Alignment and Analysis of Proteomics Data using Square Root Slope Function Framework

J. Derek Tucker[1]

[1]Department of Statistics
Florida State University
Tallahassee, FL 32306

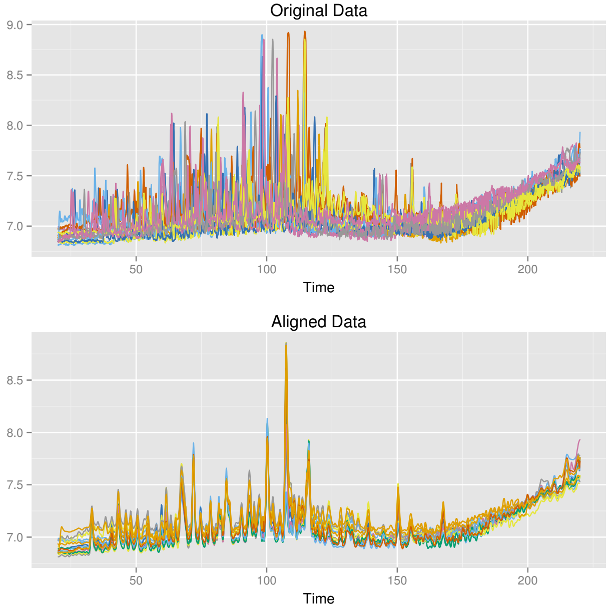
CTW: Statistics of Time Warpings and Phase Variations 2012

Problem Introduction

- Given: A collection of observed Total Icon Count (TIC) Chromoatograms
- Goals: We would like to
  - align the data
  - study their variability (FPCA)
  - develop probability models to capture their variability
  - generate random samples
- Requirement: Need a proper metric structure on the space of these functions
- Our Method: Propose phase and amplitude separation using elastic metric as presented by A. Srivastava

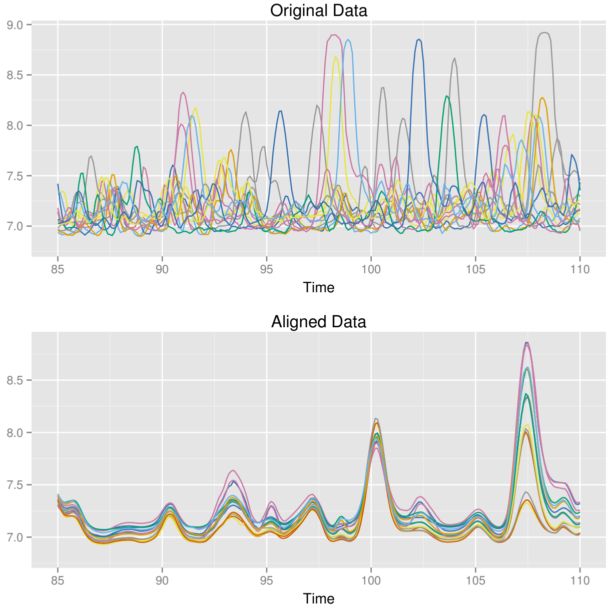
Results on Proteomics Data

- TIC (Total Ion Count) Chromatograms of blood samples. Used in protein profiling, assuming that proteins with different abundances are functionally related to disease processes.



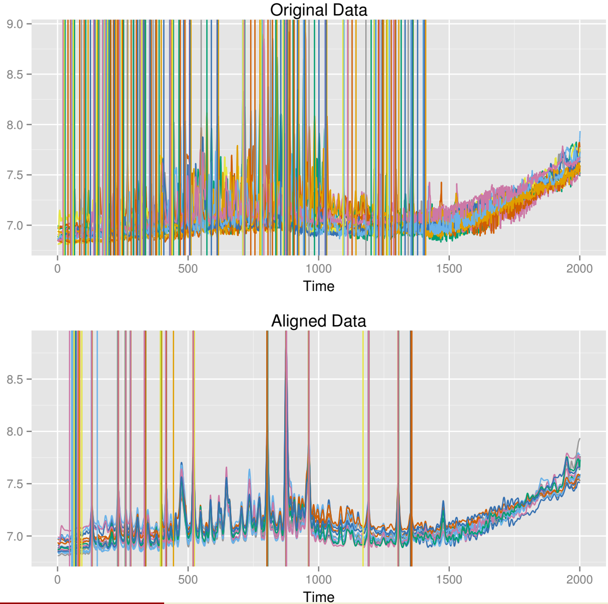
Zoom In on Alignment

- TIC (Total Ion Count) Chromatograms of blood samples. Used in protein profiling, assuming that proteins with different abundances are functionally related to disease processes.



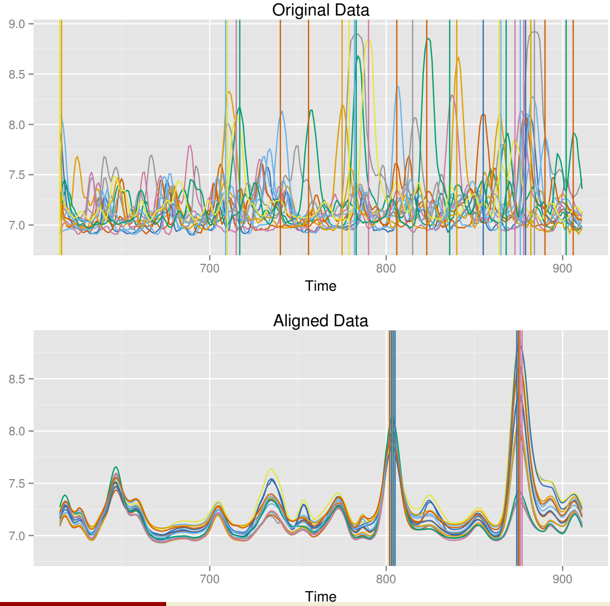
Results on Proteomics Data

- A partial "answer key" is available where several peaks have been manually identified, good alignment
- *was not used in alignment*



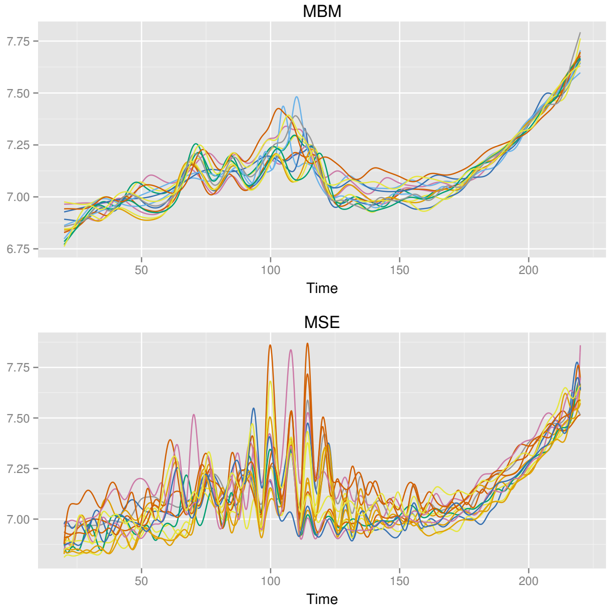
Zoom In on Alignment

- A partial "answer key" is available where several peaks have been manually identified, good alignment
- *was not used in alignment*



Comparison with Other Methods

- Comparison with MBM (James 2007) and MSE (Ramsay and Silverman 2005) methods



Alignment Performance

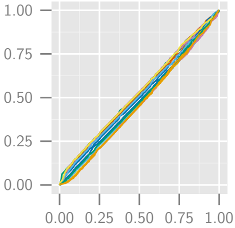
- Can also quantify the alignment performance using the decrease in the cumulative cross-sectional variance of the aligned functions

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

- Define: Original Variance $=$ Var$(\{f_i\})$, Amplitude Variance $=$ Var$(\{\tilde{f}_i\})$, Phase Variance $=$ Var$(\{\mu_f \circ \gamma_i\})$

	Original Variance	Elastic Method	MBM	MSE
Amplitude-variance	4.05	1.13	0.43	1.63
Phase-variance	0	3.04	0.80	0.51

Analysis of Warping Functions using Horizontal fPCA



- We have a collection of warping functions in the space Γ and we want to model their variability
- Γ is a nonlinear manifold and we cannot perform FPCA directly

- We choose to represent warping functions by their SRSFs as presented by A. Sravastava

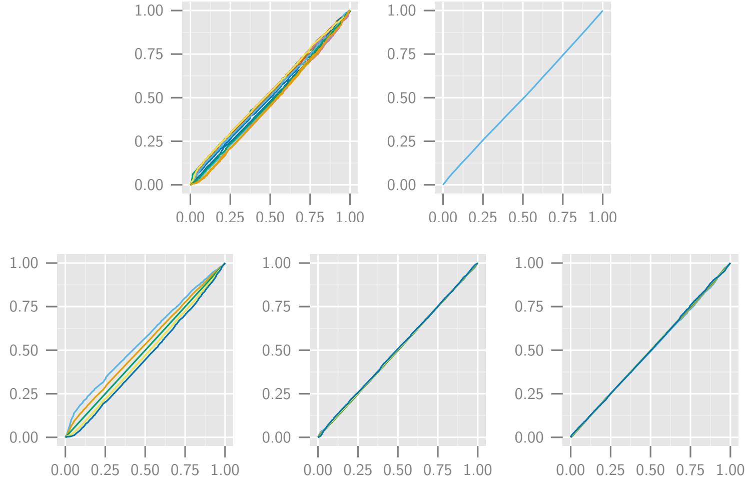
$$\psi(t) = \sqrt{\dot{\gamma}(t)}$$

- The $\mathbb{L}^2$ norm of this SRSF is:

$$\int_0^1 |\psi(t)|^2 \, dt = \int_0^1 \dot{\gamma}(t) \, dt = \gamma(1) - \gamma(0) = 1$$

- Hence, the space of such SRSFs is a unit Hilbert sphere in $\mathbb{L}^2$; call

Results on Proteomics Data



- From left to right: the observed warping functions, their Karcher mean, and the first three principal directions of the observed data.

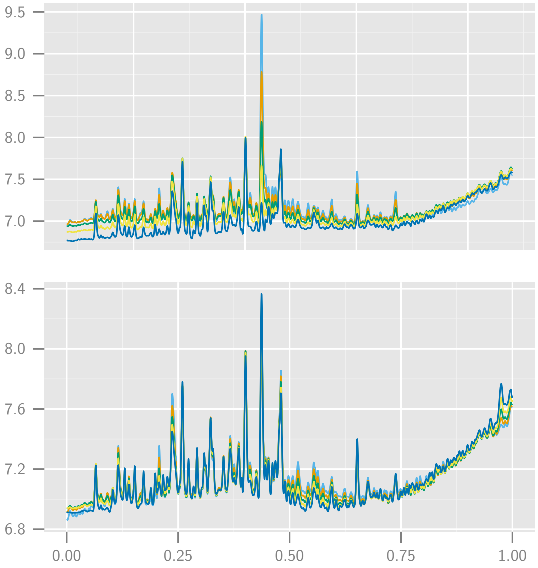
Analysis of Aligned Functions using Vertical fPCA

- The aligned can be statistically analyzed in a standard way (in $\mathbb{L}^2$) using cross-sectional computations in the SRSF space
- To properly calculate this we need to perform a joint FPCA which includes the vertical variability of $\mathcal{F}$
- a functional variable f_i is analyzed using the pair $h_i = (q_i, f_i(0))$ rather than just q_i
- Define covariance operator

$$K_h(s,t) = \frac{1}{n-1}\sum_{i=1}^{n} E[(\tilde{h}_i(s) - \mu_h(s))(\tilde{h}_i(t) - \mu_h(t))]$$

- where $\mu_h = [\mu_q \;\; \bar{f}(0)]$
- Taking the SVD, $K_h = U_h \Sigma_h V_h^{\mathsf{T}}$

Results on Proteomics Data



- First 2 vertical principal-geodesic paths
- Most of the information is captured in the first first few directions
- First 5 eigenvalues (3.89 1.94 1.49 1.10 0.95)

Modeling of Phase and Amplitude Components

- Let $c = (c_1, \ldots, c_{k_1})$ and $z = (z_1, \ldots, z_{k_2})$ be the dominant principal coefficients of the amplitude- and phase-components, respectively
- Recall that $c_j = \left\langle \tilde{h}, U_{h,j} \right\rangle$ and $z_j = \langle v, U_{\psi,j} \rangle$
- We can reconstruct the amplitude component using

$$q = \mu_q + \sum_{j=1}^{k_1} c_j U_{h,j}$$

- Similar for the phase component using $v = \sum_{j=1}^{k_2} z_j U_{\psi,j}$ and then using $\psi = \cos(\|v\|)\mu_\psi + \sin(\|v\|)\frac{v}{\|v\|}$, then

$$\gamma^s(t) = \int_0^t \psi(s)^2 ds$$

- Combining the two random quantities, we obtain a random function $f^s \circ \gamma^s$

Modeling Types

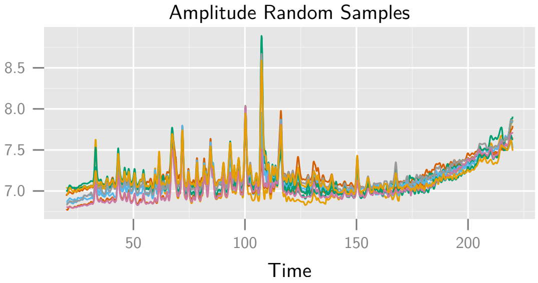
- Gaussian Models on fPCA Coefficients
  - Model $f^s(0)$, c, and z as multivariate normal random variables
  - The mean of $f^s(0)$ is $\bar{f}(0)$ while the means of c and z are zero vectors
  - Their joint covariance matrix is of the type:

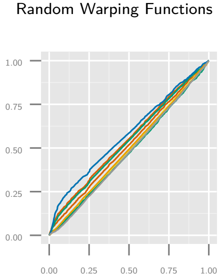
$$\begin{bmatrix} \sigma_0^2 & L_1 & L_2 \\ L_1^\intercal & \Sigma_h & S \\ L_2^\intercal & S & \Sigma_\psi \end{bmatrix} \in \mathbb{R}^{(k_1+k_2+1)\times(k_1+k_2+1)}$$

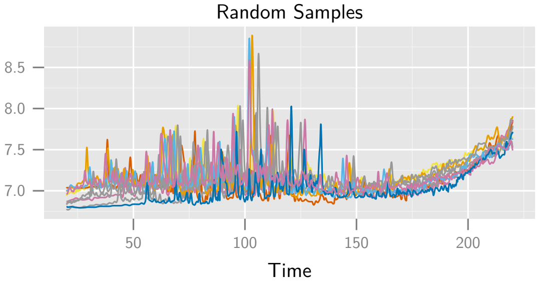
  - Here, $L_1 \in \mathbb{R}^{1\times k_1}$ captures the covariance between $f(0)$ and c, $L_2 \in \mathbb{R}^{1\times k_2}$ between $f(0)$ and z, and $S \in \mathbb{R}^{k_1\times k_2}$ between c and z
- Non-parametric Models on fPCA Coefficients
  - Use of kernel density estimation, where the density of $f^s(0)$, each of the k_1 components of c, and the k_2 components of z can be estimated using

$$p_{ker}(x) = \frac{1}{nb}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x - x_i}{b}\right)$$

Modeling Results

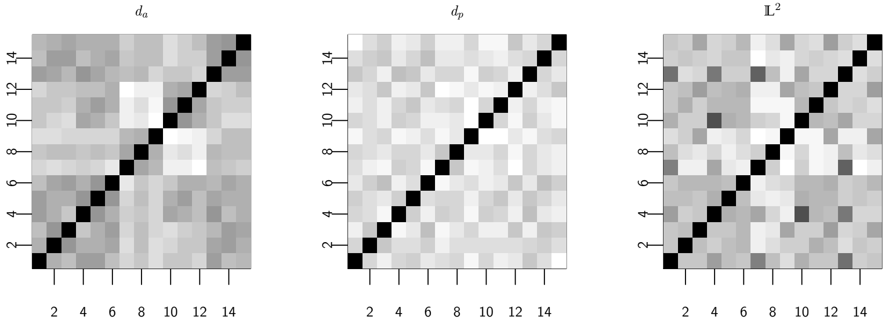






- Comparing them with the original data set we conclude that the random samples are similar to the original data

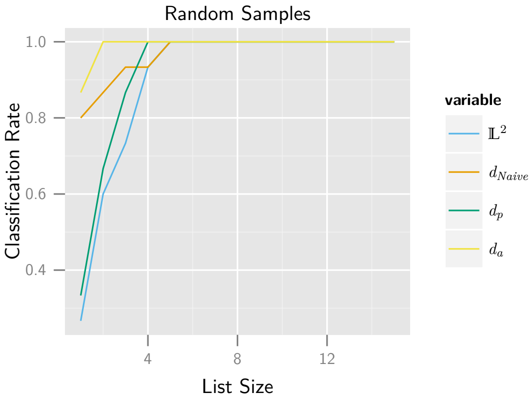
Classification using Pair-Wise Distances



- More structure to pairwise-distance matrices for d_a and d_p over standard $\mathbb{L}^2$
- Rates
  - $d_a = 87\%$ (13/15)
  - $d_p = 33\%$ (5/15)
  - $\mathbb{L}^2 = 27\%$ (4/15)

Cumulative Match Characteristic Curve

- A CMC curve plots the probability of classification against the returned candidate list size
- Also compared with a "naive" distance $d_{Naive} = \text{argmin}_{\gamma \in \Gamma} \|f_i - f_j \circ \gamma\|$



- Classification Performance of d_{Naive}: 80% (12/15)
- Our method rapidly approaches over 90% classification rate in contrast to the d_{Naive} and the standard $\mathbb{L}^2$ distances

Summary and Future Work

Conclusions

- Excellent alignment was achieved using our square-root slope function framework
- Used this framework to separate amplitude and phase of the given data
- Performed fPCA on amplitude and phase and imposed models on the components
- Verified the model using random sampling
- This theory behind this work has been submitted to Computational Statistics and Data Analysis 2012

Future Work

- Expand the analysis of classification to probabilistic models given we have more samples
- Analyze and under stand how additive noise impacts SRSFs and Karcher Mean calculation

Questions??