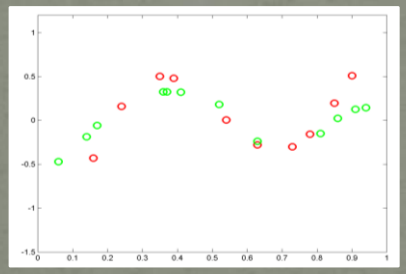
A Reproducing Kernel Hilbert Space Framework for Pairwise Time Series Distance

Zhengdong Lu, Todd Leen, Catherine Huang, Deniz Erdogmus
CSEE Department, OGI, OHSU

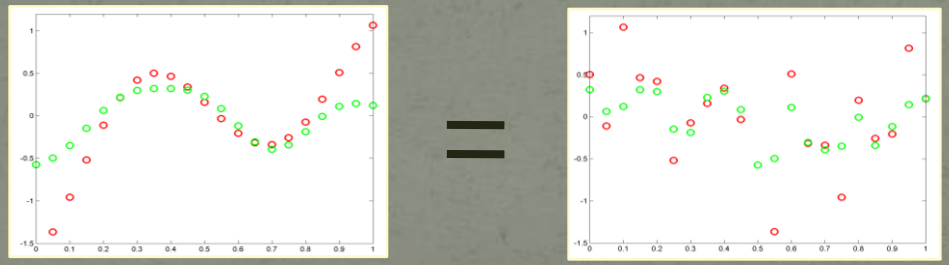
Distance for Time Series

- It is useful to calculate distance for time series
  - Retrieval, visualization, classification etc

  but often difficult
  - We often have only discrete observations made at irregular time intervals, or have different number of observations for each time series



  - We need to consider the temporal structure . Therefore even when the time series are synchronized, the point-wise distance is not desired.

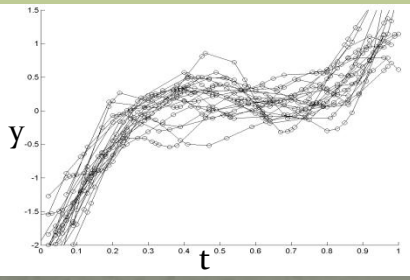


- Our approach: one way to circumvent the two difficulties

The Framework

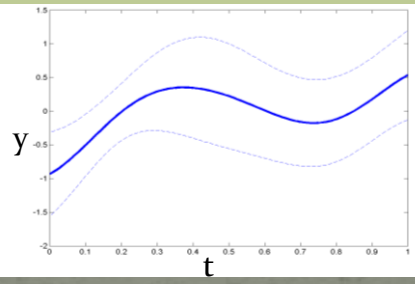
Our approach synthesizes ideas from functional data analysis, Gaussian process, Bregman divergence, and non-parametric mixed-effect model

Discrete Observations from Many Individuals

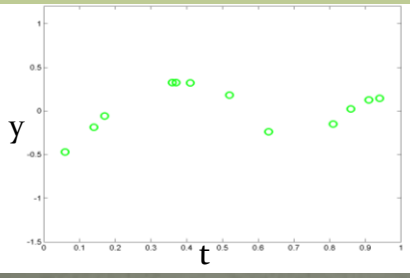


Learning with Non-parametric Mixed-effect Model

Gaussian Process

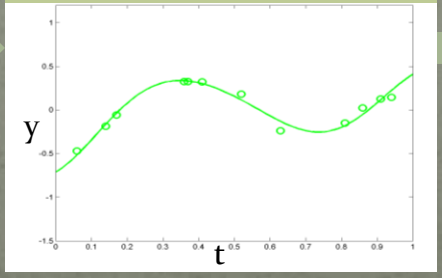


Discrete Observations from Individual Time Series *i*

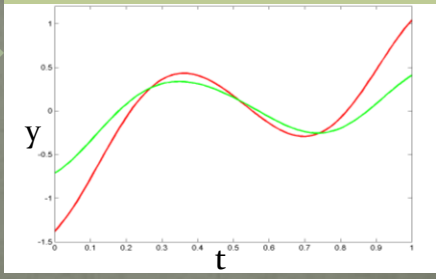


Regression

GP-based Smooth Function Representation



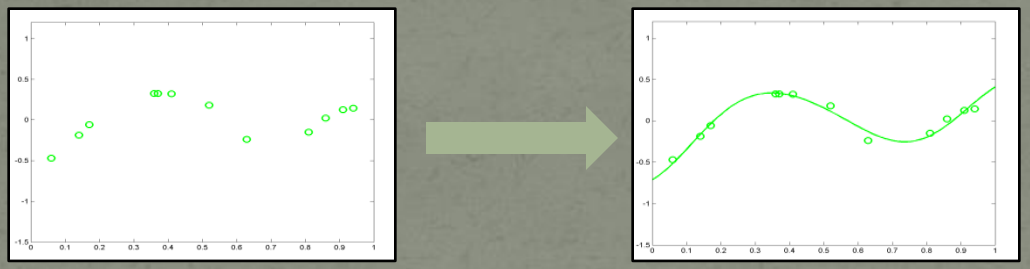
GP-based Bregman Divergence as the distance



Gaussian Processes

- Functional data analysis uses functions (curves) to represent discrete observations.

$\mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{iNi}]^T$ $f_i(t)$



- Gaussian Processes (GPs) provide a principled way for functional data analysis. Its probabilistic framework will later be exploited in deriving a distance measure and learning the regularizer (or equivalently, the kernel).
- Generative Model:
  - Prior for functions:

$$p[f] \propto \exp(-\frac{1}{2}||f - f_0||^2_{\mathcal{H}})$$ (with kernel K)

  - Observation model:

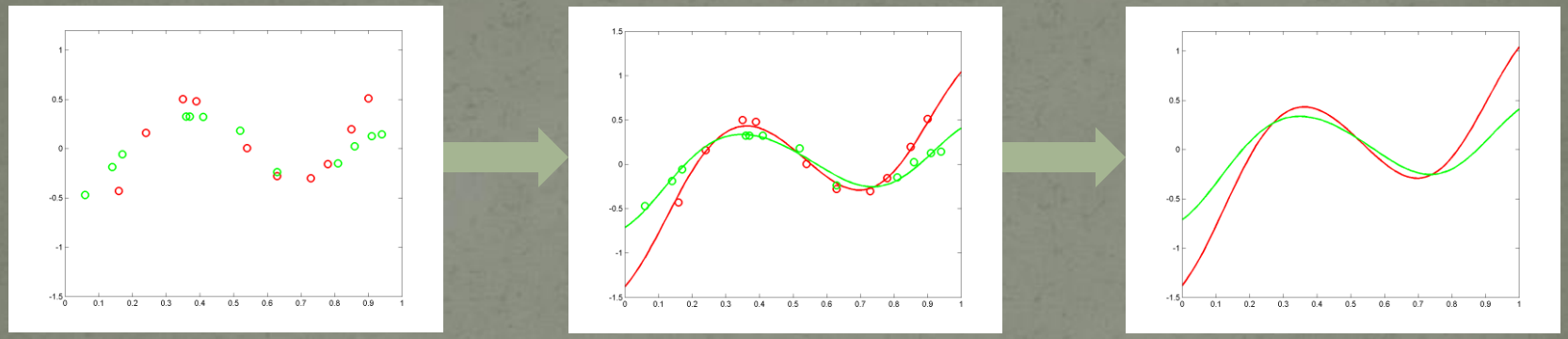
$$y_{in} = f_i(t_{in}) + \epsilon_{in}, \quad n = 1, 2, \cdots, N_i$$

- Regression: (mapping from observations to smooth curves)

$$\begin{aligned} \hat{f}_i(t) &= E[f_i(t)|\mathbf{y}_i, f_0; \mathbf{t}_i, K] \\ &= f_0 + K(t, \mathbf{t}_i)(K(\mathbf{t}_i, \mathbf{t}_i) + \sigma^2 \mathbb{I})^{-1}(\mathbf{y}_i - \mathbf{f}_{0,i}) \end{aligned}$$

Two Quesitons:

Functional data analysis uses functions (curves) to represent discrete observations.



But two questions remains:

QI : How do we calculate the distances between curves?

A: We use a distance derived from functional Bregman divergence and Gaussian processes

QII: How do we specify the Gaussian process?

A: We learn to specify the Gaussian process through non-parametric mixed-effect model, assuming there are many similar time series available

Bregman Divegence and Exponetinal Family

To answer QI: How do we calculate the distance between curves

We are going to derive a divergence measure for smooth curves based on Bregman divergence and exponential family

- Bregman divergence is a divergence measure based on a convex function $\phi(x)$

$$d_\phi(x_1||x_2) = \phi(x_1) - \phi(x_2) - \langle \nabla\phi(x_2), x_1 - x_2 \rangle$$

- The Bregman divergence can be related to exponential family distributions. More specially, any e-family distribution $p(x; \theta)$

$$p(x;\theta) = \exp(\langle x, \theta \rangle - \Phi(\theta))p_0(x),$$

can equivalently formulated as

$$\log p(x;\theta) = -d_\phi(x||\mu(\theta)) + \phi(x) + \log p_0(x)$$

where $\mu(\theta)$ is the expectation parameters corresponding to θ, and $\phi(x)$ is the conjugate function of Φ

$$\phi(x) = \sup_\theta \{\langle x, \theta \rangle - \Phi(\theta)\}$$

- We argue that the Bregman divergence $d_\phi(x_1||x_2)$ pro vides a reasonable model-weighted divergence between x_1 and x_2 associated with distribution $p(x; \theta)$.

- The Bregman divergence can be extended to space of functions.

Bregman Divergence on Space of Functions

To answer QI: How do we calculate the distance between curves ?

- Viewing Gaussian process

$$p[f] \propto \exp(-\frac{1}{2}||f - f_0||^2_{\mathcal{H}}),$$

as exponential family distribution for functions, we can calculate the corresponding (functional) Bregman divergence as

$$d_g(f_1||f_2) = g[f_1] - g[f_2] - \int Dg[f_2](f_1(t) - f_2(t))dt$$

where $g[f]$ is the corresponding seed functional and $Dg[\ ]$ is the Fr échet derivative.

- For Gaussian process, we simply have

$$d_{\mathcal{H}}(f_1||f_2) = \frac{1}{2}||f_1 - f_2||^2_{\mathcal{H}}$$

which will be used as the squared distance for curves f_1 and f_2.

- We can write the distance between two time series as the distance of two corresponding representing curves

$$d_{ij} = \frac{1}{2}||\hat{f}_i - \hat{f}_j||^2_{\mathcal{H}} = \frac{1}{2}\left\langle \hat{f}_i - \hat{f}_j, \hat{f}_i - \hat{f}_j \right\rangle_{\mathcal{H}} = \frac{1}{2}\mathbf{v}_i^T K(\mathbf{t}_i, \mathbf{t}_i)\mathbf{v}_i + \frac{1}{2}\mathbf{v}_j^T K(\mathbf{t}_i, \mathbf{t}_i)\mathbf{v}_j - \mathbf{v}_i^T K(\mathbf{t}_i, \mathbf{t}_j)\mathbf{v}_j.$$

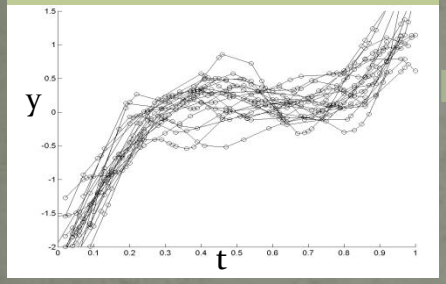
where

$$\mathbf{v}_i = (K(\mathbf{t}_i, \mathbf{t}_i) + \sigma^2 \mathbb{I})^{-1}(\mathbf{y}_i - \mathbf{f}_{0,i})$$

Learning GP through Non-parametric Mixed-effect Model

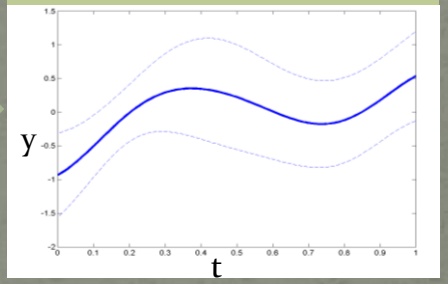
To answer Q II: How do we specify the Gaussian process.

Discrete Observations from Many Individuals

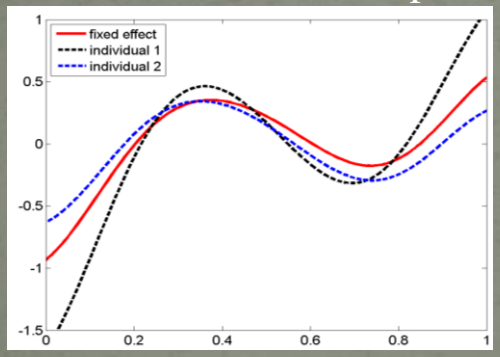


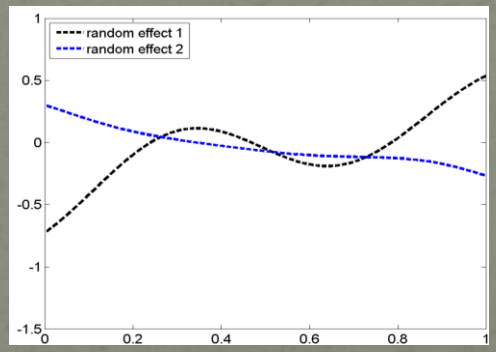
Learning with Non-parametric Mixed-effect Model

Gaussian Process



- We use a non-parametric mixed-effect model to learn the Gaussian process. Mixed-effect model describes a population of regression models by assuming every individual model consists of two pieces:





•The central piece is called fixed-effect

•The individual deviation is called random effect

- We get non-parametric mixed-effect models by using Gaussian process to model both fixed-effect and random effect

Non-parametric Mixed-effect Model

- **Generative Model**
  We assume the observations are generated by k smooth curves $\{f_1 \ f_2 \ \dots \ f_k\}$ fluctuating around a mean function f_0 (fixed-effect). We use

  $$\widetilde{f}_i = f_i - f_0$$

  to denote the deviation (random effect) of f_i from f_0, both effects are assumed zero-mean Gaussian processes:
  - Fixed effect:

    $$p_0[f_0] \propto \exp(-\frac{1}{2}||f_0||^2_{\mathcal{H}_0})$$

    The RKHS H_0 (or equivalently the kernel K_0) is predetermined, but f_0 is unknown
  - Random effect

    $$p_f[\widetilde{f}_i] \propto \exp(-\frac{1}{2}||\widetilde{f}_i||^2_{\mathcal{H}}) \quad i = 1, 2, \cdots, k$$

    Both f and H are unknown. Generally H is different from H_0
- **Observation Model**
  The discrete observations $\mathbf{y}_i$ are sampled from f_i with noise of unknown variance σ^2.

  $$y_{in} = f_i(t_{in}) + \epsilon_{in}, \quad n = 1, 2, \cdots, N_i$$
- **Parameter s** The unknown model parameters consist of

  $$\mathcal{M} = \{f_0, K, \sigma\}$$

Fitting Non-parametric Mixed-effect Model

- Our learning task is find *M* that maximizes the following probability

$$p(\mathbf{Y}|f_0; K, \sigma)p_0[f_0] = p_0[f_0]\prod_{i=1}^{k}\int \mathcal{D}f_i\{p(\mathbf{y}_i|\widetilde{f}_i, f_0; \sigma)p_f[\widetilde{f}_i]\}$$ (functional integral)

which (thanks to the Gaussian property) can be simplified to

$$p(\mathbf{Y}|f_0; K, \sigma)p_0[f_0] = p_0[f_0]\prod_{i=1}^{k}\int d\mathbf{f}_i\{p(\mathbf{y}_i|\mathbf{f}_i, f_0; \sigma)p(\mathbf{f}_i; K)\}$$ (standard integral)

- Non-parametric mixed-effect model can be fit using the EM-algorithm with $\{\mathbf{f}_1,\mathbf{f}_2,\ldots,\mathbf{f}_k\}$ as the latent variables

$$\begin{aligned}&\textbf{E-step} \quad Q(\mathcal{M}, \mathcal{M}^g) = E_{\{\mathbf{f}_i\,|\mathbf{Y};\mathcal{M}^g\}}[\log\{p(\mathbf{Y}, \{\mathbf{f}_i\}; \mathcal{M})p_0[f_0]\}]\\ &\textbf{M-step} \quad \mathcal{M}^* = \arg\max_{\mathcal{M}} Q(\mathcal{M}, \mathcal{M}^g),\end{aligned}$$

We have two different modeling choices for *K*

- Parametric

$K(t, t') = K(t, t'; \theta)$ e.g. RBF kernel or convex combination of known kernels

Appropriate for sparse observations or unsynchronized time series

- Non-parametric

$\mathbf{K} \equiv K(\mathbf{t}, \mathbf{t})$ covariance matrix evaluated on common observation times **t**

Good at fully exploiting the data, but works only on synchronized time series

More on the Optimization

- In each E-step we have

$$Q(\mathcal{M}, \mathcal{M}^g) = -\frac{1}{2}||f_0||^2_{\mathcal{H}_0} - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i} E_{\{\mathbf{f}_i|\mathbf{Y};\mathcal{M}^g\}}[(y_{ij} - \widetilde{f}_i(t_{ij}) - f_0(t_{ij}))^2] + \sum_{i=1}^{k}\int \mathrm{d}\mathbf{f}_i \log p(\mathbf{f}_i;\mathcal{M}) p(\mathbf{f}_i|\mathbf{y}_i;\mathcal{M}^g)$$

about f_0 and σ — about K

- In each M-step, we find a new K

$$\begin{aligned} K &= \arg\max_{K\in\mathcal{K}} \sum_{i=1}^{k}\int \mathrm{d}\mathbf{f}_i \log p(\mathbf{f}_i;K) p(\mathbf{f}_i|\mathbf{y}_i;K^g) \\ &= \arg\max_{K\in\mathcal{K}} -\sum_{i=1}^{k}\{\frac{1}{2}\log|K(\mathbf{t}_i,\mathbf{t}_i)| + \frac{1}{2}tr(K(\mathbf{t}_i,\mathbf{t}_i)^{-1}(\mathbf{C}_i^g + \mu_i^g(\mu_i^g)^T))\} \end{aligned}$$

where $\mu_i = K(\mathbf{t}_i,\mathbf{t}_i)(K(\mathbf{t}_i,\mathbf{t}_i) + \sigma^2\mathbb{I})^{-1}(\mathbf{y}_i - \mathbf{f}_{0,i})$

and $\mathbf{C}_i = K(\mathbf{t}_i,\mathbf{t}_i) - K(\mathbf{t}_i,\mathbf{t}_i)(K(\mathbf{t}_i,\mathbf{t}_i) + \sigma^2\mathbb{I})^{-1}K(\mathbf{t}_i,\mathbf{t}_i)$

- when we adopt a non-parametric *K*, we have closed form solution for **K**

$$\mathbf{K} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{C}_i^g + \mu_i^g(\mu_i^g)^T)$$

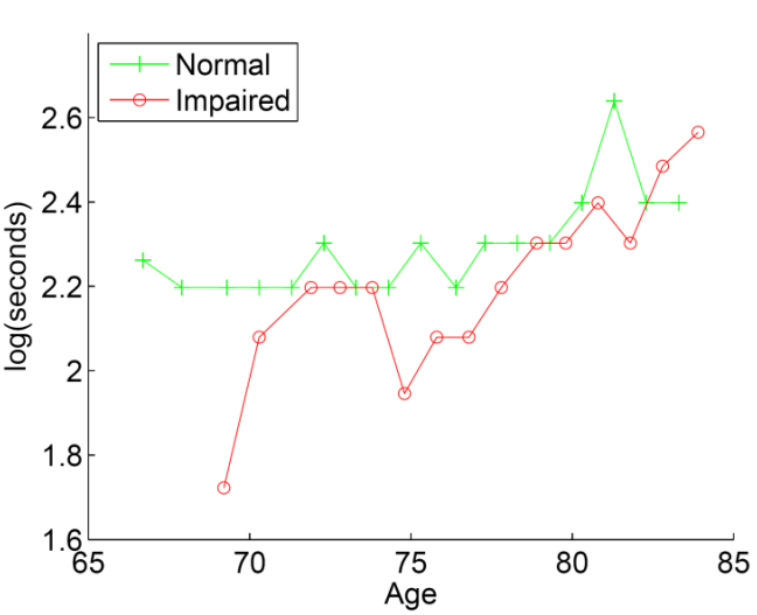
- when we adopt a parametric $K(t,t';\theta)$, we optimize over the parameter θ

$$\theta^* = \arg\max_{\theta} -\sum_{i=1}^{k}\{\frac{1}{2}\log|K(\mathbf{t}_i,\mathbf{t}_i;\theta)| + \frac{1}{2}tr(K(\mathbf{t}_i,\mathbf{t}_i;\theta)^{-1}(\mathbf{C}_i^g + \mu_i^g(\mu_i^g)^T))\}$$

Experiment (Cognitive Decline Detection I)

- We try to predict whether an aged person will decline into cognitive impairment based his/her longitudinal clinical records on motor ability.
- We considered four different motor tests:

seconds	# of seconds the subject takes to walk 9 m
steps	# of steps the subject takes to walk 9 m
tappingD	# of the tappings the subject does in 10 seconds with the dominant hand
tappingN	# of the tappings the subject does in 10 seconds with the non-dominant hand



•For each subject, the motor ability are measured with irregular intervals (usually 0.5~1 year)

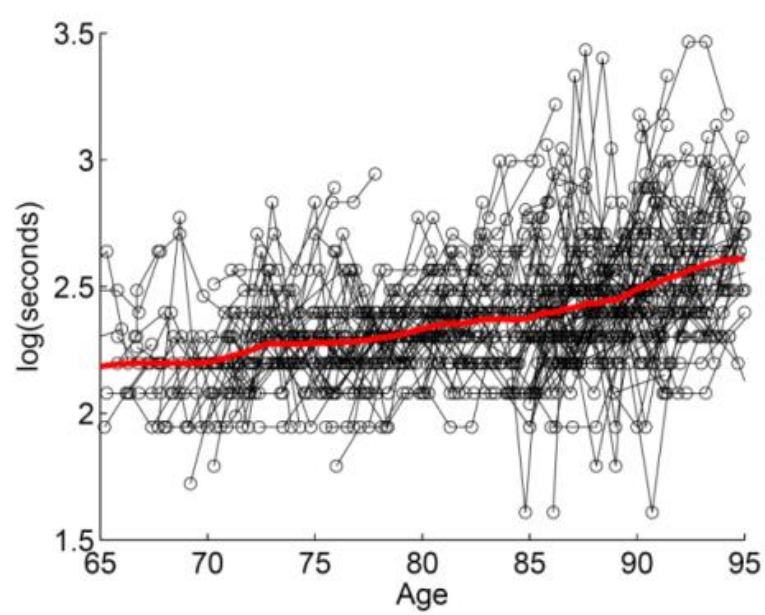
•Different subjects have their clinical visits on different schedules, with even different number of available tests.

•For people from impaired group, we use only the readings before a clinical diagnosis of dementia is reached.

Experiment (Cognitive Decline Detection I)

- We try to predict whether an aged person will decline into cognitive impairment based his/her longitudinal clinical records on motor ability.
- We considered four different motor tests:

seconds	# of seconds the subject takes to walk 9 m
steps	# of steps the subject takes to walk 9 m
tappingD	# of the tappings the subject does in 10 seconds with the dominant hand
tappingN	# of the tappings the subject does in 10 seconds with the non-dominant hand



•Both K_0 (the kernel for fixed effect f_0) and K (the kernel for the random effect) are parameterized

$$K_0(t_1,t_2) = \exp(\frac{||t_1-t_2||^2}{2s_0^2}),$$
$$K(t_1,t_2;\{a,s\}) = a\exp(\frac{||t_1-t_2||^2}{2s^2}),$$

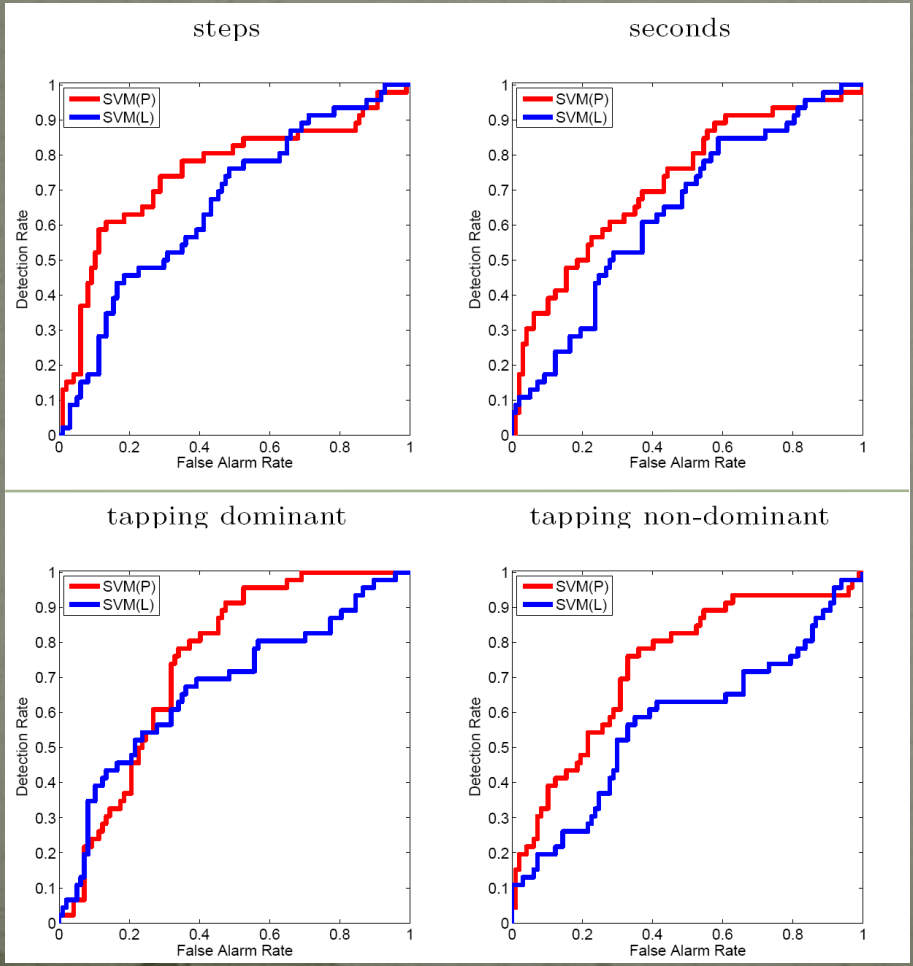
• Parameters to fit { f_0, a, s, σ }

• The fit fixed-effect (red curve) shows the general trend of deterioration of motor ability with age

Experiment (Cognitive Decline Detection II)

- We use SVM with the RBF kernel based on the proposed distance measure

$$\mathbf{G}_{ij} = \exp(-\frac{d_{ij}}{2r^2})$$



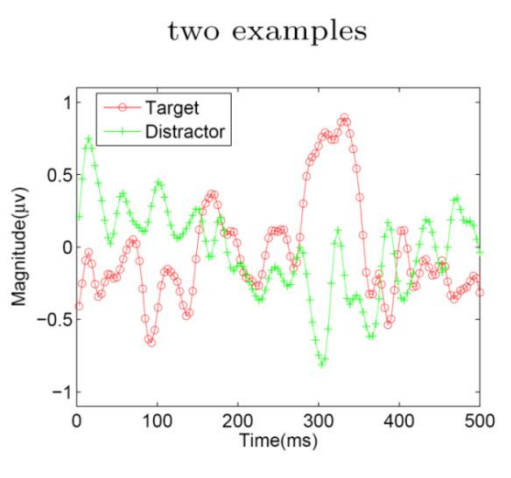
•We compared it with the SVM with the LSQ fit coefficients (polynomial) of individual time series as the feature vector

•We compare the ROC curve generated from the different classifiers.

•The ROC associated with the proposed distance measure (red) is obviously better than the one with LSQ feature (blue)

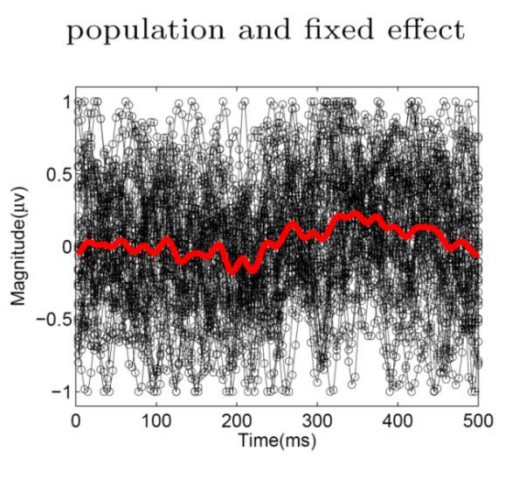
Experiment (EEG-based Image Target Detection)

- We examined the human expert's EEG signal to tell whether he has seen a target (e.g. golf course) in satellite images.



•After proper alignment and sampling, we get time series with 4128 synchronized observations.

•Previous research typically treat each time series as a vector and calculate the point-wise (Euclidean) distances.



• We directly fit **K** (*N*x*N* matrix) and **f** (*N*-dim vector) only evaluated on the observation times

•Experiments shows the proposed distance measure outperforms point-wise distance in the SVM classifier as well as linear classifier.

Summary

- Use smooth curve to represent time series (based on Gaussian process)
- Use the distance (derived from GP and Bregman divergence) between representing curves as the distance for corresponding time series
- Learn the Gaussian process

- Works well on classification of real world problems

Thank You