A Reproducing Kernel Hilbert Space Framework for Pairwise Time Series Distances

Zhengdong Lu ZHENGDON@CSEE.OGI.EDU
Todd K. Leen TLEEN@CSEE.OGI.EDU
Yonghong Huang HUANG@CSEE.OGI.EDU
Deniz Erdogmus DERDOGMUS@IEEE.ORG
OGI School, Oregon Health & Science University, 20000 NW Walker Rd., Beaverton, OR 97006 USA

Abstract

A good distance measure for time series needs to properly incorporate the temporal structure, and should be applicable to sequences with unequal lengths. In this paper, we propose a distance measure as a principled solution to the two requirements. Unlike the conventional feature vector representation, our approach represents each time series with a summarizing smooth curve in a reproducing kernel Hilbert space (RKHS), and therefore translate the distance between time series into distances between curves. Moreover we propose to learn the kernel of this RKHS from a population of time series with discrete observations using Gaussian process-based non-parametric mixed-effect models. Experiments on two vastly different real-world problems show that the proposed distance measure leads to improved classification accuracy over the conventional distance measures.

1. Introduction

Time series classification is a supervised learning problem aimed at labeling temporally structured sequences of variable length. The most common approach reduces time series classification to a static problem by suitably transforming the set of multivariate input sequences into vectors in Euclidean space. One can either summarize each time series with attributes pertinent to classification (called *feature extraction*)(Keogh & Pazzani, 1998), or use a properly sampled and aligned subsequence (called *sampling*)(Parra et al., 2003). Unfortunately, the feature extraction method is still more art than science, and the performance of the classifiers depends heavily on the designer's prior knowledge and the particular heuristic implemented. The sampling method, although preserving most of the information, is accused of ignoring the important temporal structure of the series. Indeed, the sampled sequences, if treated as vectors in Euclidean space, leads to the same classifiers after any permutation of the vector entries. Moreover, the sampling strategy does not apply to situations where we have only sparse observations that are made at different times.

In this paper, we propose a principled non-parametric distance measure for time series by representing each time series with a smooth curve in a reproducing kernel Hilbert space (RKHS) with a kernel learnt from data. This new distance measure circumvents the limitations of the two above mentioned strategies.

Paper Roadmap In Section 2, we give the background of the Bregman divergence, and then generalize it to function space for a proper distance measure of smooth curves. In Section 3 we propose a family of new distance measures for time series with only discrete observations. Section 4 is devoted to the non-parametric mixed-effect model, which helps to further specify the proposed distance measure. In Section 5, we apply the proposed distance measure to two real-world time series classification problems. Finally we discuss the related work in Section 6.

2. Gaussian Processes and Functional Bregman Divergence

The Bregman divergence is a natural generalization of squared Euclidean distance and KL-divergence. A Bregman divergence corresponding to a strictly convex function $\phi(x)$ (called seed function) is defined as

$$d_\phi(x_1||x_2) = \phi(x_1) - \phi(x_2) - \langle \nabla\phi(x_2), x_1 - x_2 \rangle \,. \quad (1)$$

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

Bregman divergence is closely connected to exponential family (Banerjee et al., 2005). For any distribution in exponential family

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

we know that the log likelihood can be re-written as

$$\log p(x;\theta) = -d_\phi(x||\mu(\theta)) + \phi(x) + \log p_0(x), \quad (2)$$

where ϕ is the conjugate function of Φ

$$\phi(x) = \sup_\theta\{\langle x, \theta\rangle - \Phi(\theta)\} \quad (3)$$

and $\mu(\theta) = \nabla\Phi(\theta)$ is the expectation parameter corresponding to θ. We go one step further to argue that $d_\phi(x_1||x_2)$ should be a proper model-weighted divergence measure between any x_1 and x_2. It is straightforward to show that for mulit-variate Gaussian distribution $\mathcal{N}(a, \Sigma)$, the corresponding Bregman divergence is given by

$$d_\phi(x_1||x_2) = \frac{1}{2}(x_1 - x_2)^T\Sigma^{-1}(x_1 - x_2), \quad (4)$$

which is also suggested in (Tipping, 1999) as a model-weighed distance for Gaussian distribution.

2.1. Extension to Function Space

We will generalize our discussion on the Bregman divergence and the exponential family to function space. To facilitate our discussion, we adopt the language of functional integral, which, although allegedly not rigorously defined, provides a powerful technique for describing the probability on functions (Simon, 1979).

Gaussian processes (GPs) (Rasmussen & Williams, 2006) generalize the multivariate Gaussian distribution to function space, which model any function f with the following probability [1]

$$p[f] \propto \exp(-\frac{1}{2}||f - f_0||^2_{\mathcal{H}}), \quad (5)$$

with f_0 being the *mean function* and $||\cdot||_{\mathcal{H}}$ the norm for the reproducing kernel Hilbert space $\mathcal{H}$. We use K to denote the corresponding reproducing kernel, which will also be noted as the covariance function for the Gaussian process expressed in Eq.(5) (Seeger, 2004). In regularization theory, the norm $||\cdot||_{\mathcal{H}}$ is often related to a particular type of smoothness of function, with large (even infinite) $||f||_{\mathcal{H}}$ for non-smooth function f.

After generalizing Eq.(1) to the functional case (Frigyik et al., 2006), we get the Bregman divergence between function f_1 and f_2, with a seed functional $g[\cdot]$

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

where $\nabla g[f_2]$ is the Fréchet derivative. The Gaussian process expressed in Eq.(5) can be viewed as a member of the exponential family extended to distributions on functions (Altun et al., 2004). Then a direct generalization of Eq.(3) leads to $g[f] = \frac{1}{2}||f||^2_{\mathcal{H}}$, which gives a GP-related divergence for smooth functions

$$d_{\mathcal{H}}(f_1||f_2) = \frac{1}{2}||f_1 - f_2||^2_{\mathcal{H}}. \quad (6)$$

3. Distance for Time Series

We consider k time series, using $\mathbf{y}_i$ to denote the N_i observations from the i^{th} time series made at times $\mathbf{t}_i$

$$\mathbf{y}_i \doteq [y_{i1}, \cdots, y_{iN_i}]^T, \quad \mathbf{t}_i \doteq [t_{i1}, \cdots, t_{iN_i}]^T.$$

The subscript i on $\mathbf{t}_i$ and N_i indicates that the observation times and even the number of observations are generally different for each individual. The time series are called *synchronized* if all the $\mathbf{t}_i$ are the same.

We can define a distance measure for such time series by associating the observations $\{\mathbf{t}_i, \mathbf{y}_i\}$ with a (smooth) curve. We assume the observations for each individual i is generated from a independent Gaussian process f_i with the same covariance function K (and therefore $\mathcal{H}$) and mean f_0. The observation is modeled as

$$y_{in} = f_i(t_{in}) + \epsilon_{in}, \quad n = 1, 2, \cdots, N_i, \quad (7)$$

where ϵ_{in} is a white observation noise with standard deviation σ for all i and n.

We choose to summarize each individual time series i with the expectation of $f_i(t)$ given the discrete noisy observation $\{\mathbf{t}_i, \mathbf{y}_i\}$.

$$\begin{aligned}\hat{f}_i(t) &= E[f_i(t)|\mathbf{y}_i, f_0; \mathbf{t}_i, K] & (8)\\ &= f_0 + K(t, \mathbf{t}_i)(K(\mathbf{t}_i, \mathbf{t}_i) + \sigma^2\mathbb{I})^{-1}(\mathbf{y}_i - \mathbf{f}_{0,i}) & (9)\end{aligned}$$

where $\mathbf{f}_{0,i} \doteq [f_0(t_{i1}), f_0(t_{i2}), \cdots, f_0(t_{iN_i})]^T$ is the values of f_0 at times $\mathbf{t}_i$, and $K(\mathbf{t}_i, \mathbf{t}_i)$ is the $N_i \times N_i$ matrix with the (n, m) entry being $K(t_{in}, t_{im})$. With a smooth f_0, we have $||\hat{f}_i||_{\mathcal{H}} < +\infty$, which can be loosely interpreted as that $\hat{f}_i$ is smooth according to K. In Fig.1, we give an example of using such a curve $\hat{f}$ to represent the noisy observations (black crosses).

We then use the distance between $\hat{f}_i$ and $\hat{f}_j$ as the distance between time series $\{\mathbf{t}_i, \mathbf{y}_i\}$ and $\{\mathbf{t}_j, \mathbf{y}_j\}$ [2],

[1] In the remainder of the paper, we use the square brackets [ ] to distinguish functionals from common functions.

[2] Although $E[||f_i - f_j||^2_{\mathcal{H}}|\mathbf{y}_i, \mathbf{y}_j; \mathbf{t}_i, \mathbf{t}_j]$ seems to be a reasonable measure of distance, it goes to infinity since with probability one a sample f from the a Gaussian process with covariance function K has $||f||_{\mathcal{H}} = \infty$ (Seeger, 2004).

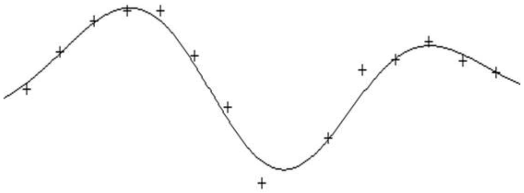

Figure 1. Using smooth curve to represent noisy discrete observations (black crosses). The smooth curve is obtained using Eq.(9) with K being a Gaussian kernel and $f_0 = 0$.

which is given by Eq.(6) as

$$d_{ij} = \frac{1}{2}||\hat{f}_i - \hat{f}_j||^2_{\mathcal{H}}. \tag{10}$$

Since $\mathcal{H}$ is the RKHS induced by the kernel K, this distance measure is well-defined

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

where $\mathbf{v}_i = (K(\mathbf{t}_i,\mathbf{t}_i) + \sigma^2\mathbb{I})^{-1}(\mathbf{y}_i - \mathbf{f}_{0,i})$. Using the reproducing kernel property

$$\forall\, t_n, t_m \quad \langle K(t_n,t), K(t_m,t)\rangle_{\mathcal{H}} = K(t_n,t_m),$$

the distance measurement can be simplified as

$$d_{ij} = \frac{1}{2}\mathbf{v}_i^T K(\mathbf{t}_i,\mathbf{t}_i)\mathbf{v}_i + \frac{1}{2}\mathbf{v}_j^T K(\mathbf{t}_i,\mathbf{t}_i)\mathbf{v}_j - \mathbf{v}_i^T K(\mathbf{t}_i,\mathbf{t}_j)\mathbf{v}_j. \tag{11}$$

It is important to note that this distance does not require all the time series to be synchronized, and is thus desirable when the sequences are of different lengths and/or the observations are made at different times, as shown in our first experiment in Section 5. When the observations for all individuals are synchronized, we have $\mathbf{t}_i = \mathbf{t} = [t_1, t_2, \cdots, t_N]^T$ with N as the total number of observations for each individual. Letting $\mathbf{K} = K(\mathbf{t},\mathbf{t})$, we can re-write d_{ij} as

$$\begin{aligned} d_{ij} &= \mathbf{v}_i^T\mathbf{K}\mathbf{v}_i + \mathbf{v}_j^T\mathbf{K}\mathbf{v}_j - 2\mathbf{v}_i^T\mathbf{K}\mathbf{v}_j & (12)\\ &= (\mathbf{v}_i - \mathbf{v}_j)^T\mathbf{K}(\mathbf{v}_i - \mathbf{v}_j) & (13)\\ &= (\mathbf{y}_i - \mathbf{y}_j)^T(\mathbf{K} + \sigma^2\mathbb{I})^{-1}\mathbf{K}(\mathbf{K} + \sigma^2\mathbb{I})^{-1}(\mathbf{y}_i - \mathbf{y}_j). & (14) \end{aligned}$$

Temporal Structure In Eq.(11)-(14), the temporal regularity is incorporated in the distance via the kernel K. It is most clear when we notice that K models the correlation of f value at different time

$$K(\mathbf{t}_i,\mathbf{t}_j) = E[(f(\mathbf{t}_i) - f_0(\mathbf{t}_i))^T(f(\mathbf{t}_j) - f_0(\mathbf{t}_j))].$$

The norm $||f_i - f_j||_{\mathcal{H}}$ measures the irregularity defined by K, in contrast to the Euclidean distance $\int(f_i(t) - f_j(t))^2 dt$ which only concerns about the point wise difference between f_i and f_j. It is also important to notice the particular temporal structure incorporated varies greatly with the choice of K imposes different type of temporal structure. For example, the widely used Matérn (including Gaussian) kernel or rational quadratic kernel promote different types and level of smoothness. On the other hand, the temporal structure is often problem specific and hard to determine beforehand. In the next section, we will discuss learning this temporal structure from the data.

4. Non-parametric Mixed-effect Model

In Section 3, we assume a Gaussian process with known mean and covariance function. However in practice it is often not the case. Instead we may want to learn the characteristic of Gaussian process from examples. One situation of interest to us is when a population of similar time series are available. This prior learning scheme is known in statistics as the empirical Bayesian or the hierarchial Beyesian (Gelman, 2004). Particularly, the model is called mixed-effect model when the hyper-prior is a Gaussian, on which the maximum likelihood (ML) solution can be found with Expectation-Maximization (EM) algorithm.

Traditional mixed-effect models are parametric, which assume a θ-parameterized regression model for each individual. Since the model parameters vary across individuals, it is natural to consider them generated by the sum of a fixed and a random piece $\theta = \alpha + \beta_i$, where α is called the *fixed effect*, and β_i, called *random effect*, is assumed distributed $\mathcal{N}(0,\mathbf{D})$ with unknown covariance $\mathbf{D}$. The fitting of mixed-effect model is to find α, $\mathbf{D}$, and the variance of observation noise.

In non-parametric mixed-effect models, the individual regression models do not take a parametric form. Instead, we assume the observations are generated by k smooth curves $\{f_1, f_2, \cdots, f_k\}$ fluctuating around a mean (fixed-effect) function f_0. We use $\widetilde{f}_i = f_i - f_0$ to denote the deviation of f_i from f_0 (random effect). The prior of both f_0 and $\widetilde{f}_i$ can be summarized with the following equations:

$$\begin{aligned} p_0[f_0] &\propto \exp(-\frac{1}{2}||f_0||^2_{\mathcal{H}_0}) & (15)\\ p_f[\widetilde{f}_i] &\propto \exp(-\frac{1}{2}||\widetilde{f}_i||^2_{\mathcal{H}}) \;\; i = 1,2,\cdots,k, & (16) \end{aligned}$$

where $\mathcal{H}$ and $\mathcal{H}_0$ are generally different Hilbert spaces, with the corresponding reproducing kernel denoted as K and K_0. Also we assume the observation noise to be white Gaussian with variance σ^2, from which follows

$$p(\mathbf{y}_i|\widetilde{f}_i, f_0; \mathbf{t}_i) \propto \prod_{j=1}^{n_i} \exp(-\frac{(y_{in} - \widetilde{f}_i(t_{in}) - f_0(t_{in}))^2}{2\sigma^2}).$$

We assume $\mathcal{H}_0$ (and thus the form of $p_0[\cdot]$) is predetermined, while the fixed effect f_0 is to be decided. Also unknown are the noise variance σ^2 and the Hilbert space $\mathcal{H}$ for random effects (or equivalently K). Our learning task is therefore to jointly optimize over $\{f_0, K, \sigma\}$ by maximizing the following probability of $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_k\}$.

$$p(\mathbf{Y}|f_0; K, \sigma)p_0[f_0] = \\ p_0[f_0] \prod_{i=1}^{k} \int \mathcal{D} f_i \{p(\mathbf{y}_i|\widetilde{f}_i, f_0; \sigma) p_f[\widetilde{f}_i]\}, \quad (17)$$

where the integral $\int \mathcal{D}\omega\, g[\omega]$ is a functional integral over ω (Simon, 1979). Using the Gaussian property, Eq.(17) can be further reduced to standard integration

$$p(\mathbf{Y}|f_0; K, \sigma)p_0[f_0] = \\ p_0[f_0] \prod_{i=1}^{k} \int d\mathbf{f}_i \{p(\mathbf{y}_i|\mathbf{f}_i, f_0; \sigma) p(\mathbf{f}_i; K)\}. \quad (18)$$

where $\mathbf{f}_i = [\widetilde{f}_i(t_{i1}), \widetilde{f}_i(t_{i2}), \cdots, \widetilde{f}_i(t_{iN_i})]^T$ collects the values of f_i on times $\mathbf{t}_i$ and $p(\mathbf{f}_i; K)$ is a standard multivariate Gaussian

$$p(\mathbf{f}_i; K) = \frac{1}{\sqrt{(2\pi)^{N_i}|K(\mathbf{t}_i, \mathbf{t}_i)|}} \exp(-\frac{1}{2}\mathbf{f}_i^T K(\mathbf{t}_i, \mathbf{t}_i)^{-1}\mathbf{f}_i). \quad (19)$$

In general, there is no unique solution of K that maximizes $p(\mathbf{Y}|f_0; K, \sigma)p_0[f_0]$. Indeed, it is easy to verify that if $K(t_{in}, t_{im}) = K'(t_{in}, t_{im})$ for any individual i and time index (n, m), we will have

$$p(\mathbf{Y}|f_0; K, \sigma)p_0[f_0] = p(\mathbf{Y}|f_0; K', \sigma)p_0[f_0].$$

This situation can be circumvented in two ways. First we can restrain K in a particular parametric family, such as the widely used Gaussian kernel. Second, we can instead optimize *only* over the entry $K(t_{in}, t_{im})$ for all individual i, and time index (n, m). Both strategies will be addressed in this paper.

4.1. Optimization with the EM Algorithm

The task is to find the set $\mathcal{M} = \{f_0, K, \sigma\}$ that maximizes the probability $p(\mathbf{Y}|f_0; K, \sigma)p_0[f_0]$. As shown in Eq.(18), we can rewrite the data likelihood $p(\mathbf{y}_i|f_0; K, \sigma)$ using the $\{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_k\}$ as the latent variables

$$p(\mathbf{y}_i|f_0; K, \sigma) = \int \mathrm{d}\mathbf{f}_i p(\mathbf{y}_i|\mathbf{f}_i, f_0, \sigma)p(\mathbf{f}_i; K), \quad (20)$$

which enables us to employ the EM algorithm in finding $\mathcal{M}$. In the following, we will give the results of the expectation step (E-step) and the maximization step (M-step).

E-step: In each EM iteration:

$$Q(\mathcal{M}, \mathcal{M}^g) = E_{\{\mathbf{f}_i\,|\mathbf{Y};\mathcal{M}^g\}}[\log\{p(\mathbf{Y}, \{\mathbf{f}_i\}; \mathcal{M})p_0[f_0]\}] \\ = \sum_{i=1}^{k} \int \mathrm{d}\mathbf{f}_i \log p(\mathbf{y}_i, \mathbf{f}_i; \mathcal{M})p(\mathbf{f}_i|\mathbf{y}_i; \mathcal{M}^g) + \log p[\mathbf{f}_0],$$

where $\mathcal{M}^g$ stands for the parameters from the last iteration. After some algebra, we can re-arrange $Q(\mathcal{M}, \mathcal{M}^g)$ into the following form

$$Q(\mathcal{M}, \mathcal{M}^g) = -\frac{1}{2}||f_0||^2_{\mathcal{H}_0} - n\log\sigma \\ - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i} E_{\{\mathbf{f}_i|\mathbf{Y};\mathcal{M}^g\}}[(y_{ij} - \widetilde{f}_i(t_{ij}) - f_0(t_{ij}))^2\,] \\ + \sum_{i=1}^{k} \int \mathrm{d}\mathbf{f}_i \log p(\mathbf{f}_i; \mathcal{M})p(\mathbf{f}_i|\mathbf{y}_i; \mathcal{M}^g). \quad (21)$$

M-step: In M-step, we find the

$$\mathcal{M}^* = \arg\max_{\mathcal{M}} Q(\mathcal{M}, \mathcal{M}^g), \quad (22)$$

and use $\mathcal{M}^*$ to update the model parameters. The optimization in Eq.(22) can be divided into two separate parts. The first three terms on the left hand side of Eq.(21) is a function of only (f_0, σ); The last (fourth) term is a function of only K. To find the solution of f_0 and σ, we need to solve the following optimization problem:

$$(\sigma^*, f_0^*) = \arg\min_{\sigma, f_0} \{\frac{1}{2}||f_0||^2_{\mathcal{H}_0} + N\log\sigma + \\ \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i} E_{\{\mathbf{f}_i|\mathbf{Y};\mathcal{M}^g\}}[(y_{ij} - \widetilde{f}_i(t_{ij}) - f_0(t_{ij}))^2]. \quad (23)$$

Particularly, with any fixed σ, maximizing $Q(\mathcal{M}, \mathcal{M}^g)$ over f_0 becomes a regularized regression problem

$$f_0^* = \arg\min_{f_0} \frac{1}{2}||f_0||^2_{\mathcal{H}_0} + \\ \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\{(y_{ij} - E_{\{\mathbf{f}_i|\mathbf{y}_i,\mathcal{M}^g\}}[\widetilde{f}_i(t_{ij})] - f_0(t_{ij}))^2\}$$

The optimization over K is

$$K = \arg\max_{K\in\mathcal{K}} \sum_{i=1}^{k} \int \mathrm{d}\mathbf{f}_i \log p(\mathbf{f}_i; K)p(\mathbf{f}_i|\mathbf{y}_i; K^g) \quad (24)$$

$$= \arg\max_{K\in\mathcal{K}} - \sum_{i=1}^{k}\{\frac{1}{2}\log|K(\mathbf{t}_i,\mathbf{t}_i)| + \frac{1}{2}tr(K(\mathbf{t}_i,\mathbf{t}_i)^{-1}(\mathbf{C}_i^g + \mu_i^g(\mu_i^g)^T))\} \quad (25)$$

where $\mathcal{K}$ is the set of feasible K, and μ_i is the posterior mean $E[\mathbf{f}_i|\mathbf{y}_i;\mathcal{M}]$ that can be calculated as

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

and $\mathbf{C}_i$ is the posterior covariance of $\mathbf{f}_i$

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

4.2. Parametric Covariance Estimation

We assume the covariance function K is of the parametric form $K(x,y;\theta)$. For example, the Gaussian kernel with scale a and kernel width s

$$K(x,y;\{a,s\}) = a\exp(-\frac{||x-y||^2}{2s^2}),$$

or as suggested in (Lanckriet et al., 2004) a convex combination of a set of kernels $\{K_1, K_2, \cdots, K_M\}$

$$K(x,y;\lambda) = \lambda_1 K_1(x,y) + \lambda_2 K_2(x,y) + \cdots + \lambda_M K_M(x,y).$$

In this case, the optimization of K in the M-step can be reduced to the following parameter estimation

$$\theta^* = \arg\max_{\theta} - \sum_{i=1}^{k}\{\frac{1}{2}\log|K(\mathbf{t}_i,\mathbf{t}_i;\theta)| + \frac{1}{2}tr(K(\mathbf{t}_i,\mathbf{t}_i;\theta)^{-1}(\mathbf{C}_i^g + \mu_i^g(\mu_i^g)^T))\} \quad (26)$$

where $p(\mathbf{f}_i;\theta) = p(\mathbf{f}_i; K(\mathbf{t}_i,\mathbf{t}_i;\theta))$. This parametric form of K is appealing in either one of the following two situations:

- when the observation are sparse, since the parametric K is generally less prone to overfitting compared to the non-parametric estimation, as will be discussed in Section 4.3.
- when the time series are not synchronized (as in Section 5.1) since the parametric K allows the out-of-sample extension.

4.3. Non-parametric Covariance Estimation

When all the time series all synchronized, we have $\mathbf{t}_i = \mathbf{t}, i = 1, 2, \cdots, k$. We can replace $K(\mathbf{t}_i,\mathbf{t}_i)$ in Eq.(25) with $\mathbf{K} \equiv K(\mathbf{t},\mathbf{t})$, and rewrite the optimization into the matrix form

$$\mathbf{K} = \arg\max_{\mathbf{K}\in\mathcal{P}} - \sum_{i=1}^{N}\{\frac{1}{2}\log|\mathbf{K}| + \frac{1}{2}tr(\mathbf{K}^{-1}(\mathbf{C}_i^g + \mu_i^g(\mu_i^g)^T))\}. \quad (27)$$

If we let $\mathcal{P}$ be the set of positive definite matrix, the solution of Eq.(27) is simple

$$\mathbf{K} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{C}_i^g + \mu_i^g(\mu_i^g)^T) \quad (28)$$

The non-parametric fitting of kernel matrix $\mathbf{K}$ is appealing since it does not assume a particular form for the covariance matrix and thus can fully exploit the information in the samples. However it can only be used when the time series are synchronized. One example of this modeling choice is given in Section 5.2.

5. Experiments

We tested the proposed distance measure on two real-world applications. The first one is an algorithm for cognitive decline detection based on longitudinal clinical observations of motor ability. The second one is an target identifier system based on electroencephalograph (EEG) signal.

In each experiment, we employ support vector machine (SVM) (Burges, 1998) with Gaussian kernel defined as follows

$$\mathbf{G}_{ij} = \exp(-\frac{d_{ij}}{2s^2}) \quad (29)$$

where d_{ij} is the squared distance between the time series i and j and the kernel width s is usually obtained using cross-validation. It is easy to see the $\mathbf{G}$ is a Mercer kernel.

5.1. Cognitive Decline Detection Based on Longitudinal Data

Research by our group and others show that motor changes, such as in walking and finger tapping rates, can effectively predict cognitive decline several years before impairment is manifest (Camicioli et al., 1998). It is highly useful to build a cognitive decline system (at least partially) based on the motor behaviors, since they can easily obtained via unintrusive in-home assessment. Our research focuses on using clinical motor behavior and data from the Oregon Brain Aging Study (OBAS) (Green et al., 2000). All 143 subjects in the cohort are healthy at entry, and when the data were drawn 46 of them had developed into mild cognitive impairment, while 97 remained cognitively healthy. We divide all the subjects into the impaired group and the normal group according to their state when the data were drawn from the database. [3] We intend

[3] This grouping is potentially inaccurate due to the possibility that those cognitively healthy subjects can later develop into dementia, which is known as right censoring in survival analysis.

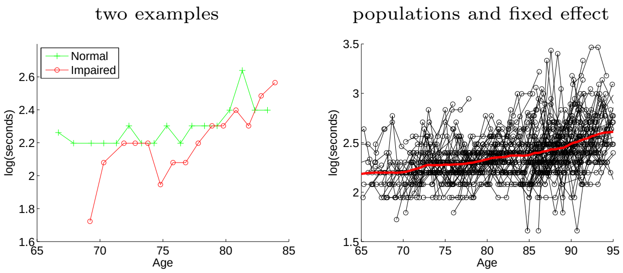

Figure 2. Left panel: sample spaghetti plots of seconds from two groups. Right panel: the population of seconds data and the fit fixed effect model (red line).

to predict whether a subject will develop into cognitive impairment based on his or her motor behavior before a clinical diagnosis (if any). In this experiment, this task reduces to predicting the group membership for each subject. This classification is difficult due to the fact that motor observations are sparse and noisy, as shown in Fig.2(left panel). We examined four motor behaviors summarized in Table 1. Usually as the subjects age or become impaired, the seconds and steps increase, while tappingD and tappingN decrease.

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

Table 1. Description of data.

We fit the non-parametric mixed-effect model to each motor behavior with the parameterized kernel

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

where s_0 is predetermined and $\{a, s\}$ are to be learnt. The right panel of Fig.2 shows the seconds time series from the 143 subjects (black $-\circ-$) and the fit fixed effect (red line). Once the model is fit, the distance between any two subjects i and j is calculated as in Eq.(11).

For comparison, we also examined a parametric feature based on the least-square (LSQ) fit coefficients for linear regression: $\mathbf{x}_i = \arg\min_{\mathbf{x}} \sum_{j=1}^{N_i}(x_0 + x_1 t_{ij} - y_{ij})^2$ with $\mathbf{x} = [x_0, x_1]^T$. This feature extraction is justified by the observation that the intercept and the slope of the motor behavior trajectory are predictor of future cognitive decline and dementia (Marquis et al., 2002). Based on the LSQ feature we get another distance measure $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$. We employ a SVM as the classifier with kernels calculated with Eq.(29). Fig.3 compares the ROC curves using the proposed distance measure and the Euclidean distance between the LSQ features. It is clear that SVM with proposed distance measure outperforms the SVM with the LSQ features in terms of the area under curve (AUC). There are two reasons for the superiority of proposed distance over the LSQ feature:

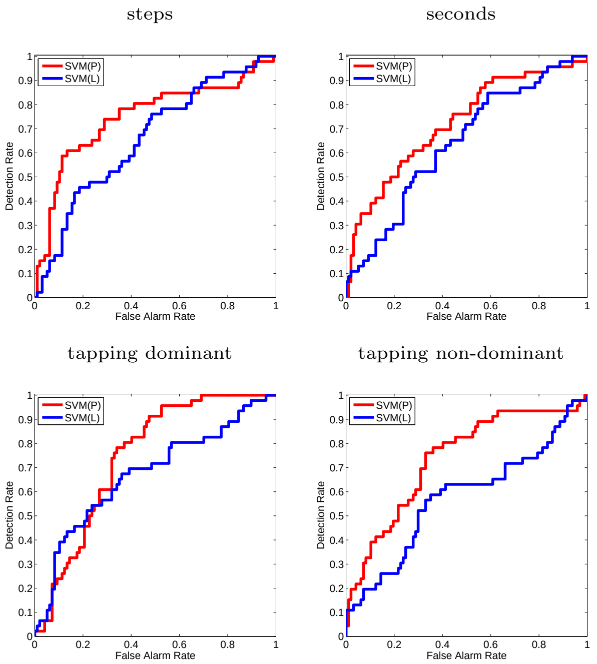

Figure 3. The ROC curve of SVM with two distance measures. SVM(P): SVM with proposed distance. SVM(L): SVM with least-square features.

- The simple heuristic features such as the intercept and the slope cannot capture enough information for the classification.
- The feature extraction is not robust enough for the sparse and noisy observations.

5.2. EEG-based Image Target Detection

The system reported here exploits the perceptual capabilities of expert humans for searching objects of interest (e.g., a golf course in a satellite image) within large image sets. The technique uses event related potentials (ERPs), neural signals linked to critical events, such as interesting/novel visual stimuli. The basic idea of the ERP-based image triage system is to collect electroencephalograph (EEG) signals from a subject's scalp when he performs visual target detection, and then detect the ERPs associated with target stimuli. We focus on the single-trial ERP detection using 32 EEG sensors, which is challenging due to the low signal-to-noise ratio.

This detection task is then boiled down to classifying the EEG segments into target-associated EPRs and distractors. After proper alignment and sampling, the EEG segments are transformed into synchronized sequence of length 4128, which are then time series, denoted $\mathbf{y}_i$ for each individual trial i. In this experiment, we collected the EEG data from three human experts, each of them performed 1 training session and 7 test sessions. In each training session, the human expert was fed with ∼600 images with ∼50 targets among them. In each test session, there are 1-4 targets within ∼3000 distractors. Fig.4 (left panel) shows single-trial EEG signals associated with a target and a distractor stimulus.

Due to the high dimensionality, the EM algorithm will be fairly slow due to the extensive use of inverse of $\mathbf{K}$ (4128×4128). To keep the computation at a reasonable level, we simplify the model by assigning a flat prior to the fixed effect f_0, or equivalently letting $||f||_{\mathcal{H}_0} = 0$ for any f. This simplifying assumption instantly leads to the following results.

- The optimal solution of $\mathbf{f}_0$ is simply the data mean $\mathbf{f}_0 = \frac{1}{k}\sum_{i=1}^{k} \mathbf{y}_i$, as shown in Fig.4 (right panel).
- The data likelihood is independent of σ^2 as long as it is less than the smallest eigenvalue of $\hat{\mathbf{K}} = \frac{1}{k}\sum_{i=1}^{k}(\mathbf{y}_i - \mathbf{f}_0)(\mathbf{y}_i - \mathbf{f}_0)^T$.

Based on the above two results, we can pick a σ and then calculate the optimal covariance $\mathbf{K}$ with Eq.(28) in one iteration.

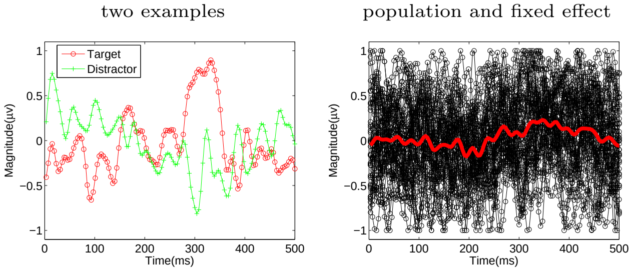

Figure 4. EEG data and the fit mixed-effect model. Left panel: Example of target-associated and distractor-associated EEG signals. Right panel: The population of EEG signals (black —○—) and the fit $\mathbf{f}_0$ (red curve).

Once the optimal $\mathbf{f}_0$ and $\mathbf{K}$ are obtained, the distance between any time series i and j can be calculated using Eq.(14). In addition to directly using the distance, we isometrically embed the time series $\{\mathbf{y}_i\}$ into Euclidean space while preserving the distance expressed in Eq.(14). The embedded vectors, called *ISO feature*, will then be used directly in linear classifiers. One obvious choice is the non-degenerated linear transformation

$$\mathbf{x}_i = \mathbf{K}^{1/2}(\mathbf{K} + \sigma^2 \mathbb{I})^{-1}\mathbf{y}_i \quad (30)$$

where $\mathbf{K}^{\frac{1}{2}}$ could be any matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ with $\mathbf{A}\mathbf{A}^T = \mathbf{K}$. We tested both the proposed distance and the (squared) Euclidean distance $||\mathbf{y}_i - \mathbf{y}_j||^2$ as the distance term d_{ij} in the Gaussian kernel $\mathbf{G}$ and compared the performance of the SVM with the two distance measures. In addition, we also tried a linear logistic classifier (LLC) with both the raw feature $\mathbf{y}_i$ and ISO feature $\mathbf{x}_i$ as the input. In our experiment, the SVM parameters and feature σ were selected using 10-fold cross validation.

Due to the extremely low probability of targets and the high cost of misdetection, we aim for a zero-miss and minimum false alarm rate (MFAR), which is defined as the percentage of false alarms among all classifications while all targets are correctly detected. We test both SVM and LLC on the 21 (=3 × 7) test sessions. Table 1 summarizes the detection results when different distance or features are used. The criteria of comparison include the average MFAR across the 21 sessions, the number of sessions with low MFAR ($\leq$ 10%)and very low MFAR ($\leq$ 2%). Clearly, the LLC with ISO features outperforms the LLC with raw feature by giving low average MFAR, more low MFAR sessions, and more very low MFAR sessions. The story is similar when using SVM as the classifier: the proposed distance outperforms the the Euclidean distance on all three criteria.

Clearly the temporal structure is important in describing the EEG signal, and thus plays a crucial role in deciding the distances between EEG time series. The proposed distance measure successfully incorporates the temporal structure information learnt with a rather simple algorithm, and yields significantly better classification than the Euclidean distance that simply adds the index-by-index differences.

	Aver. MFAR	# $\leq$ 2%	# $\leq$ 10%
LLC(I)	8.99%	12	16
LLC(R)	18.18%	2	12
SVM(P)	4.91%	13	19
SVM(E)	6.31%	7	16

Table 2. The detection results with different classifier settings. *Columns*: AverMFAR: the average MFAR across 21 sessions; #$\leq$ 2%: the number of sessions with MFAR$\leq$ 2%; #$\leq$ 10%:the number of sessions with MFAR$\leq$ 10%. *Rows:* LLC(I): LLC with the ISO feature; LLC(R): LLC with raw feature; SVM(P): SVM with the proposed distance; SVM(E): SVM with Euclidean distance;

6. Related Work

The connection between Bregman divergence and exponential family is first proposed by (Forster & Warmuth, 2000), and late used by several authors in deriving a proper distance measure for either clustering (Banerjee et al., 2005) or dimension reduction (Collins et al., 2001). our work also depends heavily on the functional Bregman divergence, an idea first fully explored in (Frigyik et al., 2006). The non-parametric mixed-effect model is a natural generalization to the hierarchical Bayesian Gaussian process proposed by (Schwaighofer et al., 2005) to functional form where synchronized and non-synchronized can be treated in a unified framework.

This work can be viewed as a particular example of the functional data analysis (Ramsay & Silverman, 1997). Particularly, in an early effort towards the functional PCA (Ramsay & Dalzell, 1991), the authors suggested to map the discrete observations $(\mathbf{t}_i, \mathbf{y}_i)$ to a smooth function through the following regularized regression

$$\hat{f}_i(t) = \arg\min_f \frac{1}{2}\sum_{n=1}^{N_i}(y_{in} - f(t_{in}))^2 + \frac{1}{2}\lambda||\mathcal{D}f||^2 \quad (31)$$

where $\mathcal{D}$ is a linear operator. The solution to Eq.(31) is the expectation in Eqn.(9) if we let $\lambda = \sigma^2$ and K be the Green's function of the operator $\mathcal{D}^*\mathcal{D}$. The difference, however, are that (1) our model also assumes a non-zero mean (fixed effect) f_0 and (2) the kernel K is learned from a population of time series.

Acknowledgments

ZL wants to thank Inderjit Dhillon, Arindam Banerjee, Jeff Kaye, and Misha Pavel for helpful discussions, and Robin Guariglia and Milar Moore for help with the OBAS data. TL, ZL, and DE were partially funded by the OHSU Behavioral and Intervention Commons project. YH is supported by NSF: ECS-0524835, ECS-0622239, and IIS-0713690.

References

Altun, Y., Hofmann, T., & Smola, A. (2004). Exponential families for conditional random fields. *Uncertainty in Artificial Intelligence(UAI).*

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with bregman divergences. *JMLR*, *6*, 1705–1749.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*, 121–167.

Camicioli, R., Howieson, D., Oken, B., Sexton, G., & Kaye, J. (1998). Motor slowing precedes cognitive impairment in the oldest old. *Neurology*, *50.*

Collins, M., Dasgupta, S., & Schapire, R. (2001). A generalization of principal component analysis to the exponential family. *NIPS 13.*

Forster, J., & Warmuth, M. K. (2000). Relative expected instantaneous loss bounds. *COLT 13.*

Frigyik, B., Srivastava, S., & Gupta, M. (2006). Functional bregman divergence and bayesian estimation of distributions. arXiv:cs/0611123.

Gelman, A. (2004). *Bayesian data analysis, second edition.* Chapman & Hall.

Green, M., Kaye, J., & Ball, M. (2000). The oregon brain aging study: Neuropathology accompanying healthy aging in the oldest old. *International Journal of Neural System*, *54*, 105–113.

Keogh, E., & Pazzani, M. (1998). An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *KDD 98* (pp. 239–241). ACM Press.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., & Jordan, M. (2004). Learning the kernel with semidefinite Programming. *Journal of Machine Learning Research*, *5*, 27–72.

Marquis, S., Moore, M., Howieson, D. B., Sexton, G., Payami, H., Kaye, J. A., & Camicioli, R. (2002). Indpendent predictors of cognitive decline in healthy elderly persons. *Arch. Neurol.*, *59*, 601–606.

Parra, L., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., & Sajda, P. (2003). Single trial detection in eeg and meg: Keeping it linear. *Neurocomputing*, *52-54*, 177–183.

Ramsay, J., & Silverman, B. (1997). *Functional data analysis.* Springer-Verlag.

Ramsay, J. O., & Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B (Methodological)*, *53*, 539–572.

Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning.* Cambridge, MA: MIT Press.

Schwaighofer, A., Tresp, V., & Yu, K. (2005). Learning gaussian process kernels via hierarchical bayes. *NIPS17.*

Seeger, M. (2004). Gaussian process for machine learning. *Internation Journal of Neural System*, *14*, 69–106.

Simon, B. (1979). *Functional integration and quantum physics.* Academic Press.

Tipping, M. (1999). Deriving cluster analytic distance functions from gaussian mixture models.