Electronic Journal of Statistics
Vol. 8 (2014) 1724–1733
ISSN: 1935-7524
DOI: 10.1214/14-EJS900B

Analysis of proteomics data: Phase amplitude separation using an extended Fisher-Rao metric*

J. Derek Tucker, Wei Wu and Anuj Srivastava

Department of Statistics
Florida State University
Tallahassee, FL, USA
e-mail: dtucker@stat.fsu.edu; wwu@stat.fsu.edu; anuj@stat.fsu.edu

Abstract: We consider the problem of alignment and classification of proteomics data, that is described in Koch et al. [4], using the Extended Fisher-Rao (EFR) framework introduced in [6]. We demonstrate this framework by separating amplitude and phase components of functional data from patients having therapeutic treatments for Acute Myeloid Leukemia (AML). Then, using individual functional principal component analysis, for both the phase and amplitude components [8], we obtain bases for principal subspaces and model the data by imposing probability models on principal coefficients. Lastly, using the distances calculated from individual components, we demonstrate a successful discrimination between responders and non-responders to treatment for AML.

Keywords and phrases: Amplitude variability, function principal component analysis, functional data analysis, phase variability.

Received August 2013.

1. Introduction

As described by Koch et al. [4], protein profiling can be used to study changes in protein expression in reference to therapeutic treatments for diseases. In this paper we analyze the protein profiles of five patients with Acute Myeloid Leukemia (AML) referred to in Koch et al. [4]. Specifically, we will develop tools for: (1) phase-amplitude separation from the given data, and (2) demonstrating metrics that can potentially assist in decision making and classification to study what different proteins are related to the disease process. The first step is performed by aligning the original functional data using nonlinear warping functions under an extended Fisher-Rao framework [6]. This results in the aligned functions (describing amplitude variability) and the warping functions (describing phase variability). Following the alignment, we utilize two metrics for data classification – one of them measures the amplitude variation and is independent of the phase components, and the other measures the phase difference while being independent of the amplitude components. With these metrics, one can also estimate the sample means and covariance on the phase and ampli-

[*] Main article 10.1214/14-EJS900.

tude components, respectively, in their appropriate spaces. While the amplitude space is a vector space and allows standard functional data analysis, the phase components need to be transformed using a square-root transformation to enable use of $\mathbb{L}^2$ norm (and cross-sectional computations) for generating summary statistics. These estimated statistics can be further used to perform functional principal component analysis (fPCA) and imposing probability models on the phase and amplitude components, respectively [8].

The rest of this paper is as follows. We briefly describe the separation, modeling, and comparison of the phase and amplitude components of any functional data in Section 2. This is followed by presentation of results on the proteomic data in Section 3, and the paper ends with the short conclusion in Section 4.

2. Approach: Extended Fisher-Rao framework

Our general framework for phase-amplitude separation and analysis of curves is adapted from ideas in shape analysis of curves [3, 7] and is described more comprehensively in [5, 6, 8]. For a broader introduction to this theory, including asymptotic results and identifiability results, we refer the reader to these papers.

Here we present a very brief review of the method used in our analysis. Let f be a real-valued function with the domain $[0, 1]$; any other domain can easily be transformed to this interval. For concreteness, only functions that are absolutely continuous on $[0, 1]$ will be considered; let $\mathcal{F}$ denote the set of all such functions. Also, let $\mathcal{H}$ be the set of boundary-preserving diffeomorphisms of the unit interval $[0, 1]$: $\mathcal{H} = \{h : [0, 1] \to [0, 1] |\ h(0) = 0,\ h(1) = 1, h \text{ is a diffeomorphism}\}$. The elements of $\mathcal{H}$ play the role of warping functions. For any $f \in \mathcal{F}$ and $h \in \mathcal{H}$, the composition $f \circ h$ denotes the time-warping of f by h. With the composition operation, the set $\mathcal{H}$ is a group with the identity element $h_{id}(t) = t$.

In our framework we represent a function using the *square-root slope function* (SRSF) and is defined as: $q(t) = \text{sign}(\dot{f}(t))\sqrt{|\dot{f}(t)|}$. For pairwise registration of functions we solve the optimization problem: $d_a(f_1, f_2) = \inf_{h \in H} \|q_1 - (q_2 \circ h)\sqrt{\dot{h}})\|$ using the dynamic programming algorithm. In the process, we evaluate d_a which measures their amplitude differences and is independent of their phases or time warpings. For aligning multiple functions, and for separating their phase-amplitude components, we first compute a Karcher mean of the given functions (denoted by μ_f in F and μ_q is in SRSF space), under the metric d_a.

$$\text{(In } \mathcal{F} \text{ space)} : \mu_f \quad = \quad \underset{f \in \mathcal{F}}{\text{argmin}} \sum_{i=1}^{n} d_a(f, f_i)^2 \tag{2.1}$$

$$\text{(In SRSF space)} : \mu_q \quad = \quad \underset{q \in \mathbb{L}^2}{\text{argmin}} \sum_{i=1}^{n} \left(\inf_{h_i \in H} \|q - (q_i, h_i)\|^2 \right). \tag{2.2}$$

(This Karcher mean has also been called by other names such as the Frechet mean, intrinsic mean or the centroid and is generalization of a Euclidean mean to metric spaces.) As described in [6], the algorithm for computing the Karcher mean also results in: (1) aligned functions $\tilde{f}_i = \{f_i \circ h_i\}$, representing the amplitude variability, and (2) the warping functions $\{h_i\}$ used in aligning the original data, representing the phase variability. For more details on this method the reader is referred to [5, 6, 8] or the companion paper [9].

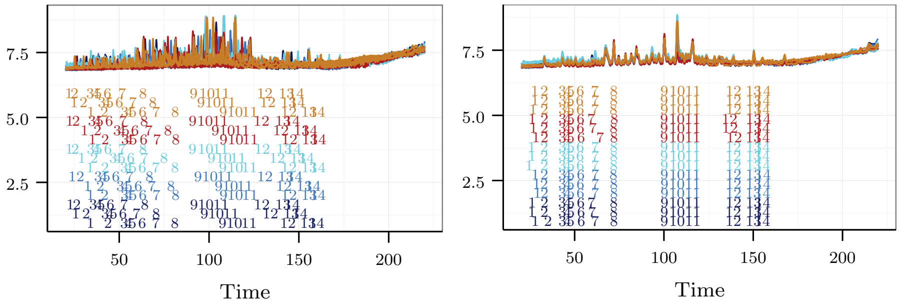
To quantify phase differences between given functions, we apply the same square-root transform to h and recognize that the set of all $\psi(t) = \sqrt{\dot{h}(t)}$ is an orthant of the unit Hilbert sphere $\mathbb{S}_\infty \subset \mathbb{L}^2$. Then, the distance between any two warping functions is exactly the arc-length between their representatives on the sphere $\mathbb{S}_\infty$. We can define a distance between a warping function and the identity function $\psi_{id}(t) = 1$ as $d_p(f_1, f_2) = \cos^{-1}(\int_0^1 \psi(t)dt)$. If h is the warping needed to align any two functions, then d_p measures the amount of warping needed to align them, and serves as a distance between their phases. One can then use these distances – d_a and d_p – for classification and further analysis. We also describe how to perform fPCA on the aligned functions (amplitude) and on the warping functions (phase) to study their variability, we will call this horizontal fPCA and veritcal fPCA, respectively. While fPCA on the aligned functions is straightforward (in the $\mathbb{L}^2$ space with its natural metric), the case of warping functions is not straightforward.

Horizontal PCA: As described in [8], we use the SRSF ψ to represent a warping function h, and since the unit Hilbert sphere is a non-linear manifold we choose a vector space tangent to the sphere for analysis; we call this the horizontal fPCA. The tangent space at any point $\psi \in \mathbb{S}_\infty$ is given by: $T_\psi(\mathbb{S}_\infty) = \{v \in \mathbb{L}^2 | \int_0^1 v(t)\psi(t)dt = 0\}$. In this tangent space we can define a sample covariance function: $(t_1, t_2) \mapsto \frac{1}{n-1}\sum_{i=1}^n v_i(t_1)v_i(t_2)$. In practice, this covariance is computed using a finite number of points, say T, on these functions and one obtains a $T \times T$ sample covariance matrix instead, denoted by K_ψ. The singular value decomposition (SVD) of $K_\psi = U_\psi \Sigma_\psi V_\psi^\mathsf{T}$ provides the estimated principal components of observed $\{\psi_i\}$: the principal directions $U_{\psi,j}$ and the observed principal coefficients $\langle v_i, U_{\psi,j}\rangle$. These components can be mapped back to $\mathcal{H}$ using the mapping $\psi \mapsto h(t) = \int_0^t \psi(s)^2 ds$.

Vertical PCA: To perform vertical fPCA on the aligned SRSFs we first add the initial value to form a larger vector: $g_i = [q_i \quad f_i(0)]$. This way, the mapping from the function space $\mathcal{F}$ to $\mathbb{L}^2 \times \mathbb{R}$ is a bijection. We can define a sample covariance operator for the aligned combined vector $\tilde{g} = [\tilde{q}_1 \quad f_i(0)]$ as $K_g = \frac{1}{n-1}\sum_{i=1}^n E[(\tilde{g}_i - \mu_g)(\tilde{g}_i - \mu_g)^\mathsf{T}] \in \mathbb{R}^{(T+1)\times(T+1)}$. Taking the SVD, $K_g = U_g \Sigma_g V_g^\mathsf{T}$ we can calculate the directions of principal variability in the given SRSFs using the first $p \leq n$ columns of U_g and can be converted back to the function space $\mathcal{F}$, via integration. This processes is called vertical fPCA and for more information on the two methods the reader is referred to [8].

3. Results on proteomics data

First, we will apply the extended Fisher-Rao framework to separate amplitude and phase components of the proteomics data. This requires the SRSF of the



(a) Original Data

(b) Aligned Data

Fig 1. *Alignment of proteomics data using the square-root slope framework with original data in Panel a) and aligned functions in Panel b). The aligned functions exhibit high level of registration of marked peaks.*

data. It should be noted that if the data is noisy some smoothing [1] can be applied before computing the derivative and SRSF. Second, we will perform fPCA on the separated amplitude and phase components, respectively. We will then construct models on the corresponding components and the models will be validated using random sampling. Lastly, we will perform classification between responders and non-responders to chemotherapy using the amplitude and phase distances, d_a and d_p, calculated during the alignment process.

3.1. Alignment

The original data with markers corresponding to the key peaks in the data is presented in Fig. 1(a). The peaks in the data are not well aligned as the corresponding markers demonstrate. The results of applying our alignment method are presented in Fig. 1(b). The aligned functions exhibit a high level of registration with almost all of the peaks are aligned. There are a few exceptions with peaks 1 and 2 which have a very low amplitude and that makes their registration difficult.

We can also quantify the alignment performance using the decrease in the cumulative cross-sectional variance of the aligned functions. For any functional dataset $\{g_i(t), i = 1, 2, \ldots, n, t \in [0, 1]\}$, let

$$\mathrm{Var}(\{g_i\}) = \frac{1}{n-1} \int_0^1 \sum_{i=1}^{n} \left(g_i(t) - \frac{1}{n} \sum_{i=1}^{n} g_i(t) \right)^2 dt,$$

denote the cumulative cross-sectional variance in the given data. For the proteomics data, we found:

$$\text{Original Variance} = \mathrm{Var}(\{f_i\}) = 4.05, \quad \text{Amplitude Variance} = \mathrm{Var}(\{\tilde{f}_i\}) = 1.13$$

$$\text{Phase Variance} = \mathrm{Var}(\{\mu_f \circ h_i\}) = 3.04.$$

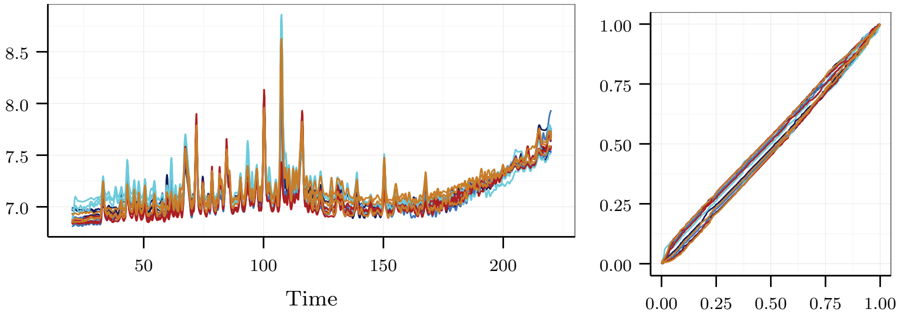


FIG 2. *The aligned proteomics data (left panel) with corresponding observed warping functions (right panel).*

where $\{f_i\}$ is the set of original functions, $\{\tilde{f}_i\}$ is the set of aligned functions, and $\{\mu_f \circ h_i\}$ is the set of applying the warping functions $\{h_i\}$ to μ_f, which is the mean function after alignment. From the decrease in the amplitude variance and increase in the phase variance we can quantify the level of alignment. Furthermore, the corresponding warping functions are presented in the right panel of Fig. 2.

Fig. 3 presents a zoom in on a region of the data from 81.4 to 111 time samples. The top panel is the original data where we see very poor alignment of the peaks. The bottom panel is the corresponding aligned data using the extended Fisher-Rao framework, where a nice alignment of the peaks and valleys have occurred.

3.2. Modeling

In this section, we present the results of applying horizontal and vertical fPCA on the warping and aligned functions, respectively. First, we perform horizontal fPCA on the warping functions and the first two principal directions are presented in Fig. 4(a) and (b), respectively. It can be see that most of the variation is contained in the first principal direction with minor perturbations contained in the next principal direction.

Next, we analyzed the aligned SRSFs by performing vertical fPCA. The first two vertical principal-geodesic paths are shown in Fig. 5 (a) and (b), respectively. The first 5 singular values for the data are: 3.89, 1.94, 1.49, 1.10, and 0.95 with the rest being negligibly small. Visually most of the variation lies in the first principal direction, which can also be attributed to the largeness of the first singular value relative to the other singular values

Once we have obtained the fPCA coefficients for the horizontal and vertical components we can then impose a probability model on the coefficients and induce a distribution on the function space $\mathcal{F}$. Using the joint Gaussian model described in [8] we randomly generate 35 domain-warping functions and 35

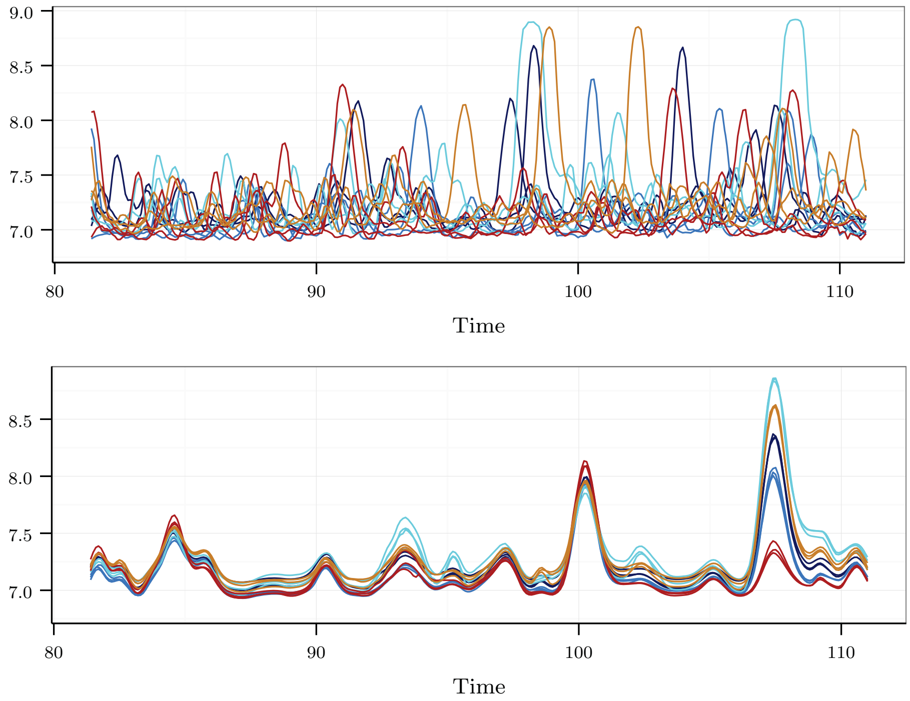
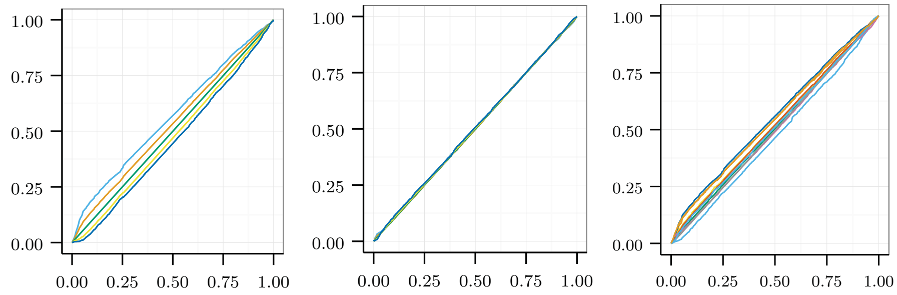


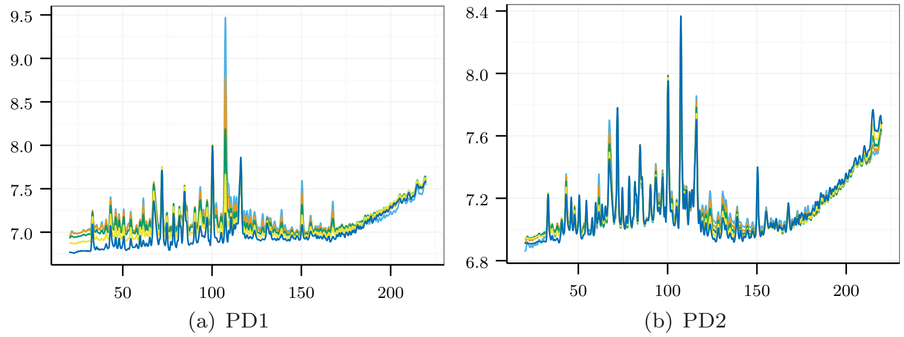
Fig 3. *Zoom in on original proteomics data (top panel) and aligned proteomics data (bottom panel). The original data peaks are varying in time locations while the aligned data has tight time alignment of all the peaks.*



(a) PD1 (b) PD2 (c) Random Samples

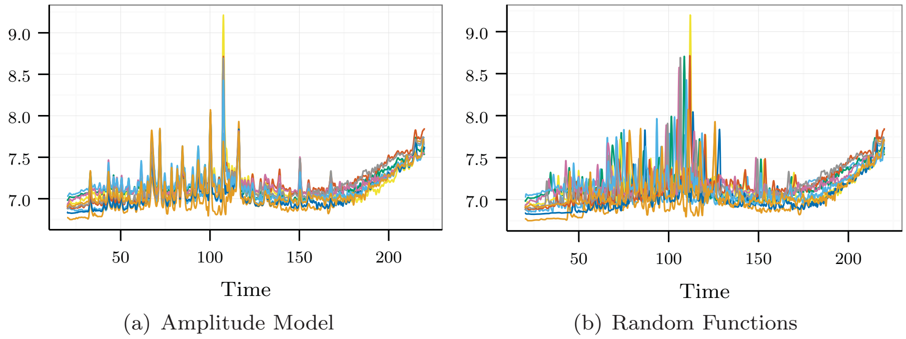
Fig 4. *The first two principal directions of the observed warping functions (a) and (b). Most of the variations is contained in the first principal direction. Corresponding random samples of warping functions from the phase model (c).*

amplitude functions. We then combine them using composition to generate a set of 35 random functions. The corresponding results are shown in Fig. 4(c) and Fig. 6. Which presents a set of random warping functions (Fig. 4(c)), with the



(a) PD1 (b) PD2

Fig 5. *The first two principal directions of the aligned functions, where most of the variation is contained.*



(a) Amplitude Model (b) Random Functions

Fig 6. *Random samples from from the amplitude model (a) and combination of the random samples with random samples from the phase model (Fig. 4(c)).*

corresponding amplitude functions (Fig. 6(a)), and the set of random samples (Fig. 6(b)). Comparing the random samples with the original data (Fig. 1(a)) we conclude that the samples are very similar to the original data and, at least under a visual inspection, the proposed models are successful in capturing the variability in the given data.

3.3. Classification

In this section, we present the results of using the amplitude distance, d_a, and phase distance, d_p, for classification between responders and non-responders to chemotherapy.

We first compute the standard $\mathbb{L}^2$ distance between each pair, i.e., $d^{ij}_{\mathbb{L}^2} = \|f_i - f_j\|,\ i, j = 1, \ldots, n$. The matrix of pairwise $\mathbb{L}^2$ distances is shown as a gray scale image in left panel in Fig. 7. This image of the pairwise distances looks unstructured, highlighting the difficulty of classification under this metric. Based on this distance matrix, we perform classification by using nearest

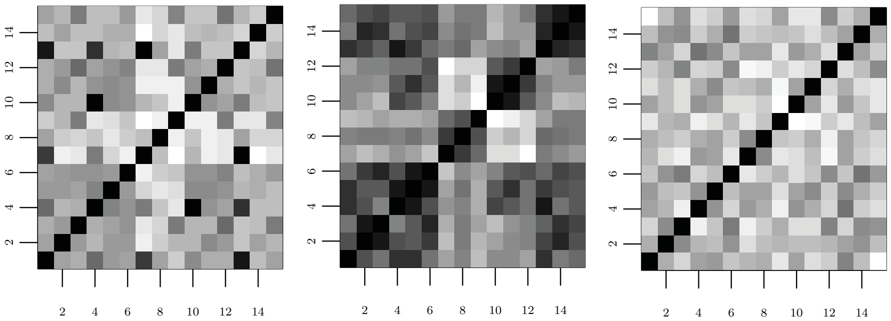


FIG 7. *The pairwise distances using the* $\mathbb{L}^2$, d_a, *and* d_p *distances, in the left, middle, and right panels, respectively.*

neighbor classifier. We then measure the error rate of this classifier using leave-one-out (LOO) cross-validation and found that the accuracy is 0.27 (4/15). Then, we computed distances d_a and d_p between all pairs of functions and these distance matrices are shown as gray scale images in the middle and right panels in Fig. 7, respectively. In the image of d_a (middle panel Fig. 7), we find that the pairwise distances are more structured than the $\mathbb{L}^2$ distances. We also perform classification using the LOO cross-validated nearest-neighbor based on the d_a distances. The accuracy turns out to be 0.87 (13/15), a significant improvement over the standard $\mathbb{L}^2$ result (0.27). We find that the d_p distances do not have strong classification performance, which can be noted by the lack of structure in the right panel of Fig. 7. The classification accuracy using d_p turns out to be 0.33 (5/15), which is only slightly higher than the standard $\mathbb{L}^2$ norm in the function space. The clear best performance of d_a is very consistent with the expectation that relative amounts of peptides should drive these differences, as stated in [4].

Since d_a and d_p each only partially describe the variability in the data, which corresponds to the phase and amplitude differences between the functions, there is a possibility of improvement if d_a and d_p are used jointly. One simple idea is to linearly combine these two distances and use the weighted distance to perform classification on the data. Define d_τ as the weighted average $d_\tau = \tau d_p + (1-\tau) d_a$, of d_a and d_p. We found an optimal of $\tau = 0.1$ provides a LOO accuracy of 0.93 (14/15), which is higher than the accuracy of the $\mathbb{L}^2, d_a$, and d_p distances. This indicates that there is some information carried in the phase than previously thought, however a larger number of samples would be required to justify this claim. Moreover, if there is more contribution from the phase it would suggest a batch effect or other nonrandom bias in the data that would have to be studied further.

For comparison, we computed a "naive" distance, d_{Naive}, which corresponds to the quantity $\min_h \|f_1 - f_2 \circ h\|$ that has often been used in the literature for function alignment. We also perform the cross-validated nearest-neighbor using

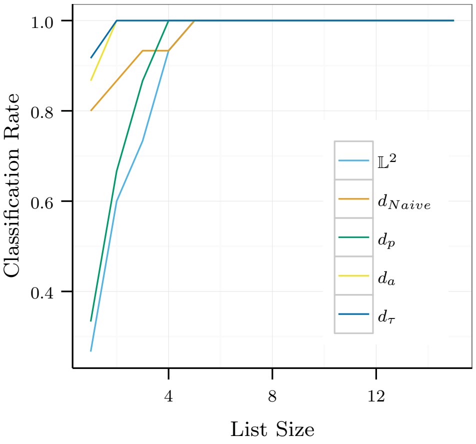


FIG 8. *CMC Comparison of* $\mathbb{L}^2$, D_{Naive}, D_x, D_y *and the weighted* D_τ $(\tau = 0.1)$ *distances.*

the d_{Naive} and find that the accuracy is 0.80 (12/15). This is better than the accuracy by d_p, but worse than that of d_a.

Next we generated a cumulative match characteristic (CMC) curve [2] for the distances $\mathbb{L}^2$, d_{Naive}, d_a, d_p, and d_τ. A CMC curve plots the probability of classification against the returned candidate list size (number of nearest neighbors) and is presented in Fig. 8. Both d_a and d_τ outperform all the other distances and very rapidly approach perfect classification for a small returned list size.

4. Conclusions

The statistical analysis and classification of functions is a challenging task, especially in the presence of phase variation. We have utilized a recent comprehensive approach that solves the problem of function alignment and analysis by using a cost function that is eventually a warping-invariant distance between the two functions. We have achieved a high level of alignment of the proteomics data using our alignment algorithm. We also studied data classification under different metrics and demonstrated a LOO performance of almost 0.90, which easily outperforms the standard $\mathbb{L}^2$ distance (0.27), and a method that uses a naive alignment (0.80).

Acknowledgments

The authors would like to thank the Mathematical Biosciences Institute for hosting the CTW: Statistics of Time Warpings and Phase Variations and Adelaide Proteomics Center for providing the data. This work was supported by NFS Grants NSF IIS-1217515 and NSF DMS-1208959.

References

[1] Bigot, J. and Gadat, S. (2010). Smoothing under diffeomorphic constraints with homeomorphic splines. *SIAM Journal on Numerical Analysis* **48** 777–785. MR2608367

[2] Bolle, R. M., Connell, J. H., Pankanti, S., Ratha, N. K. and Senior, A. W. (2005). The relation between the ROC curve and the CMC. In *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*, 15–20.

[3] Joshi, S. H., Klassen, E., Srivastava, A. and Jermyn, I. H. (2007). A novel representation for Riemannian analysis of elastic curves in $\mathbb{R}^n$. In *Proceedings of IEEE CVPR*, 1–7.

[4] Koch, I., Hoffmann, P. and Marron, J. S. (2014). Proteomics profiles from mass spectrometry. *Electronic Journal of Statistics* **8** 1703–1713, Special Section on Statistics of Time Warpings and Phase Variations.

[5] Kurtek, S., Srivastava, A. and Wu, W. (2011). Signal estimation under random time-warpings and nonlinear signal alignment. In *Proceedings of Advances in Neural Information Processing Systems (NIPS), Grenada, Spain*, 676–683.

[6] Srivastava, A., Wu, W., Kurtek, S., Klassen, E. and Marron, J. S. (2011a). Registration of functional data using Fisher-Rao metric. *arXiv:1103.3817v2 [math.ST]*.

[7] Srivastava, A., Klassen, E., Joshi, S. H. and Jermyn, I. H. (2011b). Shape analysis of elastic curves in Euclidean spaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **33** 1415–1428.

[8] Tucker, J. D., Wu, W. and Srivastava, A. (2013). Generative models for functional data using phase and amplitude separation. *Computational Statistics and Data Analysis* **61** 50–66. MR3063000

[9] Wu, W. and Srivastava, A. (2014). Analysis of spike train data: Alignment and comparisons using extended Fisher-Rao metric. *Electronic Journal of Statistics* **8** 1776–1785, Special Section on Statistics of Time Warpings and Phase Variations.