Parametrization of Datasets with Low-distortion Embeddings

François Meyer

Center for the Study of Brain, Mind and Behavior,
Program in Applied and Computational Mathematics
Princeton University

fmeyer@princeton.edu
http://www.princeton.edu/~fmeyer

Random Shapes, Tutorials, IPAM 2007

Acknowledgments

- R.R. Coifman, S. Lafon, M. Maggioni
- M. Ramírez-Vélez, D.S. Barth
- National Institutes of Health
- IPAM MGA program
- IPAM Graduate Summer School: Intelligent Extraction of Information from Graphs and High Dimensional Data

Outline

1. Introduction
   - Assumptions about the data
   - Our definition of the problem
2. Constructing a new parametrization
   - A random walk on the dataset
   - A new way to measure distances
3. The spectral connection
   - Spectral graph theory
   - From commute time to spectral geometry
4. Classification of EEG recordings

Introduction

- Explosion of high-dimensional datasets:
  web, biology, medicine, etc.
- New tools for data exploration and analysis address issues:
  - signal of interest: complicated geometry
  - data corrupted by noise,
  - algorithm complexity

Example 1: Neuroimaging

- dogma: each brain region is responsible for a specific function
- goal: delineation of functional anatomy in terms of spatial and temporal organization
- method:
  - very simple cognitive or sensory input stimulus
  - measure the output signal $\boldsymbol{x}_i$ at each voxel i inside the brain
  - detect significant changes in the signal

Example 1: fMRI of Natural Stimuli

- challenge: study the response to complex stimuli ("real life")
- example: subject watches a movie in the MRI scanner
- discover neuronal networks involved in complex tasks
- how is the analysis performed ?
- size of the problem: 200,000 time series in $\mathbb{R}^{1000}$

Example 2: Prediction of seizure from EEG

- Electroencephalogram: electrical recordings on the scalp
- seizure $\rightarrow$ time-frequency changes in the signal
- goal: predict the seizure before the onset
- best existing method: brain = nonlinear dynamical system
- neuronal synchrony: fewer independent variables needed to describe the EEG recordings ?
- Is the brain during a seizure a low dimensional dynamical system ?
- size of the problem: 100,000 brain states in $\mathbb{R}^{100}$

Outline

1. Introduction
   - Assumptions about the data
   - Our definition of the problem
2. Constructing a new parametrization
   - A random walk on the dataset
   - A new way to measure distances
3. The spectral connection
   - Spectral graph theory
   - From commute time to spectral geometry
4. Classification of EEG recordings

Assumptions about the data

- dataset: large number of internal microscopic variables
  $\rightarrow$ many degrees of freedom
- at a macroscopic scale: many variables are coupled
  $\rightarrow$ set of all possible configurations for the signals is low dimensional
- signals varies smoothly as a function of “hidden” variables
  $\rightarrow$ well defined low dimensional structure

Outline

1. Introduction
   - Assumptions about the data
   - Our definition of the problem
2. Constructing a new parametrization
   - A random walk on the dataset
   - A new way to measure distances
3. The spectral connection
   - Spectral graph theory
   - From commute time to spectral geometry
4. Classification of EEG recordings

Our definition of the problem

1. $T \times N$ dataset $\mathbf{X} = \left[\boldsymbol{x}_0 | \boldsymbol{x}_1 | \cdots | \boldsymbol{x}_{N-1}\right]$
2. goal: construction of a new parameterization
   $\phi : \mathbb{R}^T \to \mathbb{R}^K$, with $K \ll T$,
   $$\boldsymbol{x}_i \mapsto \phi(\boldsymbol{x}_i),$$
3. similar signals are mapped to the same region of the atlas:
   - $\|\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)\| \approx \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$

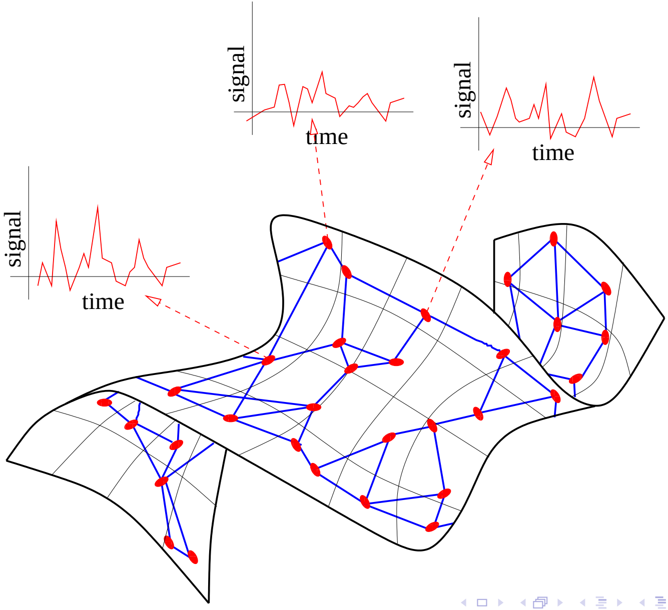
Outline

Construction of the graph

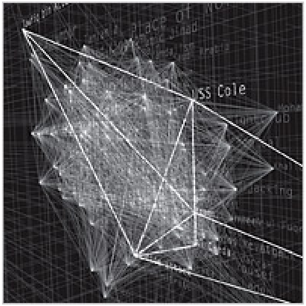
- replace the dataset by a graph G,
- vertex i of the graph $= \mathbf{x}_i$
- edges: k nearest neighbors according to $\|\mathbf{x}_i - \mathbf{x}_j\| = (\sum_{t=1}^{T} (x_i(t) - x_j(t))^2)^{1/2}$
- weight $w_{i,j}$ on the edge $\{i, j\}$: proximity between i and j,
- for instance,

$$w_{i,j} = \begin{cases} e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2}, & \text{if } \mathbf{x}_i \text{ is connected to } \mathbf{x}_j, \\ 0 & \text{otherwise.} \end{cases}$$

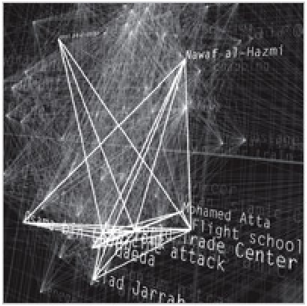
From the dataset to the graph



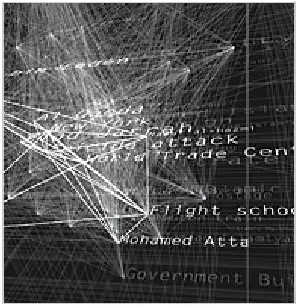
From the dataset to the graph



OCT. 12, 2000



JUNE 2000 TO JUNE 2001



SEPT. 11, 2001

A random walk on the graph

- weighted graph G, matrix $\boldsymbol{W}$, $W_{i,j} = w_{i,j}$
- random walk on the graph with transition probability $\mathbf{P}$,

$$P_{i,j} = w_{i,j}/d_i,$$

- $d_i = \sum_j w_{i,j}$: degree of the vertex i, $\mathbf{D}$ diagonal matrix,

$$D_{ii} = d_i = \sum_j W_{i,j}. \tag{1}$$

- $\boldsymbol{\pi} = \frac{1}{\sum_{i,j} w_{i,j}}[d_1, d_2, \cdots, d_N]$ stationary distribution

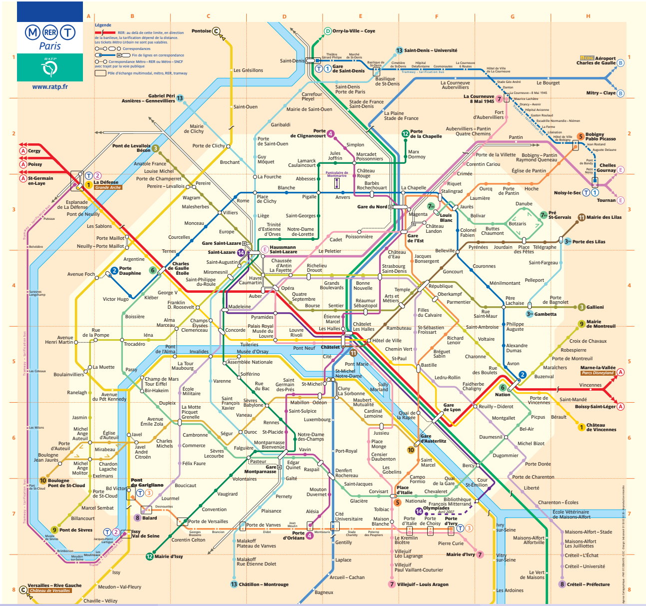
Outline

A good distance on the graph

- similarity measure between any two vertices i and j
- distinguish between strongly connected vertices and weakly connected vertices
- solution: average commute time, $\kappa(i,j) = H(j,i) + H(i,j)$
- symmetric version of the average hitting time from i to j,

$$H(i,j) = E_i[T_j] \quad \text{with} \quad T_j = \min\{n \geqslant 0; Z_n = j\}.$$

  E_i: random walk is started at i
- κ is a distance:
  1. $\kappa(i,j) = 0 \implies i = j$,
  2. $\kappa(i,j) \leqslant \kappa(i,k) + \kappa(k,j)$,
  3. $\kappa(i,j) = \kappa(j,i)$.



Paris
RATP
www.ratp.fr
Légende
Correspondances
Fin de lignes en correspondance
Pôle d'échange multimodal, métro, RER, tramway
Pontoise
Orry-la-Ville – Coye
Saint-Denis – Université
Aéroport Charles de Gaulle
Mitry – Claye
Gabriel Péri Asnières – Gennevilliers
Les Grésillons
Saint-Denis
Gare de Saint-Denis
Basilique de St-Denis
Saint-Denis Porte de Paris
Stade de France Saint-Denis
La Plaine Stade de France
La Courneuve Aubervilliers
La Courneuve 8 Mai 1945
Le Bourget
Fort d'Aubervilliers
Aubervilliers – Pantin Quatre Chemins
Bobigny Pablo Picasso
Chelles Gournay
Noisy-le-Sec
Tournan
Cergy
Poissy
St-Germain en-Laye
Pont de Levallois Bécon
La Défense Grande Arche
Esplanade de La Défense
Pont de Neuilly
Les Sablons
Porte Maillot
Porte de Clichy
Porte de Clignancourt
Porte de la Chapelle
Porte de la Villette
Bobigny – Pantin Raymond Queneau
Église de Pantin
Mairie des Lilas
Pré St-Gervais
Porte des Lilas
Gare Saint-Lazare
Haussmann Saint-Lazare
Charles de Gaulle Étoile
Porte Dauphine
Avenue Foch
Gare du Nord
Gare de l'Est
République
Opéra
Madeleine
Concorde
Châtelet Les Halles
Châtelet
Hôtel de Ville
Gallieni
Gambetta
Mairie de Montreuil
Marne-la-Vallée Parcs Disneyland
Boissy-Saint-Léger
Château de Vincennes
Vincennes
Nation
Gare de Lyon
Bastille
Gare d'Austerlitz
Invalides
Musée d'Orsay
Champ de Mars Tour Eiffel
Javel
Boulogne Pont de St-Cloud
Pont du Garigliano
Balard
Issy Val de Seine
Mairie d'Issy
Pont de Sèvres
Gare Montparnasse
Place d'Italie
Bibliothèque François Mitterrand
Olympiades
Porte d'Orléans
Châtillon – Montrouge
Villejuif – Louis Aragon
Mairie d'Ivry
Créteil – Préfecture
Versailles – Rive Gauche Château de Versailles
Chaville – Vélizy
Meudon – Val-Fleury
Bagneux
Arcueil – Cachan
Les Ardoines

How does $\kappa(i,j)$ compare to $\delta(i,j)$?

- $\kappa(i,j)$ can be compared to the standard distance δ on the graph

Theorem

If i and j are at a distance $\delta(i,j)$ on the graph, then

$$2\delta(i,j) \leqslant \kappa(i,j) \leqslant C\delta(i,j),$$

where $C = \max_{i,j} \frac{1}{\pi_i P_{i,j}} = \frac{\sum_{i,j} w_{i,j}}{\min_{i,j} w_{i,j}}$

- Markov chain is reversible, $\pi_i P_{i,j} = \pi_j P_{j,i}$
- C can be large

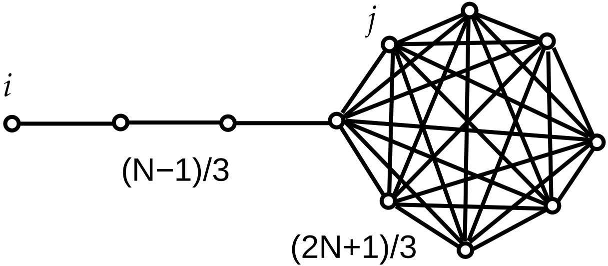
Your worst commute time in *L.A.*: Sepulveda Blvd or 405 ?

Maximum commute time: lost in the city...

- among all graphs with N vertices,
  what is the graph with the largest $\kappa(i, j)$?

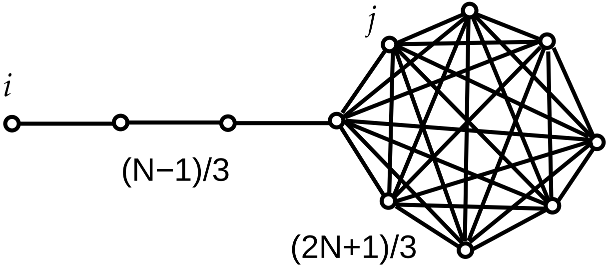
Maximum commute time: lost in the city...

- among all graphs with N vertices, what is the graph with the largest $\kappa(i,j)$?
- lollipop graph: path with $(N-1)/3$ vertices, complete subgraph with $(2N+1)/3$ vertices



Maximum commute time: lost in the city...

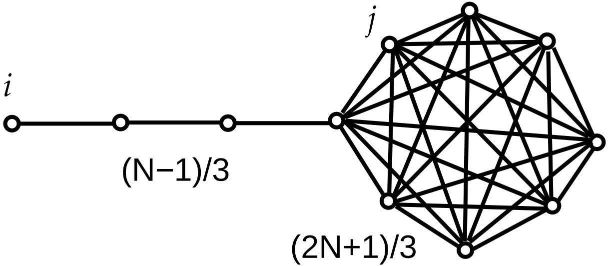
- among all graphs with N vertices,
  what is the graph with the largest $\kappa(i, j)$?
- lollipop graph: path with $(N-1)/3$ vertices,
  complete subgraph with $(2N+1)/3$ vertices



- $\kappa(i, j) = \frac{4}{27} N^3 + \mathcal{O}(N)$

Maximum commute time: lost in the city...

- among all graphs with N vertices,
  what is the graph with the largest $\kappa(i, j)$?
- lollipop graph: path with $(N-1)/3$ vertices,
  complete subgraph with $(2N+1)/3$ vertices



- $\kappa(i, j) = \frac{4}{27}N^3 + \mathcal{O}(N)$
- $\delta(i, j) = \frac{2}{3}N$, $C = 2N(2N+1)/18$
- [Jonasson, 2000]

Outline

The spectral connection

- Fundamental matrix $\mathbf{Z} = (I - (\mathbf{P} - \mathbf{\Pi}))^{-1} = I + \sum_{k \geqslant 1} \mathbf{P}^k - \mathbf{\Pi}$ with $\mathbf{\Pi}^T = [\boldsymbol{\pi}_1 | \cdots | \boldsymbol{\pi}_N]$
- $\mathbf{Z}$ is the Green function of the Laplacian, $I - \mathbf{L}$

Theorem

[Bremaud, 1999] Hitting time $E_i[T_j] = (Z_{j,j} - Z_{i,j})/\pi_j$.

- $E_i[T_j] = 1 + \sum_{k;k \neq j} P_{i,k} E_i[T_k]$
- eigenfunctions $\boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_N$ of

$$\mathbf{D}^{\frac{1}{2}} \mathbf{P} \mathbf{D}^{-\frac{1}{2}}, \tag{2}$$

with eigenvalues $-1 \leqslant \lambda_N \cdots \leqslant \lambda_2 < \lambda_1 = 1$.

The spectral connection

- commute time:

$$\kappa(i,j) = \sum_{k=2}^{N} \frac{1}{1-\lambda_k} \left(\frac{\phi_k(i)}{\sqrt{\pi_i}} - \frac{\phi_k(j)}{\sqrt{\pi_j}} \right)^2. \tag{3}$$

- define an embedding

$$i \mapsto I_k(i) = \frac{1}{\sqrt{1-\lambda_k}} \frac{\phi_k(i)}{\sqrt{\pi_i}}, \quad k = 2, \cdots, N \tag{4}$$

- Euclidean distance on the image of the embedding
  = commute time

$$\kappa(i,j) = \|I(i) - I(j)\|^2 = \sum_{k=2}^{N} |I_k(i) - I_k(j)|^2$$

Outline

The Laplacian connection

- $\boldsymbol{\phi}_k$ is also an eigenfunction of the Laplacian

$$\mathcal{L} = I - \mathbf{D}^{\frac{1}{2}} \mathbf{P} \mathbf{D}^{-\frac{1}{2}},$$

  with the eigenvalues $\beta_k = 1 - \lambda_k$.
- $\boldsymbol{\phi}_k$ minimizes the "distortion"

$$\min_{\boldsymbol{\phi}, \|\boldsymbol{\phi}\|=1} \frac{\sum_{[i,j]} w_{i,j} (\boldsymbol{\phi}(i) - \boldsymbol{\phi}(j))^2}{\sum_i d_i \boldsymbol{\phi}^2(i)}$$

  with $\boldsymbol{\phi}_k$ orthogonal to $\{\boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \cdots, \boldsymbol{\phi}_{k-1}\}$.
- Laplacian eigenmaps [Belkin and Niyogi, 2003]

The diffusion distance connection

- [Lafon, 2004, Coifman and Lafon, 2006]
- diffusion distance,

$$D_t^2(i,j) = \sum_{k=2}^{N} \lambda_k^{2t} \left(\frac{\phi_k(i)}{\sqrt{\pi_i}} - \frac{\phi_k(j)}{\sqrt{\pi_j}} \right)^2 . \tag{5}$$

- commute time = sum of the diffusion distance at all scale t

$$\sum_{t=0}^{\infty} D_{t/2}^2(i,j) = \sum_{k=2}^{N} \frac{1}{1-\lambda_k} \left(\frac{\phi_k(i)}{\sqrt{\pi_i}} - \frac{\phi_k(j)}{\sqrt{\pi_j}} \right)^2 = \kappa(i,j)$$

The spectral geometry connection

- data sampled on a n-dimensional manifold $\mathcal{M}$
- $\mathcal{M}$ embedded by its heat kernel $K_{\mathcal{M}}(t,x,y)$

Theorem

[Bérard et al., 1994]

$$\psi_t : \mathcal{M} \mapsto l^2(\mathbb{R}) \tag{6}$$

$$x \mapsto \left\{ \sqrt{2}(4\pi)^{n/4} t^{(n+2)/4} e^{-\lambda_j t/2} \phi_k(x) \right\}_{k \geqslant 1} \tag{7}$$

$\forall t > 0$, the map ψ_t is an embedding of $\mathcal{M}$ into $l^2(\mathbb{R})$.

- scale parameter t: similar to diffusion distance

The spectral geometry connection

ψ_t: composition of

1. embedding of $\mathcal{M}$ by the heat kernel:
   each point on $\mathcal{M}$ is mapped to a bump function.

$$\mathcal{M} \mapsto L^2(\mathcal{M}) \tag{8}$$

$$x \mapsto K_{\mathcal{M}}(t/2, x, .) \tag{9}$$

2. isometry given by the choice of basis, $\{\Phi_1, \Phi_2, \cdots\}$, of $L^2(\mathcal{M})$,
   each function of $L^2(\mathcal{M})$ is expanded into the basis of eigenfunctions of the Laplace-Beltrami operator

$$L^2(\mathcal{M}) \mapsto l^2(\mathbb{R}) \tag{10}$$

$$f \mapsto \{< f, \phi_k >\}_{k \geqslant 1} \tag{11}$$

Algorithm 1: Construction of the embedding

Input:

- $\mathbf{x}_i(t), t = 0, \cdots, T-1, \ i = 1, \cdots, N,$
- σ ; n_n number of nearest neighbors.
- K: number of eigenfunctions.

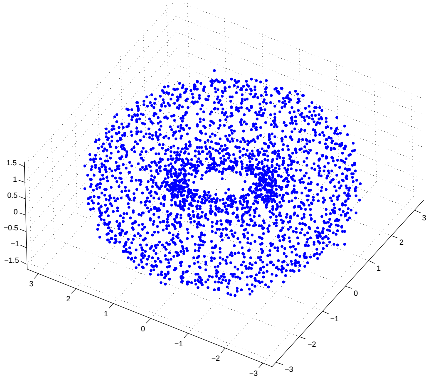
Algorithm:

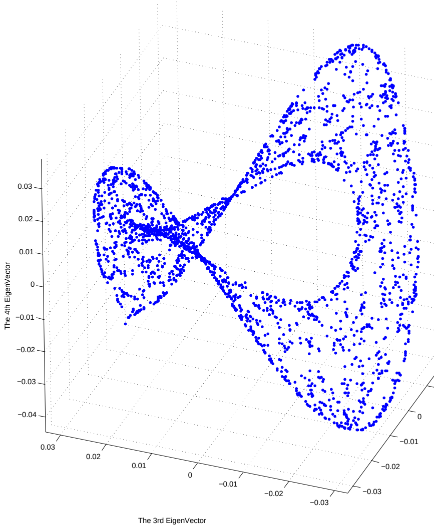
1. construct the graph defined by the n_n nearest (according to $\|\mathbf{x}_i - \mathbf{x}_j\|$) neighbors of each $\mathbf{x}_i$
2. compute $\mathbf{P}$
3. find the first K eigenfunctions, ϕ_k, of $\mathbf{D}^{\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}}$

Output: For all $\mathbf{x}_i$

- new co-ordinates of $\mathbf{x}_i$: $\left\{ \frac{1}{1-\lambda_k} \frac{\phi_k(i)}{\sqrt{\pi_i}} \right\}, \quad k = 2, \cdots, N$

A toy example



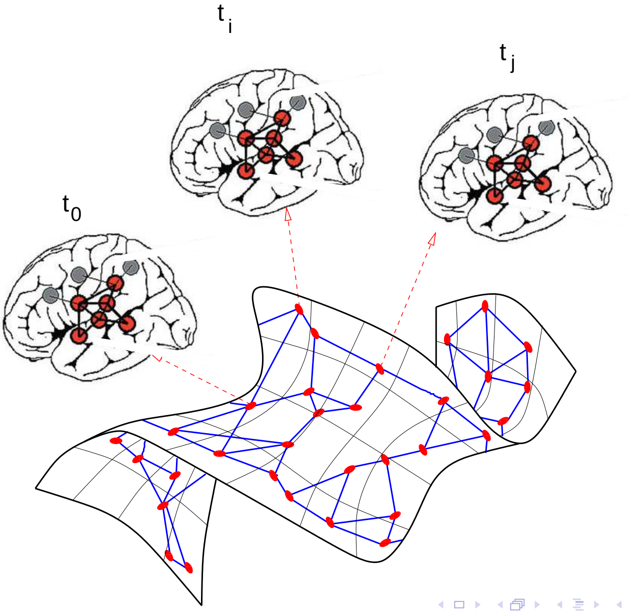


- classification of EEG recordings into baseline and ictal states
- Hypothesis: we can find a lower dimensional representation for classification
- More details can be found in [Ramírez-Vélez et al., 2006, Meyer and Shen, 2007]

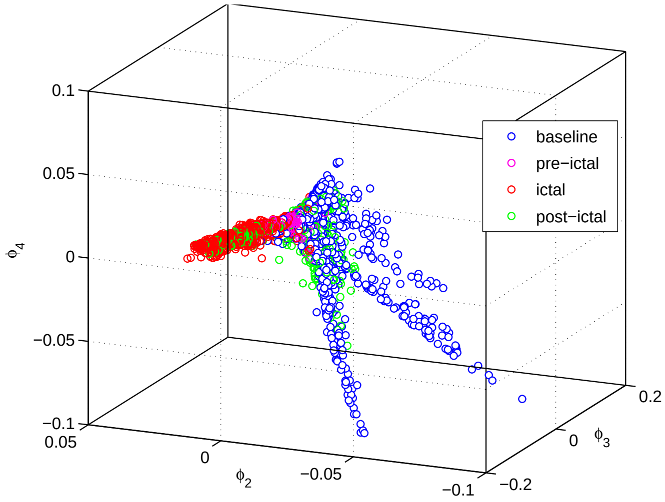
Human dataset

- scalp electroencephalograms
- 55 electrode channels, lowpass filtered at 256 Hz.
- baseline, pre-ictal, ictal,
  and post-ictal time segments
- each node of the graph is in $\mathbb{R}^{55}$

The graph of the brain dynamics

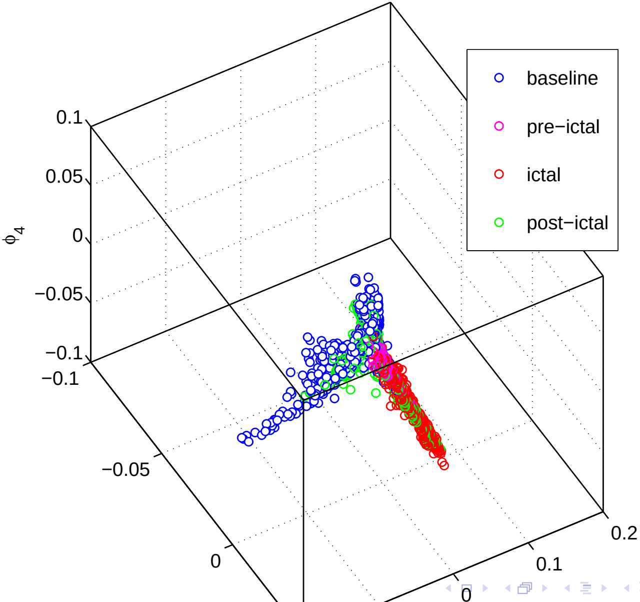


Embedding



baseline
pre−ictal
ictal
post−ictal
0.1
0.05
0
−0.05
−0.1
ϕ_4
0.05
0
−0.05
−0.1
ϕ_2
0.2
0
−0.2
ϕ_3

Embedding (2)



baseline
pre−ictal
ictal
post−ictal
ϕ_4
0.1
0.05
0
−0.05
−0.1
−0.1
−0.05
0
0
0.1
0.2

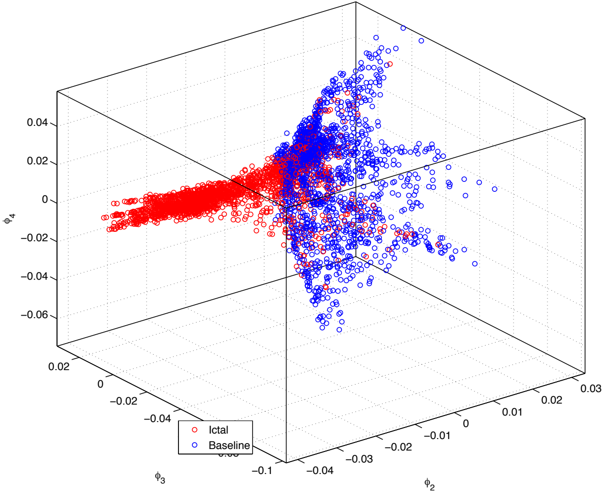
Classification

- $d = 10$ dimensions
- 10-fold cross validation

Table: Classification error (kernel ridge regression)

	Baseline	Ictal	Total
Raw Data	93.20	70.40	81.80
PCA	81.60	78.80	82.20
Random walk	100	82.80	91.40

Classification



Estimated labels: red=ictal, blue=baseline

Computational details

Curse of dimensionality:

- Fast nearest neighbors in high dimension
- Eigensolvers for large ($N = 10^5 - 10^6$) sparse matrices

Open questions

- much faster eigensolvers are needed...Matlab blows up for $N > 10,000$
- real time update ϕ_k with new incoming data ?
- ϕ_k: sensitive to σ and n_n
- how many new co-ordinates (local dimension) ?

Belkin, M. and Niyogi, P. (2003).
Laplacian eigenmaps for dimensionality reduction and data representation.
Neural Computations, 15:1373–1396.

Bérard, P., Besson, G., and Gallot, S. (1994).
Embeddings Riemannian manifolds by their heat kernel.
Geometric and Functional Analysis, 4(4):373–398.

Bremaud, P. (1999).
Markov Chains.
Springer Verlag.

Coifman, R. and Lafon, S. (2006).
Diffusion maps.
Applied and Computational Harmonic Analysis, 21:5–30.

Jonasson, J. (2000).
Lollipop graphs are extremal for commute times.
Random Structures and Algorithms, 16(2):131–142.

- Lafon, S. (2004).
  Diffusion maps and geometric harmonics.
  PhD thesis, Yale University, New Haven.

- Meyer, F. and Shen, X. (2007).
  Exploration of high dimensional biomedical datasets with low-distortion embedding.
  In *Proc. Data Mining for Biomedical Informatics Workshop, 7th SIAM International Conference on Data Mining.*
  To appear.

- Ramírez-Vélez, M., Staba, R., Barth, D., and Meyer, F. (2006).
  Nonlinear classification of EEG data for seizure detection.
  In *Proc. IEEE International Symposium on Biomedical Imaging: Macro to Nano*, pages 956–959.