Local linear regression on manifolds and its geometric implications

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).

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 $\iota : M \hookrightarrow \mathbb{R}^p$.
- Consider the following regression model

\[
Y = m(X) + \sigma(X) \epsilon,
\]

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


Local Linear Regression on Unknown Manifolds (MALLER)

Let $\{(X_i, Y_i)\}_{i=1}^n$ denote a random sample observed from the regression model (1) with $\mathcal{X} = \{X_i\}_{i=1}^n$ 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

- $\mathcal{X} = \{X_1, \ldots, X_n\}$

- $\mathcal{N}_{x,\delta}^{\mathbb{R}^p} = \{X_j \in \mathcal{X} : \|X_j - x\|_{\mathbb{R}^p} < \sqrt{\delta}\}$: the set of Euclidean $\sqrt{\delta}$-neighbors of $x$

- $d(\cdot, \cdot)$: the geodesic distance

- $\mathcal{N}_{x,\delta}^{\mathcal{M}} = \{X_j \in \mathcal{X} : d(X_j, x) < \sqrt{\delta}\}$: 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 $\mathcal{N}_{x,\delta}^{\mathbb{R}^p} \cup \{x\}$, and use the set

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

  as an estimate of $\mathcal{N}_{x,\delta}^{\mathcal{M}}$. 
Figure: $\tau$: reach, $1/\tau$: the condition number of $M$. The set of Euclidean $\sqrt{\delta}$-neighbors of $x$, $N_{x,\delta}^{\mathbb{R}^p}$, consists of both the red and green crosses. The set of geodesic $\sqrt{\delta}$-neighbors of $x$, $N_{x,\delta}^M$, consists of the red crosses but not the green crosses.
Step 3: embedded tangent plane estimation

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

Local PCA:
- \(\Sigma_x\): the sample covariance matrix of \(\mathcal{N}_{x, h_{PCA}}^{\text{true}}\).
- \(\{U_k(x)\}_{k=1}^d\): the first \(d\) eigenvectors of \(\Sigma_x\).
- Let \(B_x\) be the \(p \times d\) matrix \(B_x = \begin{bmatrix} U_1(x) & \ldots & U_d(x) \end{bmatrix}\).

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 \(\iota_* T_xM\).
Step 4: local linear regression on tangent plane estimate

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

- $h > 0$: a bandwidth

- Let

$$\hat{\beta}_x = \arg\min_{\beta \in \mathbb{R}^{d+1}} \sum_{l=1}^{n} \left( Y_l - \beta_0 - \sum_{k=1}^{d} \beta_k x_{l,k} \right)^2 I_{\mathbb{N}^{\text{true}}_{x,h}}(X_l) K_h(X_l, x),$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_d)^T$, $K_h(X_l, x) := \frac{1}{h^{d/2}} K \left( \frac{\|X_l-x\|_P}{\sqrt{h}} \right)$ and $I$ is the indicator function.
Let $\bar{X}_x = \begin{bmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{bmatrix}^T$, and

$\bar{W}_x = \text{diag}(K_h(X_1, x)I_{\mathcal{N}^\text{true}}(X_1), \ldots, K_h(X_n, x)I_{\mathcal{N}^\text{true}}(X_n))$.

The functional $\hat{\beta}_x$ can be written as

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

where $Y = (Y_1, \ldots, Y_n)^T$, if $(X_x^T \bar{W}_x X_x)^{-1}$ exists.

The estimator of $m(x)$ we propose is given by

$$
\hat{m}(x, h) := v_1^T \hat{\beta}_x = v_1^T (X_x^T \bar{W}_x X_x)^{-1} X_x^T \bar{W}_x Y,
$$

(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:

$$\hat{\iota} \ast \text{grad} m(x) := \sum_{i=1}^{d} \hat{\nabla}_{\partial_i(x)} m(x, h) U_i(x).$$  \quad (3)$$

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

$$\hat{\nabla}_{\partial_i(x)} m(x, h) := v_{i+1}^T \hat{\beta}_x,$$  \quad (4)$$

and $\{\partial_i(x)\}_{i=1}^d$ is the orthonormal basis of $T_xM$ 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 $u, v \in T_x M$, 
  \[ g_x(u, v) := \langle \iota_* u, \iota_* v \rangle. \]

- The exponential map at $x \in M$ is denoted as $\exp_x$.

- The volume form on $M$ induced from $g$ is denoted as $dV$.

- Define the set of points close to the boundary $\partial M$ with distance less than $\delta \geq 0$, where $\delta$ is small enough, as 
  \[ M_\delta(x) = \{ y \in M : \min_{y \in \partial M} d(x, y) \leq \delta \}, \]
  where $d(x, y)$ is the geodesic distance between $x$ and $y$.

- Denote by $\nabla$ the Levi-Civita connection, $\Delta$ the Laplace-Beltrami operator, and Hess the second order covariant derivative operator on $(M, g)$.
Probability density function of the random vector $X : \Omega \to \iota(M)$:

- $X$: a measurable function with respect to the probability space $(\Omega, \mathcal{F}, P)$
- $\tilde{\mathcal{B}}$: the Borel sigma algebra of $\iota(M)$.
- $\tilde{P}_X$: 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 $dV$ so that $d\tilde{P}_X(x) = f(\iota^{-1}(x))\iota_*dV(x)$ for some $f \in C^2(M)$. That is, for an integrable function $\zeta : \iota(M) \to \mathbb{R}$,

$$
E\zeta(X) = \int_{\Omega} \zeta(X(\omega))dP(\omega) = \int_{\iota(M)} \zeta(x)d\tilde{P}_X(x)
= \int_{\iota(M)} \zeta(x)f(\iota^{-1}(x))\iota_*dV(x) = \int_{M} \zeta(\iota(y))f(y)dV(y).
$$

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 $C^2(M)$ and satisfies

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

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

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

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

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

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

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

For $x \in M_{\sqrt{h}}$ and $h > 0$, define

$$
\nu_{i,x} := \begin{bmatrix}
\nu_{i,x,11} & \nu_{i,x,12} \\
\nu_{i,x,12}^T & \nu_{i,x,22}
\end{bmatrix}
$$

$$
:= \begin{bmatrix}
\int \frac{1}{\sqrt{h}} D(x) K^i(\|u\|) du & \int \frac{1}{\sqrt{h}} D(x) K^i(\|u\|) u^T du \\
\int \frac{1}{\sqrt{h}} D(x) K^i(\|u\|) u du & \int \frac{1}{\sqrt{h}} D(x) K^i(\|u\|) uu^T du
\end{bmatrix},
$$

$$
D(x) := \exp^{-1}_x (B^M_{\sqrt{h}}(x) \cap M) \subset T_x M,
$$

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

Here, $I_k$ denotes the $k \times k$ identity matrix for any $k \in \mathbb{N}$. 
**Theorem 2.** Suppose \( x \in M \sqrt{h} \), \( h_{\text{PCA}} \asymp n^{-2/(d+1)} \) and \( h \geq h_{\text{PCA}} \).

The conditional MSE of the estimator \( \hat{m}(x, h) \) is

\[
\text{MSE}\{\hat{m}(x, h) | \mathcal{X}\} = \frac{h^2}{4} \left[ \text{tr} \left( \text{Hess} m(x) \nu_{1,x,22} \right) \right]^2 + \frac{\nu_1^T \nu_{1,x}^{-1} \nu_{2,x} \nu_{1,x}^{-1} \nu_1 \sigma^2(x)}{nh^{d/2} f(x)}
\]

\[+ O_p \left( h^{3/2} h_{\text{PCA}}^{3/4} + h^{5/2} \right) + O_p \left( \frac{1}{n^{1/2} h^{d/4-2}} + \frac{1}{nh^{d/2-1/2}} + \frac{1}{n^{3/2} h^{3d/4}} \right).\]

**Corollary 1.** Suppose \( \partial M \) is smooth, \( x \in M \sqrt{h} \), \( h_{\text{PCA}} \asymp n^{-2/(d+1)} \) 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 \( m \):

\[
E\{\hat{m}(x, h) - m(x) | \mathcal{X}\} = \frac{h}{2} \sum_{k=1}^{d} c_k(x) \nabla_{\partial_k, \partial_k}^2 m(x)
\]

\[+ O_p \left( h^{1/2} h_{\text{PCA}}^{3/4} + h^{3/2} \right) + O_p \left( \frac{1}{n^{1/2} h^{d/4-1}} \right),\]

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

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

$$= h^2 \left[ \frac{\mu_{1,2}}{d} \nabla_{\partial_i} f(x) \Delta m(x) - \frac{\mu_{1,2}}{d} \int_{S^{d-1}} \theta^T \text{Hess} m(x) \theta \nabla_{\theta} f(x) d\theta \right] + \frac{1}{nh^{d/2}+1} \frac{d\mu_{2,2} \sigma^2(x) f(x)}{\mu_{2,2}} + O_p(h_5^5 + h^3 h_{PCA}^3) + O_p\left(\frac{1}{n^{1/2} h^{d-3/2}} + \frac{1}{n h^{d/2}} + \frac{1}{n^{3/2} h^{3d/4}+1}\right),$$

where $\{\partial_i(x)\}_{i=1}^d$ is an orthonormal basis of $T_xM$. 
Theorem 4. Suppose \( x \in M_{\sqrt{h}} \), \( h_{\text{PCA}} \asymp n^{-2/(d+1)} \) and \( h \geq h_{\text{PCA}} \). The conditional MSE for the estimator \( \widehat{\nabla_{\partial_i(x)} m(x, h)} \) given in (4) is

\[
\text{MSE}\{\widehat{\nabla_{\partial_i(x)} m(x, h)}|X\} = h \left( \frac{v_{i+1}^T \nu_{1,x}^{-1}}{2} \int_{\frac{1}{\sqrt{h}}D(x)} K(\|u\|) u^T \text{Hess}m(x) u \left[ \begin{array}{c} 1 \\ u \end{array} \right] du \right)^2 \\
+ \frac{v_{i+1}^T \nu_{1,x}^{-1} \nu_{2,x} \nu_{1,x}^{-1} \nu_{i+1} \sigma^2(x)}{nh^{d+1}_2} + O_p \left( h^2 h_{\text{PCA}}^3 + hh_{\text{PCA}}^2 \right) \\
+ O_p \left( \frac{1}{n^2 h^{d-3/2}_4} + \frac{1}{nh^{d+1/2}_2} + \frac{1}{n^{3/2} h^{3d/4}} \right),
\]

where \( \{\partial_i(x)\}_{i=1}^d \) is an orthonormal basis of \( T_x M \).
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 $\{(X_j, Y_j)\}_{j \in J}$.

- The mGCV bandwidth, denoted as $h_{mGCV, \hat{m}}$, is chosen to be the value of $h$ in $\mathcal{H}_{mGCV}$ which minimizes

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

where $\text{atr}_J(h) := \frac{1}{|J|} \sum_{j \in J} \nu_1^T (X_j^T W_j X_j)^{-1} \nu_1 h^{-d/2} K(0)$. 
II. Estimate the value of the conditional variance $\sigma^2$ at $x$:

- Define the residuals as $\hat{r}_l := (Y_l - \hat{m}(X_l, h_{mGCV, \hat{m}}))^2$, $l = 1, \ldots, n$.

- Let $(\hat{\alpha}(x), \hat{\beta}(x))$ be the minimizer of the following function of $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^d$:

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

where $h_{mGCV, \hat{r}}$ 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 $\sigma^2$ 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

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

$$\hat{v}(x, h) = \nu_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x \hat{S}_x W_x X_x (X_x^T W_x X_x)^{-1} \nu_1,$$

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 (Tenenbaum, 2000):

- There are 698 $64 \times 64$ images, denoted as $\{I_i^{64}\}_{i=1}^{698}$, labeled with three variables: 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 $\{I_i^k\}_{i=1}^{698}$, 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 $\{I_i^7\}_{i=1}^{698}$ 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.

<table>
<thead>
<tr>
<th></th>
<th>RASE</th>
<th>computational time</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALLER</td>
<td>1.320 ± 0.992</td>
<td>13.429 ± 4.920</td>
</tr>
<tr>
<td>NEDE</td>
<td>1.785 ± 1.212</td>
<td>34.461 ± 4.585</td>
</tr>
<tr>
<td>NALEDE</td>
<td>1.776 ± 1.200</td>
<td>170.709 ± 28.819</td>
</tr>
<tr>
<td>NEDEP</td>
<td>1.869 ± 1.241</td>
<td>53.721 ± 8.359</td>
</tr>
<tr>
<td>NALEDEP</td>
<td>2.810 ± 3.653</td>
<td>187.375 ± 31.262</td>
</tr>
</tbody>
</table>

**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 $\{I_i^7\}_{i=1}^{698}$. 
Next, we carried out another 200 replications of the same experiment but with $k = 14, 21, \text{ or } 28$.

When $k = 14, 21$ or 28, it takes a long time to compute the methods by Aswani, Bickel, and Tomlin (2011).

<table>
<thead>
<tr>
<th>RASE</th>
<th>$k = 14$</th>
<th>$k = 21$</th>
<th>$k = 28$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$1.048 \pm 0.645$</td>
<td>$1.185 \pm 1.583$</td>
<td>$1.014 \pm 0.697$</td>
</tr>
<tr>
<td>computational time</td>
<td>$17.229 \pm 5.826$</td>
<td>$18.782 \pm 5.636$</td>
<td>$33.439 \pm 16.601$</td>
</tr>
</tbody>
</table>

**Table**: The averages and standard deviations over 200 replications of RMSE and computational time (in seconds) for our estimator using the resized data $\{l_i^{14}\}_{i=1}^{698}$, $\{l_i^{21}\}_{i=1}^{698}$, or $\{l_i^{28}\}_{i=1}^{698}$. 
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.
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 volume 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 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.
<table>
<thead>
<tr>
<th>Method</th>
<th>Estimation Error (cm)</th>
<th>Q90</th>
<th>F(1)</th>
<th>Computational Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALLER</td>
<td>$1.726 \pm 3.26$</td>
<td>3.55</td>
<td>47.42%</td>
<td>$3.1 \pm 0.52$</td>
</tr>
<tr>
<td>NN(1, 3)</td>
<td>$1.84 \pm 3.06$</td>
<td>3.8</td>
<td>45.56%</td>
<td>$3 \times 10^{-3} \pm 0.19 \times 10^{-3}$</td>
</tr>
<tr>
<td>NN(6, 3)</td>
<td>$1.95 \pm 3.39$</td>
<td>4.03</td>
<td>42.81%</td>
<td>$4.2 \times 10^{-3} \pm 0.15 \times 10^{-3}$</td>
</tr>
<tr>
<td>NEDE</td>
<td>$3.386 \pm 4.247$</td>
<td>8.06</td>
<td>29.77%</td>
<td>$5.93 \pm 0.86$</td>
</tr>
<tr>
<td>NALEDE</td>
<td>$3.275 \pm 4.113$</td>
<td>7.73</td>
<td>30.16%</td>
<td>$11.31 \pm 2$</td>
</tr>
<tr>
<td>NEDEP</td>
<td>$3.388 \pm 4.258$</td>
<td>8.06</td>
<td>29.77%</td>
<td>$9.29 \pm 1.35$</td>
</tr>
<tr>
<td>NALEDEP</td>
<td>$3.276 \pm 4.113$</td>
<td>7.73</td>
<td>30.15%</td>
<td>$14.66 \pm 2.26$</td>
</tr>
</tbody>
</table>

**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 $n \times n$ diagonal 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

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

so $A$ is the smoothing matrix of $\hat{m}_{NW}(\cdot, h)$. 
When \( m \in C^3(M) \) and \( X_i \notin M_\sqrt{h} \), as \( n \to \infty \),

\[
(Am)(i) = m(X_i) + h^{\mu_{1,2}} \left( \Delta m(X_i) + 2 \frac{m(X_i) \Delta f(X_i)}{f(X_i)} \right) \\
+ O(h^2) + O_p \left( \frac{1}{n^{1/2} h^{d/4-1/2}} \right),
\]

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

Define

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

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

\[
(L_1 m)(i) = \frac{\mu_{1,2}}{2d} \Delta m(X_i) + O(h) + O_p \left( \frac{1}{n^{1/2} h^{d/4+1/2}} \right).
\]
\[ \int_M \|\nabla m\|^2 = -\int_M (\Delta m)m \text{ for twice differentiable } m : M \to \mathbb{R}. \]

The minimizer of \( \int_M \|\nabla m\|^2 \) subject to \( \|m\| = 1 \) is given by the eigenfunctions of the Laplace-Beltrami operator \( \Delta \).

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 $\partial M$ is non-empty and smooth. When $X_i \in M, \sqrt{h},$

$$(D^{-1}_1 W_1 m)(i) = m(X_0) + \sqrt{h} C_1 \partial \nu m(X_0) + O(h) + O_p \left( \frac{1}{n^{1/2} h^{d/4-1/2}} \right),$$

where $C_1 = O(1)$, $X_0 \in \partial M$ is the point on $\partial M$ closest to $X_i$, and $\nu$ 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 $L_1$:

$$\begin{cases} 
\Delta m(x) = \lambda m(x) & \text{when } x \in M \\
\frac{\partial m}{\partial \nu}(x) = 0 & \text{when } x \in \partial M 
\end{cases}$$
Our method
For given $h > 0$, consider the proposed MALLER and define

$$A_p = \begin{bmatrix}
\nu_1^T (X_1 \mathbb{X} X_1 \mathbb{W} X_1)^{-1} X_1^T \mathbb{W} X_1 \\
\vdots \\
\nu_1^T (X_n \mathbb{X} X_n \mathbb{W} X_n)^{-1} X_n^T \mathbb{W} X_n
\end{bmatrix},$$

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

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

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

Thus the matrix $L_p$ can be used to construct an estimator of the Laplace-Beltrami operator $\Delta$. 
Suppose $M$ is compact, smooth, and its boundary $\partial M$ is nonempty and smooth. For $X_i \in M_{\sqrt{h}}$, Corollary 1 leads to

$$(L_p m)(i) = \frac{1}{2} \sum_{k=1}^{d} c_k(X_i) \nabla_{\partial_k}^2 m(X_i) + O_p(h^{-1/2} h_{PCA}^{3/4} + h_{PCA}^{1/2}) + O_p\left(\frac{1}{n^{1/2} h^{d/4}}\right).$$

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 $h = 0.09$.

Figure: Bar plots of the first 30 eigenvalues of $L_p$. The first eigenvalue of $\Delta$ is zero for $S^2$, $S^3$ and $S^4$, and the multiplicities of the first few eigenvalues of $\Delta$ 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 \( \theta_l \) 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: $X_i \rightarrow (\lambda_1^t \phi_1(i), \lambda_2^t \phi_2(i))$, where $L_p \phi_j = \lambda_j \phi_j$, $j = 1, 2$. 