{3/2}) + O_p\Big(\frac{1}{n^{1/2}h^{d/4-1}}\Big),$$

where $\{\partial_k\}_{k=1}^d$ is a normal coordinate around x and $c_k(x)$ is uniformly bounded for all $k = 1, \ldots, d$.

Theorem 3. Suppose $x \in \mathrm{M} \backslash \mathrm{M}_{\sqrt{h}}$, $h_{\mathrm{PCA}} \asymp n^{-2/(d+1)}$ and $h \geq h_{\mathrm{PCA}}$. The conditional MSE for the estimator $\widehat{\nabla_{\partial_i(x)} m}(x,h)$ given in (4) is

$$\mathrm{MSE}\{\widehat{\nabla_{\partial_i(x)} m}(x,h) | \mathcal{X}\}$$

$$= h^2 \left[\frac{\mu_{1,2}}{d} \frac{\nabla_{\partial_i} f(x)}{f(x)} \Delta m(x) - \frac{\mu_{1,2} d \int_{S^{d-1}} \theta^T \mathrm{Hess} m(x) \theta \theta \nabla_\theta f(x) \mathrm{d}\theta}{|S^{d-1}| f(x)} \right]^2$$

$$+ \frac{1}{n h^{\frac{d}{2}+1}} \frac{d \mu_{2,2} \sigma^2(x) f(x)}{\mu_{1,2}^2} + O_p(h^{\frac{5}{2}} + h^{\frac{3}{2}} h_{\mathrm{PCA}}^{\frac{3}{4}})$$

$$+ O_p\Big(\frac{1}{n^{\frac{1}{2}} h^{\frac{d}{4}-\frac{3}{2}}} + \frac{1}{n h^{\frac{d}{2}}} + \frac{1}{n^{\frac{3}{2}} h^{\frac{3d}{4}+1}} \Big),$$

where $\{\partial_i(x)\}_{i=1}^d$ is an orthonormal basis of $T_x \mathrm{M}$.

Theorem 4. Suppose $x \in \mathrm{M}_{\sqrt{h}}$, $h_{\mathrm{PCA}} \asymp n^{-2/(d+1)}$ and $h \geq h_{\mathrm{PCA}}$. The conditional MSE for the estimator $\widehat{\nabla_{\partial_i(x)} m}(x, h)$ given in (4) is

$$\begin{aligned}
&\mathrm{MSE}\{\widehat{\nabla_{\partial_i(x)} m}(x,h)|\mathcal{X}\} \\
&= h\Big(\frac{v_{i+1}^T \nu_{1,x}^{-1}}{2} \int_{\frac{1}{\sqrt{h}}\mathcal{D}(x)} K(\|u\|) u^T \mathrm{Hess} m(x) u \begin{bmatrix} 1 \\ u \end{bmatrix} \mathrm{d}u\Big)^2 \\
&+ \frac{v_{i+1}^T \nu_{1,x}^{-1} \nu_{2,x} \nu_{1,x}^{-1} v_{i+1}}{n h^{\frac{d}{2}+1}} \frac{\sigma^2(x)}{f(x)} + O_p\Big(h^{\frac{1}{2}} h_{\mathrm{PCA}}^{\frac{3}{4}} + h h_{\mathrm{PCA}}^{\frac{1}{2}}\Big) \\
&+ O_p\Big(\frac{1}{n^{\frac{1}{2}} h^{\frac{d}{4}-\frac{3}{2}}} + \frac{1}{n h^{\frac{d}{2}+\frac{1}{2}}} + \frac{1}{n^{\frac{3}{2}} h^{\frac{3d}{4}}}\Big),
\end{aligned}$$

where $\{\partial_i(x)\}_{i=1}^d$ is an orthonormal basis of $T_x\mathrm{M}$.

Bandwidth selection

I. Pilot bandwidth

The modified generalized cross-validation (mGCV) suggested in Bickel and Li (2007).

- For each X_l, choose a block of data points $\{(X_j, Y_j)\}_{j\in\mathcal{J}}$.
- The mGCV bandwidth, denoted as $h_{\text{mGCV},\hat{m}}$, is chosen to be the value of h in $\mathcal{H}_{\text{mGCV}}$ which minimizes

$$\text{mGCV}(h) = \Big(1 + 2\text{atr}_{\mathcal{J}}(h)\Big)\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\Big(Y_j - \hat{m}(X_j, h)\Big)^2,$$

where $\text{atr}_{\mathcal{J}}(h) := \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} v_1^T(\mathbb{X}_{X_j}^T\mathbb{W}_{X_j}\mathbb{X}_{X_j})^{-1}v_1 h^{-d/2}K(0)$.

II. Estimate the value of the conditional variance σ^2 at x:

- Define the residuals as $\hat{r}_l := \left(Y_l - \hat{m}(X_l, h_{\mathrm{mGCV},\hat{m}})\right)^2$, $l = 1, \ldots, n$.
- Let $(\hat{\alpha}(x), \hat{\beta}(x))$ be the minimizer of the following function of $\alpha \in \mathbb{R},\ \beta \in \mathbb{R}^d$:

$$\sum_{X_l \in \mathcal{N}^{true}_{h_{\mathrm{mGCV},\hat{r}}}} \left(\log(\hat{r}_l + 1/n) - \alpha - \beta^T B_x^T (X_l - x)\right)^2 K_{h_{\mathrm{mGCV},\hat{r}}}(X_l, x),$$

  where $h_{\mathrm{mGCV},\hat{r}}$ is the bandwidth determined by minimizing the mGCV upon the data set $\{(X_l, \log(\hat{r}_l + 1/n))\}_{l=1}^n$.
- The estimated σ^2 at x, $\hat{\sigma}^2(x)$, is then defined by

$$\hat{\sigma}^2(x) := e^{\hat{\alpha}(x)} \left[\frac{1}{n}\sum_{l=1}^{n} \hat{r}_l e^{-\hat{\alpha}(X_l)}\right]^{-1}.$$

III. Bandwidth for $\hat{m}(x, h)$:

- Estimate the conditional bias and the conditional variance of $\hat{m}(x, h)$ respectively by

$$\hat{b}(x, h) = \frac{\hat{m}(x, h) - \hat{m}(x, h/2)}{1/2}$$

$$\hat{v}(x, h) = v_1^T(\mathbb{X}_x^T \mathbb{W}_x \mathbb{X}_x)^{-1} \mathbb{X}_x^T \mathbb{W}_x \hat{S}_x \mathbb{W}_x \mathbb{X}_x (\mathbb{X}_x^T \mathbb{W}_x \mathbb{X}_x)^{-1} v_1,$$

  where $\hat{S}_x = \text{diag}\{\hat{\sigma}^2(X_1), \ldots, \hat{\sigma}^2(X_n)\}$.

- The conditional MSE of $\hat{m}(x, h)$ is estimated by

$$\widehat{\text{MSE}}(x, h) := \hat{b}(x, h)^2 + \hat{v}(x, h).$$

- The value of h which minimizes $\widehat{\text{MSE}}(x, h)$, denoted as $\hat{h}_{\text{opt}}(x)$, is selected to approximate the optimal bandwidth.

Isomap face data

Isomap face data (Tenenbaum, 2000):

- There are 698 64×64 images, denoted as $\{I_l^{64}\}_{l=1}^{698}$, labeled with three variable: the horizontal orientation, the vertical orientation, and the illumination direction.
- The dataset was sampled from a 3-dimensional manifold embedded in $\mathbb{R}^{64\times 64}$, which is parametrized by the above three variables.
- Denote the resized images of size $k \times k$ as $\{I_l^k\}_{l=1}^{698}$, where $k \in [1, 64] \cap \mathbb{Z}$.

We performed 200 replications of the following experiment, which was suggested by Aswani, Bickel, and Tomlin (2011).

- Fix $k = 7$. We randomly split $\{I_l^7\}_{l=1}^{698}$ into a training set consisting of 688 images and a testing set consisting of 10 images.
- The horizontal orientation of the images in the testing set was then estimated from the training set.

	RASE	computational time
MALLER	1.320 ± 0.992	13.429 ± 4.920
NEDE	1.785 ± 1.212	34.461 ± 4.585
NALEDE	1.776 ± 1.200	170.709 ± 28.819
NEDEP	1.869 ± 1.241	53.721 ± 8.359
NALEDEP	2.810 ± 3.653	187.375 ± 31.262

Table : The averages and standard deviations, over 200 replications, of RASE and computational time in seconds for different estimators tested on the resized Isomap face data $\{I_l^7\}_{l=1}^{698}$.

- Next, we carried out another 200 replications of the same experiment but with $k = 14, 21$, or 28.
- When $k = 14, 21$ or 28, it takes a long time to compute the methods by Aswani, Bickel, and Tomlin (2011).

	$k = 14$	$k = 21$	$k = 28$
RASE	1.048 ± 0.645	1.185 ± 1.583	1.014 ± 0.697
computational time	17.229 ± 5.826	18.782 ± 5.636	33.439 ± 16.601

Table : The averages and standard deviations over 200 replications of RMSE and computational time (in seconds) for our estimator using the resized data $\{I_l^{14}\}_{l=1}^{698}$, $\{I_l^{21}\}_{l=1}^{698}$, or $\{I_l^{28}\}_{l=1}^{698}$.

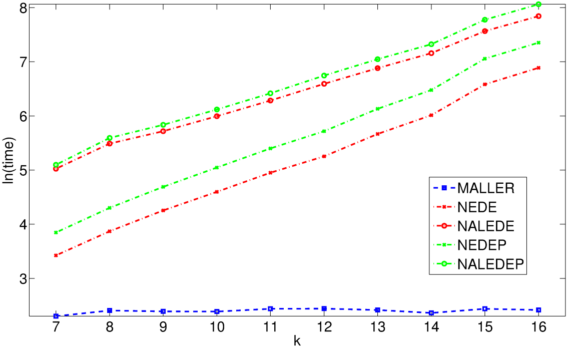


Figure : The running time for MALLER, NEDE, NALEDE, NEDEP and NALEDEP when $k = 7, 8, \ldots, 16$. The y-axis is in the natural log scale.

Registration Problem with Computed Tomography data

A. Frank and A. Asuncion. UCI machine learning repository, 2010.

- There are 53500 2D CT images from 97 volumes scanned from 71 different patients.
- There are s_i slices in the i-th volume. So, $\sum_{i=1}^{97} s_i = 53500$.
- The age of the patients ranges from 4 to 86 years old.
- The collection covers the complete area between the top of the head to the end of the coccyx. Each patient contributed no more than 1 thorax and 1 neck scan.
- Then 53500 feature vectors in $\mathbb{R}^{384}$ are determined based on the radial image descriptor (Graf et al., 2011).

Two-level nearest neighbors search (Graf et al., 2011):

- PCA is applied to the 384-dim feature vectors to project them onto the first 50 dominant principal components.
- Let x be the PCA vector corresponding to the test image.
- First, find the $k_1 \in \mathbb{N}$ nearest neighbors of x in each volumn and get $N \times k_1$ vectors, denoted as S.
- Then, find the $k_2 \in \mathbb{N}$ nearest neighbors of x in S and their associated ground truth, denoted as y_l, $l = 1 \ldots, k_2$.
- The estimate of the true location of the test image is given by $\frac{1}{k_2} \sum_{l=1}^{k_2} y_l$.
- We call this the $\mathrm{NN}(k_1, k_2)$ algorithm.

Application of MALLER to CT data

- We followed the same PCA dimension reduction, two-level estimation, and leave-one-volume-out schemes.
- Following Graf et al. (2011), we set the dimension of the PCA vectors as 50.
- It may occur that some of the images in S actually come from different anatomical sections from the location of the test image, so we included the corresponding location information in step 2 of MALLER.
- We took $k_1 = 6$ to build up S in order to speed up the computation for clinical purpose, and to ensure that the number of points is not too small.

	estimation error (cm)	Q90	F(1)	computational time (sec)
MALLER	1.726 ± 3.26	3.55	47.42%	3.1 ± 0.52
NN(1, 3)	1.84 ± 3.06	3.8	45.56%	$3 \times 10^{-3} \pm 0.19 \times 10^{-3}$
NN(6, 3)	1.95 ± 3.39	4.03	42.81%	$4.2 \times 10^{-3} \pm 0.15 \times 10^{-3}$
NEDE	3.386 ± 4.247	8.06	29.77%	5.93 ± 0.86
NALEDE	3.275 ± 4.113	7.73	30.16%	11.31 ± 2
NEDEP	3.388 ± 4.258	8.06	29.77%	9.29 ± 1.35
NALEDEP	3.276 ± 4.113	7.73	30.15%	14.66 ± 2.26

Table : CT Data. $F(1)$: the proportion of the estimation errors being less than 1cm; Q90: the 90% quantile of the estimation errors.

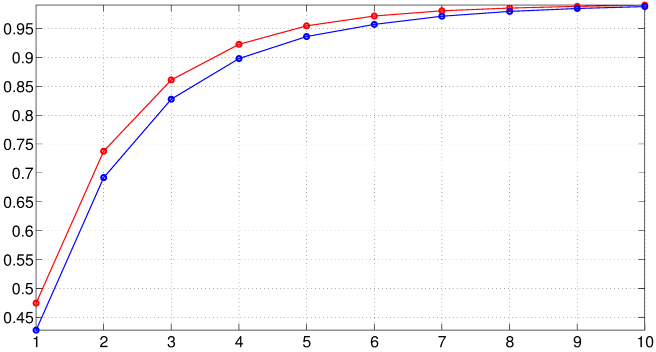


Figure : The cumulative proportion of the estimation errors of MALLER (red) and NN(1, 3) (blue). The unit in the x-axis is cm.

Application to Manifold Learning

Diffusion map

For a fixed bandwidth $h > 0$, define $n \times n$ matrix W and $n \times n$ diagnal matrix D by

$$W(i,j) = K\left(\frac{\|X_i - X_j\|_{\mathbb{R}^p}}{\sqrt{h}}\right) \quad \text{and} \quad D(i,i) = \sum_{j=1}^{n} W(i,j).$$

Then $A := D^{-1}W$ is a Markov transition matrix of a random walk over the sample points $\{X_i\}_{i=1}^n$.
Given the regression model (1), define the Nadaraya-Watson type estimator $\hat{m}_{NW}$ of m at X_i as

$$\hat{m}_{NW}(X_i, h) := (AY)(i) = \frac{\sum_{j=1}^n K\left(\frac{\|X_i - X_j\|_{\mathbb{R}^p}}{\sqrt{h}}\right) Y_j}{\sum_{j=1}^n K\left(\frac{\|X_i - X_j\|_{\mathbb{R}^p}}{\sqrt{h}}\right)}, \; i = 1, \ldots, n,$$

so A is the smoothing matrix of $\hat{m}_{NW}(\cdot, h)$.

When $m \in C^3(\mathrm{M})$ and $X_i \notin \mathrm{M}_{\sqrt{h}}$, as $n \to \infty$,

$$\begin{aligned}(Am)(i) &= m(X_i) + h\frac{\mu_{1,2}}{2d}\left(\Delta m(X_i) + 2\frac{m(X_i)\Delta f(X_i)}{f(X_i)}\right) \\ &\quad + O(h^2) + O_p\left(\frac{1}{n^{1/2}h^{d/4-1/2}}\right),\end{aligned}$$

where $m = (m(X_1), \ldots, m(X_n))^T$.
Define

$$W_1 = D^{-1}WD^{-1}, \quad D_1(i,i) = \sum_{j=1}^{n} W_1(i,j), \quad L_1 = h^{-1}\left(D_1^{-1}W_1 - I_n\right).$$

When $n \to \infty$, it is shown by Coifman and Lafon (2006) that

$$(L_1 m)(i) = \frac{\mu_{1,2}}{2d}\Delta m(X_i) + O(h) + O_p\left(\frac{1}{n^{1/2}h^{d/4+1/2}}\right).$$

- $\int_{\mathrm{M}} \|\nabla m\|^2 = -\int_{\mathrm{M}} (\Delta m) m$ for twice differentiable $m : \mathrm{M} \to \mathbb{R}$.
- The minimizer of $\int_{\mathrm{M}} \|\nabla m\|^2$ subject to $\|m\| = 1$ is given by the eigenfunctions of the Laplace-Beltrami operator Δ.
- The diffusion map is $\psi_t : V \to \mathbb{R}^d$ such that $\psi_t(v) = (\lambda_1^t \psi_1 v, \ldots, \lambda_d^t \psi_d v) \in \mathbb{R}^d$, where $\psi_1, \ldots, \psi_d$ are the first d eigenvectors of L_1 and $\lambda_1 \ldots, \lambda_d$ are the corresponding eigenvalues.

Suppose M is compact, smooth and ∂M is non-empty and smooth. When $X_i \in \mathrm{M}_{\sqrt{h}}$,

$$(D_1^{-1} W_1 m)(i) = m(X_0) + \sqrt{h} C_1 \partial_\nu m(X_0) + O(h) + O_p\Big(\frac{1}{n^{1/2} h^{d/4-1/2}}\Big),$$

where $C_1 = O(1)$, $X_0 \in \partial\mathrm{M}$ is the point on ∂M closest to X_i, and ν is the normal direction at X_0.
If the $\sqrt{h}$-order term is non-zero, the estimator $(L_1 m)(i)$ blows up when $h \to 0$. Thus, the Neuman's boundary condition $\frac{\partial m}{\partial \nu} = 0$ is necessary for L_1:

$$\begin{cases} \Delta m(x) = \lambda m(x) & \text{when } x \in \mathrm{M} \\ \frac{\partial m}{\partial \nu}(x) = 0 & \text{when } x \in \partial\mathrm{M} \end{cases}$$

Our method

For given $h > 0$, consider the proposed MALLER and define

$$A_p = \begin{bmatrix} v_1^T(\mathbb{X}_{X_1}^T \mathbb{W}_{X_1} \mathbb{X}_{X_1})^{-1} \mathbb{X}_{X_1}^T \mathbb{W}_{X_1} \\ \vdots \\ v_1^T(\mathbb{X}_{X_n}^T \mathbb{W}_{X_n} \mathbb{X}_{X_n})^{-1} \mathbb{X}_{X_n}^T \mathbb{W}_{X_n} \end{bmatrix},$$

$$L_p = h^{-1}\big(A_p - I_n\big).$$

Then, for any $m \in C^3(\mathrm{M})$ and $X_i \notin \mathrm{M}_{\sqrt{h}}$, from Theorem 1 we have directly

$$(L_p m)(i) = \frac{\mu_{1,2}}{2d} \Delta m(X_i) + O(h^{1/2}) + O_p\Big(\frac{1}{n^{1/2} h^{d/4}}\Big).$$

Thus the matrix L_p can be used to construct an estimator of the Laplace-Beltrami operator Δ.

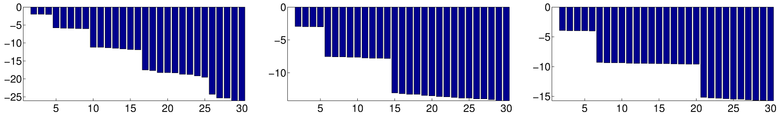
Suppose M is compact, smooth, and its boundary ∂M is nonempty and smooth.
For $X_i \in \mathrm{M}_{\sqrt{h}}$, Corollary 1 leads to

$$
\begin{aligned}
(L_p m)(i) &= \frac{1}{2}\sum_{k=1}^{d} c_k(X_i)\nabla^2_{\partial_k,\partial_k} m(X_i) + O_p(h^{-1/2}h_{\mathrm{PCA}}^{3/4} + h_{\mathrm{PCA}}^{1/2}) \\
&\quad + O_p\Big(\frac{1}{n^{1/2}h^{d/4}}\Big).
\end{aligned}
$$

Thus, we know that when X_i is near the boundary, the estimator L_p does not blow up when $h \to 0$, and a different boundary condition can be imposed.

Example: spheres

We sampled 1000 points uniformly from the 2-dim sphere S^2 embedded in $\mathbb{R}^3$, 2000 points uniformly from the 3-dim sphere S^3 embedded in $\mathbb{R}^4$, and 4000 points uniformly from the 4-dim sphere S^4 embedded in $\mathbb{R}^4$, and built the matrix L_p with $h = 0.09$.



(a) S^2 (b) S^3 (c) S^4

Figure : Bar plots of the first 30 eigenvalues of L_p. The first eigenvalue of Δ is zero for S^2, S^3 and S^4, and the multiplicities of the first few eigenvalues of Δ of S^k are $1, 3, 5, 7 \ldots$ when $k = 2$, are $1, 4, 9, 16 \ldots$ when $k = 3$, and are $1, 5, 14, 30 \ldots$ when $k = 4$.

Example: half circle

We sampled 2000 points $\{(\cos(\theta_l), \sin(\theta_l))\}_{l=1}^{2000}$ from the half circle embedded in $\mathbb{R}^2$, where θ_l were uniformly sampled from $[0, \pi]$, and evaluated the eigenvectors of L_p built on $\{(\cos(\theta_l), \sin(\theta_l))\}_{l=1}^{2000}$.

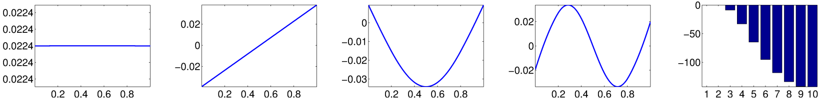


Figure : The first four eigenvectors of L_p and the first 10 eigenvalues of L_p. The first two eigenvalues are zero. Notice that the second, third and fourth eigenvectors can not happen if the Laplace-Beltrami operator satisfies the Neuman's condition.

Example: Swiss roll

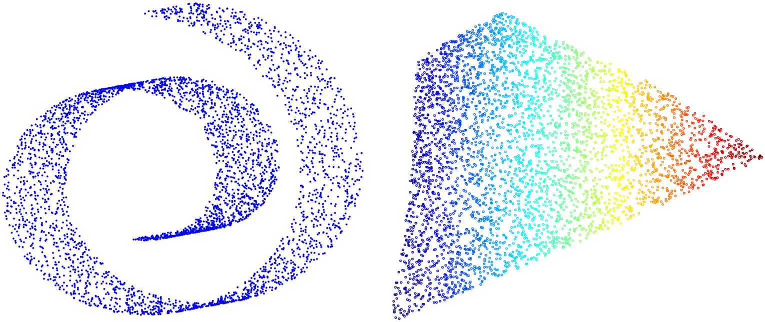


Figure : Visualization of Swiss roll data. Left panel: data $X_1, \ldots, X_n$. Right panel: $X_i \to (\lambda_1^t \phi_1(i), \lambda_2^t \phi_2(i))$, where $L_p \phi_j = \lambda_j \phi_j$, $j = 1, 2$.