Signal Ensemble Classification on Manifolds

Linh Lieu, *Naoki Saito*, and *Julien van Hout*

Department of Mathematics
University of California, Davis

The 2nd Reunion Conference on
Mathematics of Knowledge and Search Engines
Lake Arrowhaed
June 11, 2010

Outline

Outline

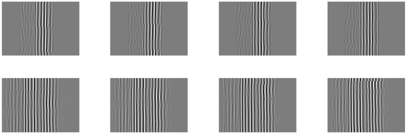
Acknowledgment

- Raphy Coifman (Yale)
- Quyen Hyunh (NSWC)
- Yosi Keller (Bar-Ilan Univ., Israel)
- Stéphane Lafon (Google)
- Bradley Marchand (UC Davis $\Longrightarrow$ NSWC, Panama City, FL)
- Hrushikesh Mhaskar (Cal State LA)
- NSF
- ONR

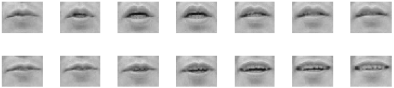
Outline

Signal Ensemble Classification Problems

- We want to classify ensembles of signals, not individual signals.
- Examples include: Underwater object classification using sonar waveforms; Classification of video clips, ...



(a) Sonar Waveforms



(b) Video Clips of Digit Speaking Lips

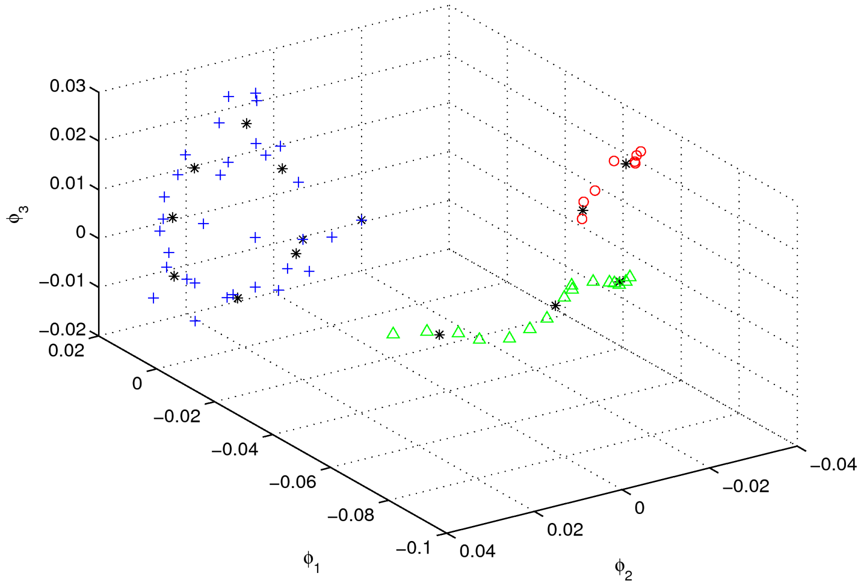
- Let $X := \bigcup_{i=1}^{M} X^i \subset \mathbb{R}^d$ be a collection of M training ensembles. Each X^i consists of m_i individual signals, i.e., $X^i := \{\mathbf{x}_1^i, \ldots, \mathbf{x}_{m_i}^i\}$, and has a unique label among C possible labels. Let $m_\star := \sum_{i=1}^{M} m_i$. Let $Y := \bigcup_{j=1}^{N} Y^j \subset \mathbb{R}^d$ be a collection of test (i.e., unlabeled) ensembles where $Y^j := \{\mathbf{y}_1^j, \cdots, \mathbf{y}_{n_j}^j\}$. Our goal is to classify each Y^j to one of the possible C classes given the training ensembles X. This task is different from classifying each signal $\mathbf{y}_k^j \in Y$ individually.

Outline

Our Proposed Algorithm

- Training Stage (X is given)
  1. Preset a large enough initial dimension $1 \leq s_0 \ll \min(d, m_\star)$.
  2. Construct a low-dimensional embedding map $\Psi : \mathbb{R}^d \to \mathbb{R}^{s_0}$.
  3. For $i = 1 : M$, construct a signature P^i using $\Psi(X^i)$.
  4. Determine the appropriate dimension $1 \leq s \leq s_0$ and re-adjust each signature P^i in Step 1.3.
- Test Stage (Now Y is fed)
  1. Extend the learned map Ψ to the test ensembles Y to embed them in $\mathbb{R}^s$.
  2. Construct a signature Q^j for each Y^j, $j = 1 : N$.
  3. For $j = 1 : N$, measure the distance $d(P^i, Q^j)$, and find $i_j := \arg\min_{1 \leq i \leq M} d(P^i, Q^j)$. Assign the label of X^{i_j} to Y^j . In other words, apply 1-nearest neighbor classifier with the base distance $d(\cdot, \cdot)$ in the reduced embedding space $\mathbb{R}^s$.

Signatures in the Reduced Embedding Space



Outline

Dimensionality Reduction/Low-Dimensional Embedding

- Many techniques, proposals, algorithms exist.
- In this talk, we only deal with
  - Classical Multidimensional Scaling (CMDS) $\equiv$ PCA
  - Laplacian Eigenmap
  - Diffusion Map
- CMDS/PCA is a linear technique whereas LE/DM are nonlinear.

Notation

- Let X be the training data matrix, $X := (\mathbf{x}_1, \dots, \mathbf{x}_{m_\star}) \in \mathbb{R}^{d \times m_\star}$.
- Let $\widetilde{X} := X(I - \mathbf{1}\mathbf{1}^T/m_\star)$, i.e., the centered data matrix (the mean of the column vectors $\bar{\mathbf{x}}$ is subtracted from each column vector).
- Let $\Psi : \mathbb{R}^d \to \mathbb{R}^s$ be a low-dimensional embedding map.
- Let $\Psi(X) = (\Psi(\mathbf{x}_1), \dots, \Psi(\mathbf{x}_{m_\star})) \in \mathbb{R}^{s \times m_\star}$

Classical (Multidimensional) Scaling and PCA

- Define the similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ by the centered correlation

$$\alpha(\mathbf{x}_i, \mathbf{x}_j) := (\mathbf{x}_i - \bar{\mathbf{x}})^T (\mathbf{x}_j - \bar{\mathbf{x}}).$$

- Then, the classical scaling seeks the low-dimensional representation that preserves the pairwise similarities in X as well as possible by minimizing

$$J_{\mathrm{CS}}(\Psi) := \sum_{i,j} (\alpha(\mathbf{x}_i, \mathbf{x}_j) - \alpha(\Psi(\mathbf{x}_i), \Psi(\mathbf{x}_j)))^2 = \left\| \widetilde{X}^T \widetilde{X} - \Psi(\widetilde{X})^T \Psi(\widetilde{X}) \right\|_F^2 .$$

- We can find this map using the SVD of $\widetilde{X} = U \Sigma V^T$ as

$$\Psi(\widetilde{X}) = U_s^T \widetilde{X} = \Sigma_s V_s^T ,$$

which is exactly the same as using the first s components of PCA!
- A drawback: too global and not incorporating local geometry

Laplacian Eigenmaps (Belkin & Niyogi, 2001–3)

- Incorporating local geometric information in $\mathbb{R}^d$ for the embedding
- Define the proximity weight $w(\mathbf{x}_i, \mathbf{x}_j)$, e.g., $w_\epsilon(\mathbf{x}_i, \mathbf{x}_j) := \mathrm{e}^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\epsilon^2}$.
- Now, seek Ψ that minimizes the following

$$J_{\mathrm{Lap}}(\Psi) := \sum_{i,j} \|\Psi(\mathbf{x}_i) - \Psi(\mathbf{x}_j)\|^2 w_\epsilon(\mathbf{x}_i, \mathbf{x}_j).$$

- This leads to the following optimization problem:

$$\min_{\Psi(X) \in \mathbb{R}^{s \times m_\star}} \mathrm{tr}\left(\Psi(X) L \Psi(X)^T\right) \quad \text{subject to } \Psi(X) D \Psi(X)^T = I,$$

where the matrices are defined as

$$W := \left(w_\epsilon(\mathbf{x}_i, \mathbf{x}_j)\right), \quad D := \mathrm{diag}\left(\sum_j w_\epsilon(\mathbf{x}_1, \mathbf{x}_j), \ldots, \sum_j w_\epsilon(\mathbf{x}_{m_\star}, \mathbf{x}_j),\right).$$

The matrix $L := D - W$ is called the (unnormalized) graph Laplacian.

Laplacian Eigenmaps ...

- This leads to the following generalized eigenvalue problem:

$$L\Psi(X)^T = D\Psi(X)^T\Lambda; \quad L \in \mathbb{R}^{m_\star \times m_\star},\ \Lambda \in \mathbb{R}^{s\times s},$$
$$\Updownarrow$$
$$L_{\mathrm{rw}}\Psi_{\mathrm{rw}}(X)^T = \Psi_{\mathrm{rw}}(X)^T\Lambda_{\mathrm{rw}}; \quad L_{\mathrm{rw}} := D^{-1}L = I - D^{-1}W.$$

- $\Psi_{\mathrm{rw}}(X) \in \mathbb{R}^{s\times m_\star}$ is the Laplacian Eigenmap of X.
- Another possibility is:

$$L_{\mathrm{sym}}\Psi_{\mathrm{sym}}(X)^T = \Psi(X)^T_{\mathrm{sym}}\Lambda_{\mathrm{sym}}; \quad L_{\mathrm{sym}} := D^{-\frac{1}{2}}LD^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}.$$

- Both L_{rw} and L_{sym} are called the normalized graph Laplacians (rw = 'random walk'; sym = 'symmetric').

$$\Psi_{\mathrm{rw}}(X) = \Psi_{\mathrm{sym}}(X)D^{-\frac{1}{2}}, \quad \Lambda_{\mathrm{rw}} = \Lambda_{\mathrm{sym}}.$$

- Eigenvalues are sorted in increasing order; $L_{\mathrm{rw}}\mathbf{1} = \mathbf{0}$.
- A drawback: sensitive to sampling density on a manifold.

Diffusion Maps (Coifman & Lafon 2004–6)

- Focus on the normalized weighted adjacency matrix $A_{\mathrm{rw}} := D^{-1}W$.
- Interpret A_{rw} as the transition matrix of a random walk on X or the diffusion operator on X. A_{rw}^t = running the random walk t steps.
- Perform density invariant normalization on W, i.e., $\widetilde{W} := D^{-1}WD^{-1}$ first. Then, do the row-stochastic normalization, i.e., $\widetilde{A}_{\mathrm{rw}} := \widetilde{D}^{-1}\widetilde{W}$ where $\widetilde{D}$ is the degree matrix (diagonal) of $\widetilde{W}$.
- Finally perform the eigenanalysis:

$$\widetilde{A}_{\mathrm{rw}}\Psi_{\mathrm{DM}}(X)^T = \Psi_{\mathrm{DM}}(X)^T\Lambda_{\mathrm{DM}},$$

  where the eigenvalues are sorted in decreasing order; $\widetilde{A}_{\mathrm{rw}}\mathbf{1} = \mathbf{1}$.
- Diffusion map is defined as:

$$\Psi_{\mathrm{DM}}^t(X) := \Lambda_{\mathrm{DM}}^t\Psi_{\mathrm{DM}}(X).$$

- Relationship with the Laplacian eigenmap (if $\widetilde{W}$ is used in L_{rw}):

$$\Psi_{\mathrm{DM}}^1(X) = \Psi_{\mathrm{rw}}(X); \quad \Lambda_{\mathrm{DM}} = I - \Lambda_{\mathrm{rw}}.$$

Remarks

- Can use SVD or symmetric eigenvalue solver for computing these embedding maps
- Choosing a good scale parameter ϵ for both LE and DM is not easy $\Longrightarrow \epsilon =$ the mean of the k-nearest neighbor distances. But how to choose k? $\Longrightarrow$ Cross validation, etc.
- For DM, choosing t or when to stop the diffusion is another subtle question, which is quite dependent on ϵ and the decay of the eigenvalues.
- Choosing an appropriate value of s is yet another problem $\Longrightarrow$ Elongated K-means algorithm:
  G. Sanguinetti, J. Laidler, and N. D. Lawrence, "Automatic determination of the number of clusters using spectral algorithms," *Proc. 15th IEEE Workshop on Machine Learning for Signal Processing*, pp55–60, 2005.

Extension of Maps for Test Data

- For PCA, it is quite easy; simply the multiplication of U_s^T to Y.
- For LE/DM, it is more involved and the following geometric harmonics multiscale extension algorithm is necessary:
  S. Lafon, Y. Keller, R. R. Coifman, "Data fusion and multicue data matching by diffusion maps," *IEEE Trans. Pattern Anal. Machine Intell.*, vol.28, no.11, pp.1784–1797, 2006.

Outline

Earth Mover's Distance (EMD)

- Originated from the Monge-Kantorovich optimal transport problem
- Used successfully in image retrieval from large databases, image registration and warping, etc.
- Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *Intern. J. Comput. Vision*, vol.40, no.2, pp.99–121, 2000.
- S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent, "Optimal mass transport for registration and warping," *Intern. J. Comput. Vision*, vol.60, no.3, pp.225–240, 2004.
- More robust (for our classification problems) than the Hausdorff distance (HD) between two ensembles $\Psi(X^i)$, $\Psi(Y^j)$ in the reduced embedding space, which was used by Lafon-Keller-Coifman.

$$d_H(\Psi(X^i), \Psi(Y^j)) := \max\left(\max_{\mathbf{y}\in\Psi(Y^j)} \min_{\mathbf{x}\in\Psi(X^i)} \|\mathbf{x}-\mathbf{y}\|, \max_{\mathbf{x}\in\Psi(X^i)} \min_{\mathbf{y}\in\Psi(Y^j)} \|\mathbf{x}-\mathbf{y}\| \right).$$

Earth Mover's Distance (EMD) ...

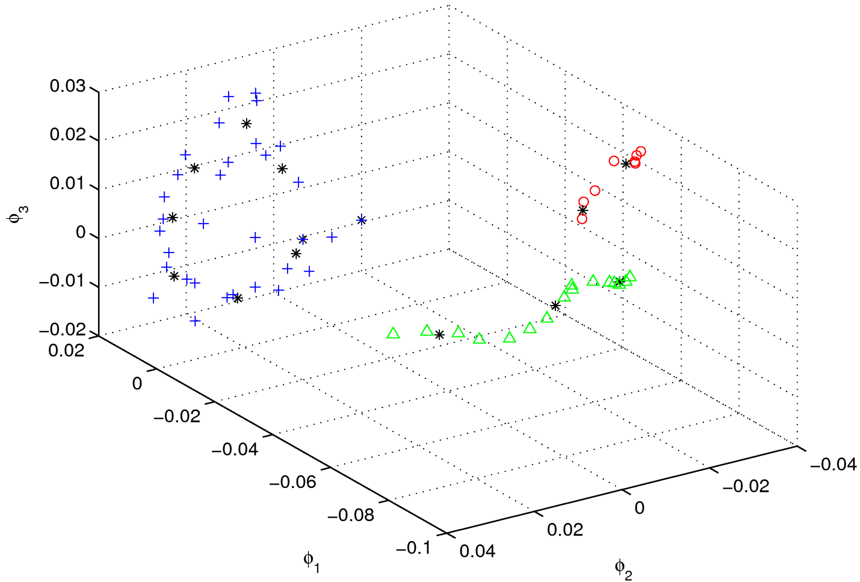
Let $P = \{(\mathbf{x}_1, p_1), \ldots, (\mathbf{x}_m, p_m)\}$ and $Q = \{(\mathbf{y}_1, q_1), \ldots, (\mathbf{y}_n, q_n)\}$ be two signatures characterizing two classes or objects of interest. $\mathbf{x}_i, \mathbf{y}_j \in \mathbb{R}^s$ are cluster centers and p_i, q_j are populations (or mass) of the corresponding clusters. Then, the Earth Mover's Distance (EMD) is defined by

$$\mathrm{EMD}(P, Q) := \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} c_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}},$$

where

- c_{ij} is the cost of moving one unit mass from the ith cluster in P to the jth cluster in Q. A typical example: $c_{ij} = \frac{1}{2}\|\mathbf{x}_i - \mathbf{y}_j\|^2$.
- $f_{ij} \geq 0$: the optimal flow between two distributions that minimizes the total cost $\sum_{i=1}^m \sum_{j=1}^n f_{ij} c_{ij}$, subject to the following constraints:
  - $\sum_{i=1}^m f_{ij} \leq q_j$, $j = 1, \ldots, n$;
  - $\sum_{j=1}^n f_{ij} \leq p_i$, $i = 1, \ldots, m$;
  - $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min\{\sum_{i=1}^m p_i, \sum_{j=1}^n q_j\}$.

Signatures in the Reduced Embedding Space (again)

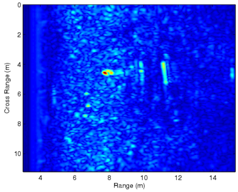


Outline

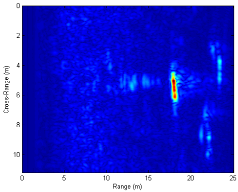
Outline

Underwater Object Classification

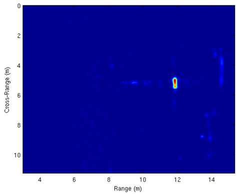
- Sonar waveforms in the acoustic scattering experiments were collected in a fresh water pond at Naval Surface Warfare Center (NSWC), Panama City, FL.
- Three experiments on different days were performed. Each time, there were two objects in the pond.
  1. C1: Buried Al cylinder; S1: Fe Sphere filled with air
  2. C2: Proud Al cylinder; S2: Fe Sphere filled with silicone oil
  3. C3: Shorter proud Al cylinder; S3 = S2
- Source: frequency 20kHz; sinusoidal shape; 0.2msec duration
- Received waveforms were sampled at rate 500kHz



(a) Buried



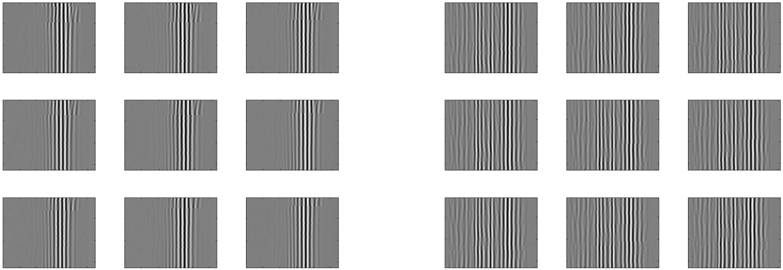
(b) Proud



(c) Short Proud

Underwater Object Classification . . .

- Our objective is to classify objects according to their material compositions independent of shapes, sizes, buried or proud.
- Each data point is in $\mathbb{R}^{17\times600}$; The number of data points in C1, C2, C3, S1, S2, S3 are 8, 8, 16, 32, 32, 32, respectively.
- Pick one of these 6 ensembles as a test ensemble $Y = Y^1$ whereas the other 5 ensembles are used as training ensembles $X = \bigcup_{i=1}^{5} X^i$. Then do the classification of Y.
- Repeat this process 5 more times.



Underwater Object Classification: Results

Object		$C1$	$C2$	$C3$	$S1$	$S2$	$S3$
True Label		**Al**	**Al**	**Al**	**IA**	**IS**	**IS**
PCA	EMD	**Al**	**Al**	**Al**	**IS**	**IS**	**IA**
	HD	**Al**	**Al**	**Al**	**IS**	**IS**	**IA**
$\mathrm{LE}_{\mathrm{rw}}$	EMD	**Al**	**Al**	**Al**	**Al**	**IS**	**IS**
	HD	**Al**	**Al**	**Al**	**Al**	**Al**	**IS**
$\mathrm{LE}_{\mathrm{sym}}$	EMD	**Al**	**Al**	**Al**	**Al**	**IS**	**IS**
	HD	**Al**	**Al**	**Al**	**Al**	**IS**	**IS**
DM	EMD	**Al**	**Al**	**Al**	**Al**	**IS**	**IS**
	HD	**Al**	**Al**	**Al**	**Al**	**IS**	**IS**

Al = Aluminum; IA = Iron-Air; IS = Iron-Silicone Oil

Underwater Object Classification: EMD vs HD

EMD and HD values in the $\mathrm{LE}_{\mathrm{rw}}$ coordinates between $S2$ and all other objects

Object	$C1$	$C2$	$C3$	$S1$	$S3$
EMD	0.0070	0.0064	0.0057	0.0085	0.0053
HD	0.1917	0.2374	0.1237	0.1500	0.1684

Outline

Video Clip Classification: Lip Reading

- Lips speaking five digits, 'one', . . . , 'five' were captured by a camcorder with the rate 60 frames/second.
- Each video frame is cropped to have 55×70 pixels.
- A single speaker spoke each digit 10 times (i.e., totally 50 video clips).
- Each video clip consists of $30 \sim 63$ video frames.
- Split the whole data randomly into the training and test ensembles as $X = \bigcup_{i=1}^{25} X^i$, $Y = \bigcup_{j=1}^{25} Y^j$. Then, do the classification.
- Repeat this process 99 times more.

Lip-Reading total recognition errors (averaged over 100 trials)

PCA EMD	PCA HD	$\mathrm{LE}_{\mathrm{rw}}$ EMD	$\mathrm{LE}_{\mathrm{rw}}$ HD	$\mathrm{LE}_{\mathrm{sym}}$ EMD	$\mathrm{LE}_{\mathrm{sym}}$ HD	DM EMD	DM HD
5.3%	9.4%	36.1%	36.1%	26.0%	27.6%	24.1%	25.2%

Outline

Conclusions & Future Plan

- The key for the signal ensemble classification was to use the appropriate dimensionality reduction techniques with the robust distance measure like EMD;
- The best choice of the dimensionality reduction depends on the data; this is particularly so for the real data.
- Global (PCA) vs Local (LE/DM): Lip-reading video clips involve more global trajectories while sonar waveforms involve more localized clusters.
- Robustness of EMD was important compared to HD.
- Comparison with the other ideas of ours: node connectivity matching that do not require the eigenvalue/eigenvector computations;
- Comparison with explicit feature extraction techniques such as Local Discriminant Basis

Outline

e

References

- Laplacian Eigenfunction Resource Page
  http://www.math.ucdavis.edu/~saito/lapeig/ contains
  - All the talk slides of the special session on "Kernel Methods in Data Analysis, " which Yosi and I organized at IEEE Workshop on Statistical Signal Processing; and
  - My Course Note (elementary) on "Laplacian Eigenfunctions: Theory, Applications, and Computations"
- The following article is available at
  http://www.math.ucdavis.edu/~saito/publications/
  - L. Lieu and N. Saito: "Signal ensemble classification using low-dimensional embeddings and Earth Mover's Distance," to appear in *Wavelets: Old and New Perspectives* (J. Cohen and A. Zayed, eds.), Birkhuser, 2010.

Thank you very much for your attention!