Multivariate and Functional Principal Components without Eigenanalysis

Jim Ramsay, McGill University

Workshop on Registration
Mathematical Biosciences Institute
Columbus, OH 13 November 2012

Outline

1. Overview
2. Beyond Eigenanalysis
3. Parameter Cascading
4. Parameter cascading for PCA
5. Parameter Cascading for Registered PCA

PCA: The essential idea

- We have a N by n data matrix $\mathbf{X}$.
- We propose the reduced rank K bilinear model

  $$\mathbf{X} = \mathbf{FA}$$

  where
  - $\mathbf{A}$ is a K by n matrix of principal component coefficients, with $K << n$
  - $\mathbf{F}$ is a N by K matrix of principal component scores
- Usually $N >> n$, and the factor scores are interesting, but it's $\mathbf{A}$ that tells us what the core K components of variation are, to within a full rank linear transformation.
- The fundamental goal of PCA is to identify a linear subspace $\mathcal{R}^{\mathcal{K}}$.

What I'd like to do with PCA

- Provide GLM capability: PCA for mixtures of types of variables, using fitting criteria appropriate to each data type.
- Define a fitting strategy that recognizes PC scores **F** as nuisance parameters and PC components in **A** as structural parameters.
- Generalize PCA:
  - synthesize the treatment of both multivariate and functional data
  - implement partial least squares: an approximation of an external vector **y** via a K dimensional subspace $\mathcal{R}^{\mathcal{K}}$
  - combine PCA with the registration of functional data

Outline

1. Overview
2. **Beyond Eigenanalysis**
3. Parameter Cascading
4. Parameter cascading for PCA
5. Parameter Cascading for Registered PCA

Eigenanalysis and PCA

- The singular value decomposition yields both **A** and **F**,
- But the usual procedure is to extract **A** from the eigenanalysis of $N^{-1}\mathbf{X}'\mathbf{X}$ or the correlation matrix **R**
- and then use regression analysis to obtain the least squares estimate

$$\mathbf{F} = \mathbf{X}\mathbf{A}'(\mathbf{A}'\mathbf{A})^{-1}$$

Why eigenanalysis gets in the way

- Eigenanalysis forces us to use least squares fitting for all variables.
- Eigenanalysis treats the estimation of **F** and **A** symmetrically, but **A** contains structural parameters and **F** contains nuisance parameters. They require different estimation strategies.
- Eigenalysis inappropriately highlights the basis system rather than the subspace that it defines.
- Eigenalysis cannot accommodate extensions such as registration of functional data.

Outline

1. Overview
2. Beyond Eigenanalysis
3. **Parameter Cascading**
4. Parameter cascading for PCA
5. Parameter Cascading for Registered PCA

Structural versus nuisance parameters

- Nuisance parameters are required in a model to capture important variation, but are seldom themselves of direct interest. A well-known example are random effect parameters in a mixed effects (ME) model.
- The number of nuisance parameters often depends on the configuration or design of the data.
- Structural parameters are typically of direct interest, for example fixed effect parameters for ME models.
- Their number is usually fixed, and typically much smaller than the number of nuisance parameters.
- Estimating nuisance and structural parameters using the same strategy risks burning up large number of degrees of freedom and rendering the structural parameter estimates unnecessarily unstable. ME model estimation recognizes this.

The parameter cascading strategy

- Parameter cascading is a method for estimating large and varying numbers of nuisance parameters $\mathbf{c}$ in the presence of a small fixed number of structural parameters $\boldsymbol{\theta}$.
- Parameter cascading defines nuisance parameters as *smooth* functions $\mathbf{c}(\boldsymbol{\theta})$ of structural parameters.
- Imposing smoothness or regularizing $\mathbf{c}(\boldsymbol{\theta})$ keeps nuisance parameters from burning up large numbers of degrees of freedom, and therefore stabilizes the structural parameter estimates.
- Nuisance parameter function $\mathbf{c}(\boldsymbol{\theta})$ is often defined by an inner optimization of a criterion $J(\mathbf{c}|\boldsymbol{\theta})$ each time $\boldsymbol{\theta}$ is changed in an outer optimization cycle.
- The outer optimization $H(\boldsymbol{\theta})$ is frequently different from $J(\mathbf{c}|\boldsymbol{\theta})$.

The parameter cascading strategy and the Implicit Function Theorem

- The total derivative or gradient of H with respect to θ requires the use of the Implicit Function Theorem:

$$\frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} - \frac{\partial H}{\partial \mathbf{c}} \left[\frac{\partial^2 J}{\partial^2 \mathbf{c}^2} \right]^{-1} \frac{\partial^2 J}{\partial \mathbf{c} \partial \theta}$$

- The total Hessian is also available in this way.

Outline

1. Overview
2. Beyond Eigenanalysis
3. Parameter Cascading
4. **Parameter cascading for PCA**
5. Parameter Cascading for Registered PCA

The parameter cascading strategy for multivariate PCA

- We add smoothness to the least squares criterion for **F** given **A** by attaching penalty terms:

$$J(\mathbf{F}|\mathbf{A},\mathbf{X}) = \|\mathbf{X} - \mathbf{F}\mathbf{A})\|^2 + \lambda_1\|\mathbf{F}'\mathbf{P}_1\mathbf{F}\|^2 + \lambda_2\|\mathbf{F}\mathbf{P}_2\mathbf{F}'\|^2.$$

- The minimizer $\hat{\mathbf{F}}(\mathbf{A})$ has a closed form expression.
- Order K matrix $\mathbf{P}_1$ and order N matrix $\mathbf{P}_2$ are often projectors onto complements of some pre-defined subspaces or special patterns.
- Smoothing parameters $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ allow us to control the emphasis that we place on the PC scores having these particular structures.

The fitting criterion for A

- This is defined in terms of only the PC coefficients **A.**
- Consequently, we can choose our fitting criteria freely, such as

$$H(\mathbf{A}) = -\sum_{j}^{n} \ln L_j(\mathbf{A}|\mathbf{x}_j)$$

where $-\ln L_j$ is the negative log likelihood appropriate to variable j and defined by data N-vector $\mathbf{x}_j$.

- The gradient of G will depend on **A** both directly through its the partial derivative, and also via the N functions $\mathbf{f}_i(\mathbf{A})$

$$\frac{dH}{d\mathbf{A}} = \frac{\partial H}{\partial \mathbf{A}} + \sum_{i}^{N} \frac{\partial H}{\partial F_i} \frac{dF_i}{d\mathbf{A}}$$

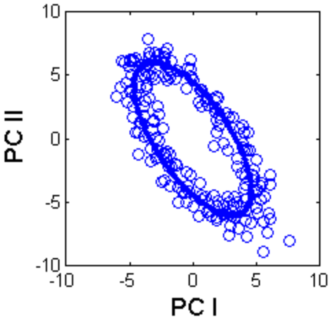
- PCA is now estimates Kn parameters instead of $K(N+n)$ parameters.

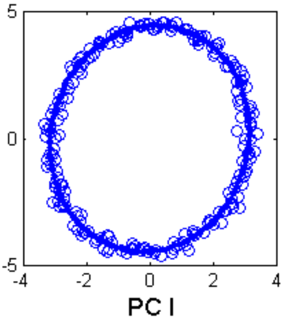
Evaluating the fit

- Without regularization, $\mathbf{A}$ and $\mathbf{F}$ are defined to within a nonsingular linear transformation $\mathbf{W}$ of order K: $\mathbf{FWW}^{-1}\mathbf{A}$ provides the same fit to the data.
- Regularization may remove some of this unidentifiability, but some will inevitably remain.
- Consequently, we cannot assess fit in term of $\mathbf{A}$, but must rather focus our attention on:
  - predictive criteria assessing fit at the data level
  - geometric measures of conformity between the K-dimensional estimated subspace and some true or population subspace.
- Canonical correlation methodology serves these purposes well.

A simulated data example

- $N = 200, n = 5, K = 2$, $\mathbf{F}$ contains 200 equally spaced points on a circle, $\mathbf{A} = [[1, 1, 1, 1, 1]; [-2, -1, 0, 1, 2]]$ $\mathbf{X} = \mathbf{FA}$ plus i.i.d. Gaussian error, $\sigma = 0.5$.
- The unregularized fit to the data yielded squared canonical correlations $\rho_1^2 = 0.99991$ and $\rho_2^2 = 0.99775$ between the true and estimated subspaces.
- The following figure compares unregularized and regularized estimates of $\mathbf{F}$ where the regularization penalized departure from the true values.


PC II
PC I
10
5
0
-5
-10
-10
-5
0
5
10


PC I
5
0
-5
-4
-2
0
2
4

The parameter cascading strategy for functional PCA

- The data are now N functions $x_i(t)$
- The principal coefficients are now functions $a_k(t), k = 1, \ldots, K$.
- The inner criterion J is now:

$$J(\mathbf{F}|\mathbf{a},\mathbf{x}) = \sum_i \int [x_i(t) - \sum_k f_{ik} a_k(t)]^2 dt + \lambda_1 \|\mathbf{F}'\mathbf{P}_1\mathbf{F}\|^2 + \lambda_2 \|\mathbf{F}\mathbf{P}_2\mathbf{F}'\|^2$$

- Structural parameter $\mathbf{A}$ is now a K by L matrix of coefficients for a basis function of each a_k in terms of L basis functions.

- The outer criterion could be

$$H(\mathbf{A}|\mathbf{x}) = \sum_i \int [x_i(t) - \sum_k f_{ik} a_k(t)]^2 dt + \lambda_3 \mathrm{trace}(\mathbf{AUA}')$$

where penalty matrix **U** defines a roughness penalty for the a_k's.

Outline

1. Overview
2. Beyond Eigenanalysis
3. Parameter Cascading
4. Parameter cascading for PCA
5. **Parameter Cascading for Registered PCA**

PCA and registration

- PCA is designed to decompose amplitude variation into a small number of components.
- Unfortunately, the presence of phase variation adds additional components to accommodate phase variation, and complicates the interpretation of amplitude variation components.
- By explicitly allowing for phase variation, we can avoid this problem.
- Let time warping function $h_i(t)$ be a smooth strictly increasing function of time specific to observation i.
- We modify the principal component function $a_k(t)$ by replacing t by $h_i(t)$, that is, we use $a_k[h_i(t)]$ to define the fit to the ith observed function $x_i(t)$.
- We define $h_i(t)$ in terms of of a basis function expansion with coefficients in vector $\mathbf{d}_i$: that is, $h_i(t|\mathbf{d}_i)$.

The inner optimization criterion

- It is known that the inner optimization criterion $J(\mathbf{D}|\mathbf{A})$ ought not to be a least squares measure.
- Instead, we minimize the smallest–eigenvalue–of–crossproduct criterion (SEofC)

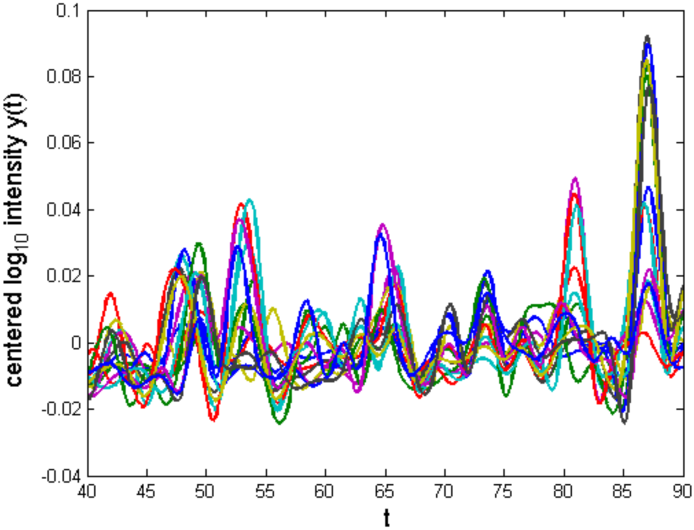
$$J(\mathbf{D}|\mathbf{A}) = \sum_{i}^{N} \mathtt{MINEIG} \begin{bmatrix} \int \{x_i(t)\}^2 \, dt & \int x_i(t)\, \hat{x}_i[h(t|\mathbf{d}_i)] \, dt \\ \int x_i(t)\, \hat{x}_i[h(t|\mathbf{d}_i)] \, dt & \int \{\hat{x}_i[h(t|\mathbf{d}_i)]\}^2 \, dt \end{bmatrix}$$

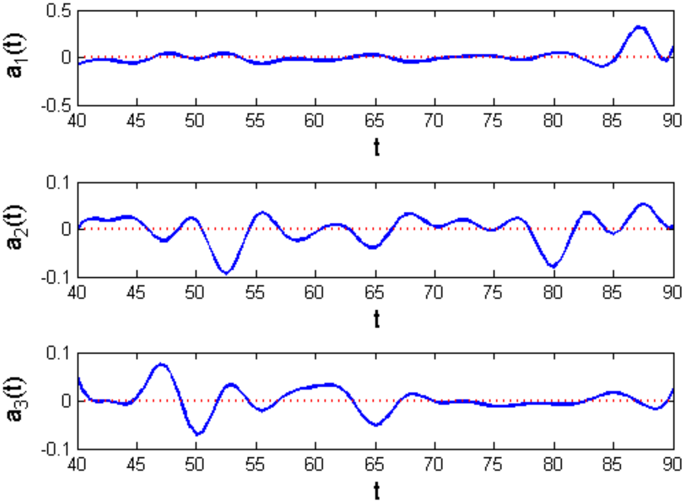
$$+\lambda_1 \mathbf{d}_i' \mathbf{V} \mathbf{d}_i .$$

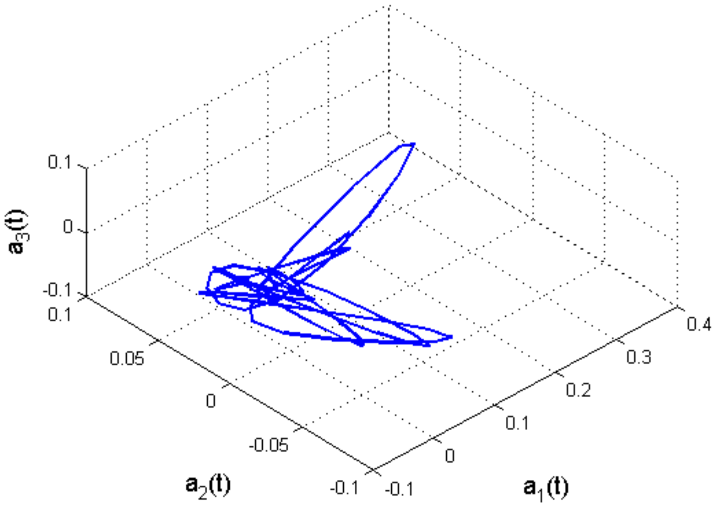
- The second term penalizes the curvature in the warping functions h_i.

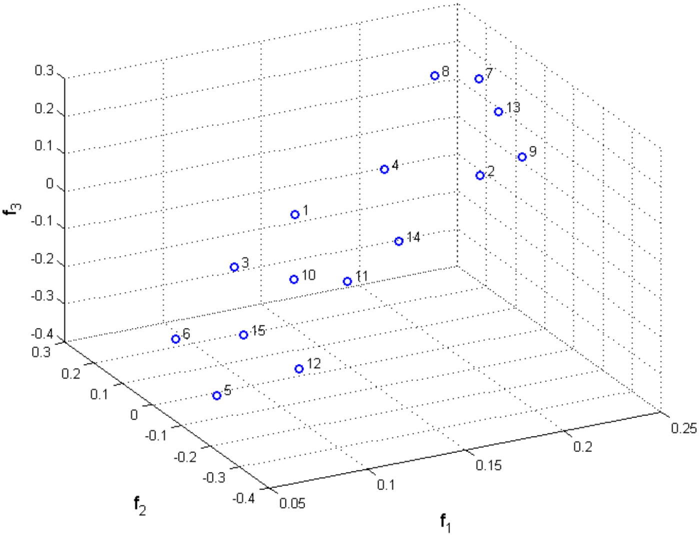
Registered PCA of the proteomics data

- In order to reduce computation and focus on the part of the showing most variation, I selected the data between 40 and 90 minutes, and used the centered log intensities.
- The data were pre-registered by using three landmarks along with a polygonal warping functions. A good deal of local phase variation remained in the data, however.
- The 510 observations were represented by a linear combination 512 B-spline basis functions fit by slight smoothing.
- Three principal components were chosen.
- The principal component coefficient functions a_k were represented using 23 B-spline basis functions.
- The warping functions h_i were represented using 13 B-spline basis functions with a penalty on the curvature of the log of their derivatives.

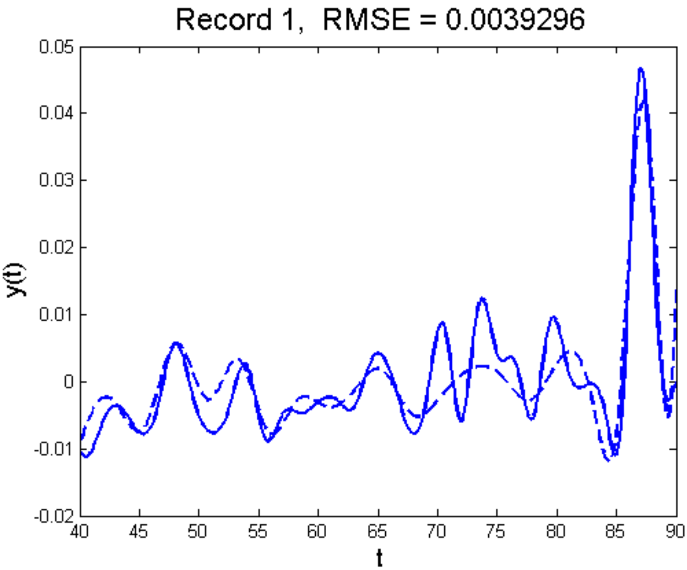

0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
40
45
50
55
60
65
70
75
80
85
90
t
centered $\log_{10}$ intensity y(t)

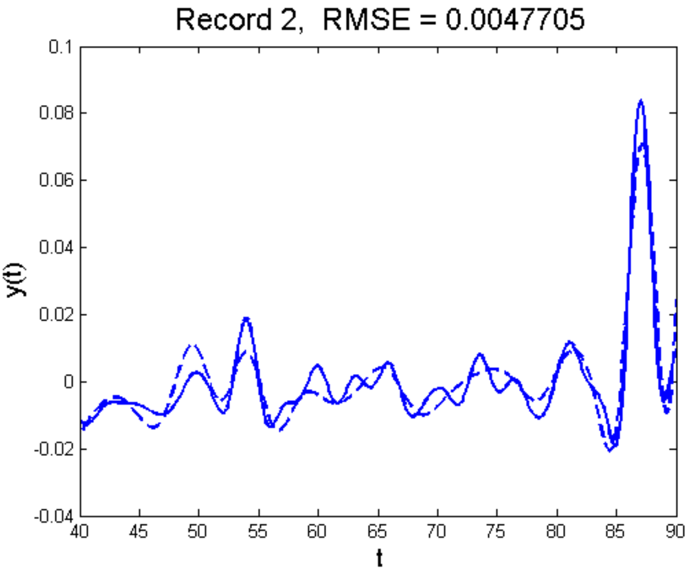

$a_1(t)$
0.5
0
-0.5
40 45 50 55 60 65 70 75 80 85 90
t
$a_2(t)$
0.1
0
-0.1
40 45 50 55 60 65 70 75 80 85 90
t
$a_3(t)$
0.1
0
-0.1
40 45 50 55 60 65 70 75 80 85 90
t

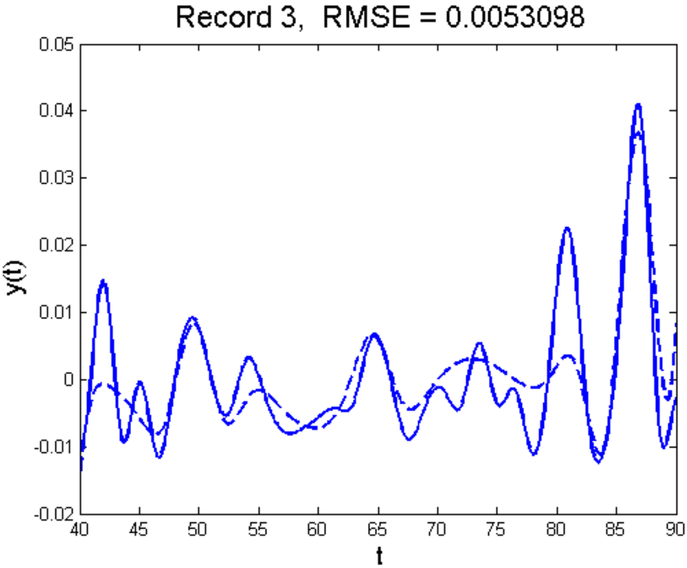

0.1
0
-0.1
$a_3(t)$
0.1
0.05
0
-0.05
-0.1
$a_2(t)$
-0.1
0
0.1
0.2
0.3
0.4
$a_1(t)$

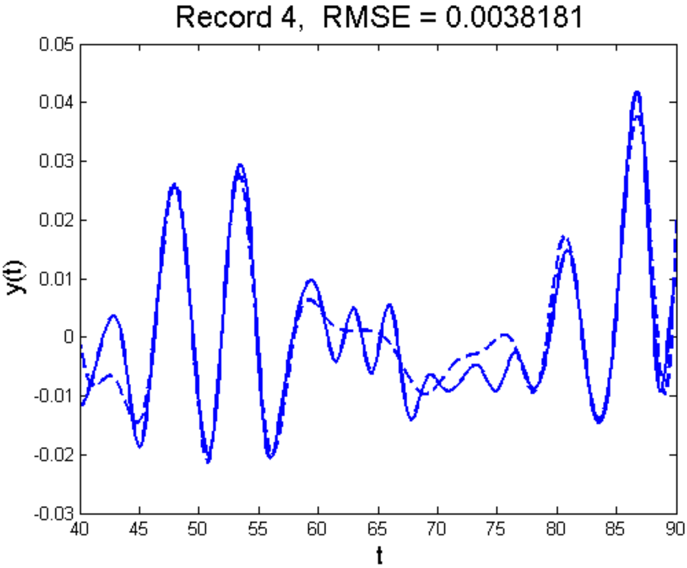


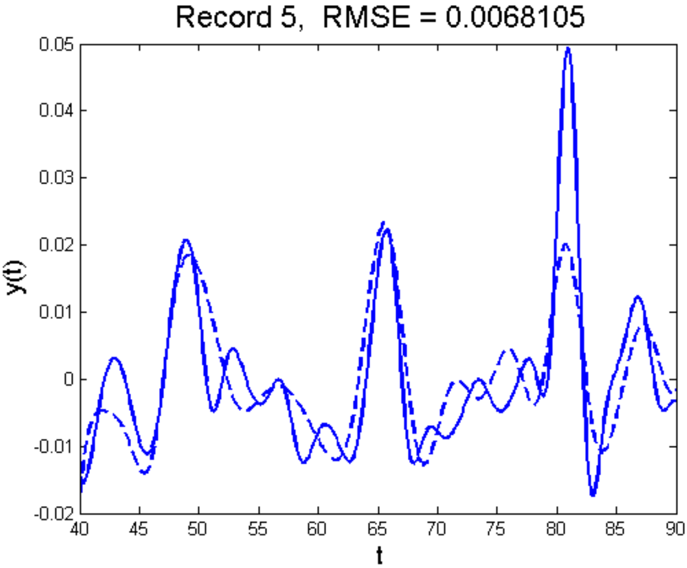
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
f_3
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
f_2
0.05
0.1
0.15
0.2
0.25
f_1
8
7
13
9
4
2
1
14
3
10
11
6
15
12
5

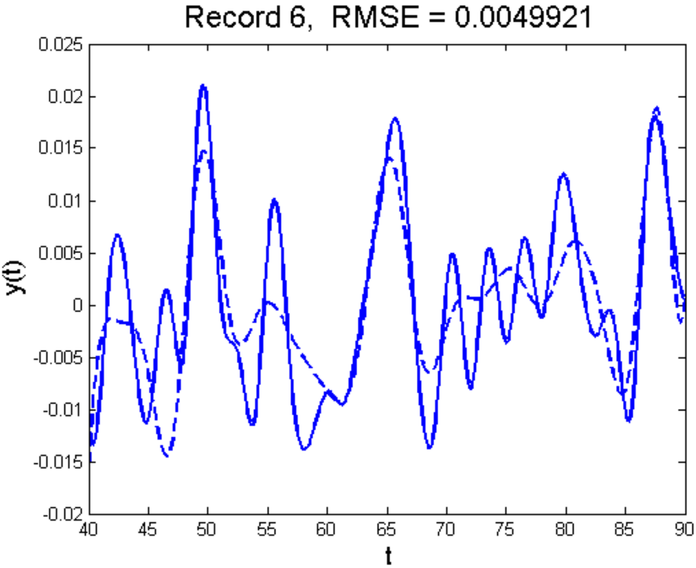

Record 1, RMSE = 0.0039296
y(t)
t
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
40
45
50
55
60
65
70
75
80
85
90

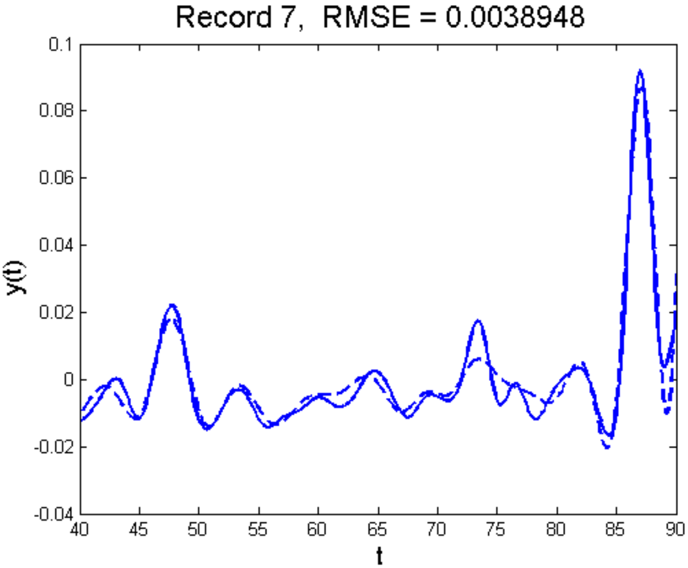

Record 2, RMSE = 0.0047705
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
y(t)
40
45
50
55
60
65
70
75
80
85
90
t

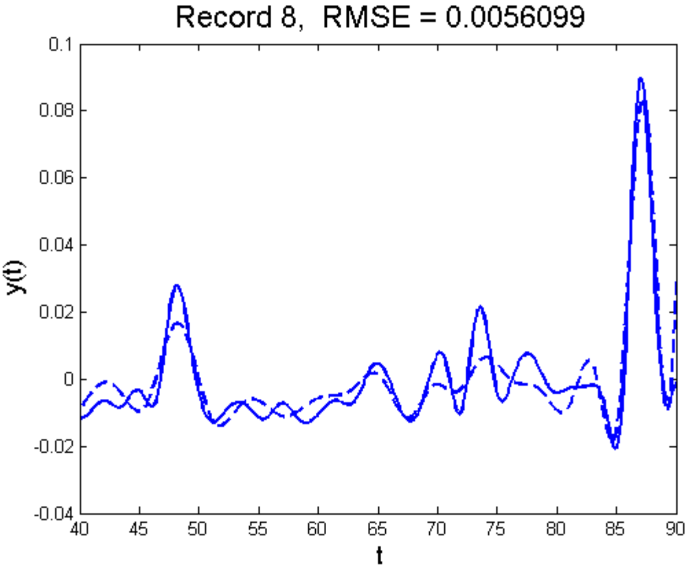

Record 3, RMSE = 0.0053098
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
y(t)
40
45
50
55
60
65
70
75
80
85
90
t

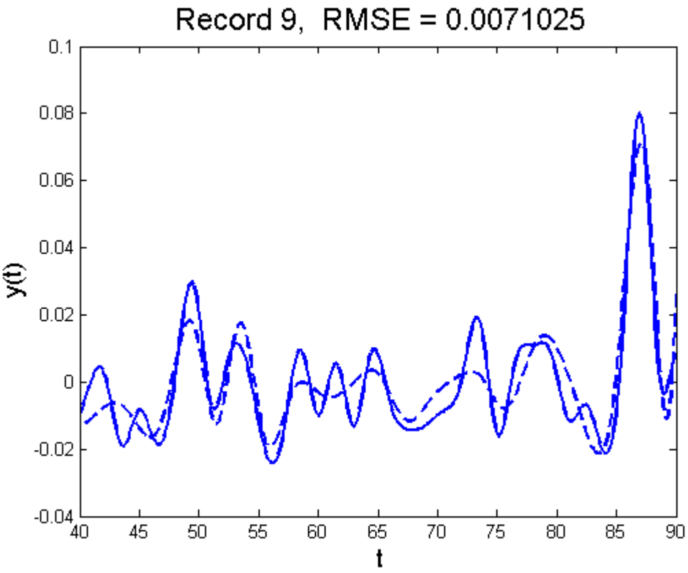

Record 4, RMSE = 0.0038181
y(t)
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
40
45
50
55
60
65
70
75
80
85
90
t

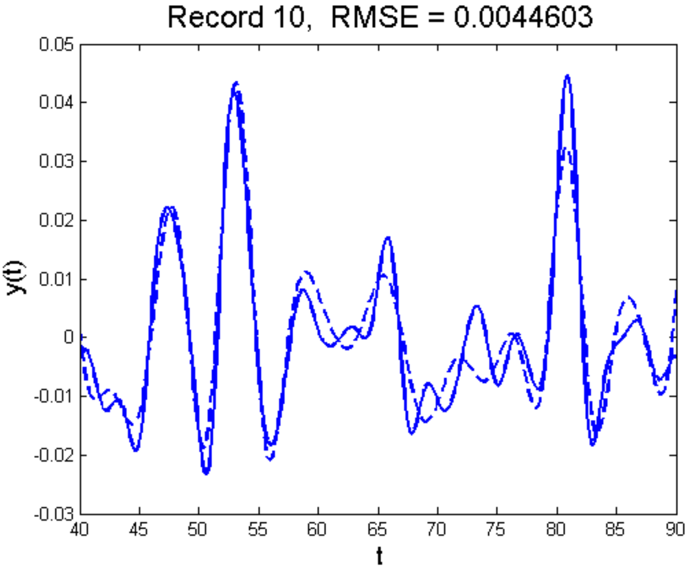

Record 5, RMSE = 0.0068105
y(t)
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
40
45
50
55
60
65
70
75
80
85
90
t

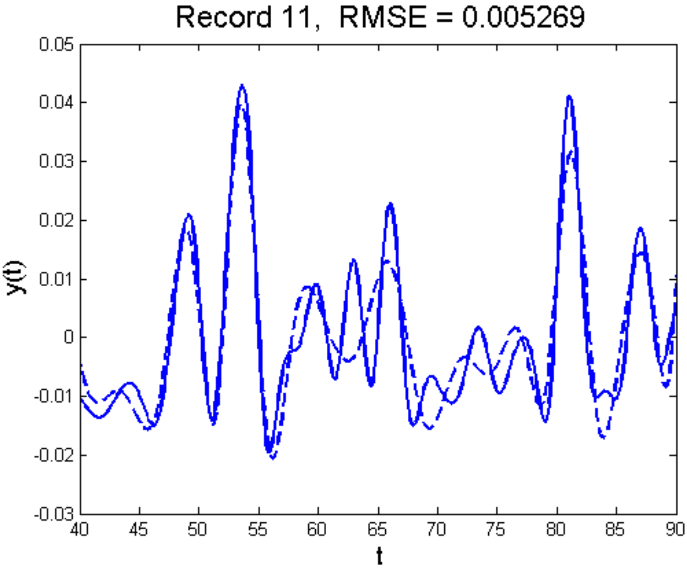

Record 6, RMSE = 0.0049921
0.025
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
y(t)
40
45
50
55
60
65
70
75
80
85
90
t

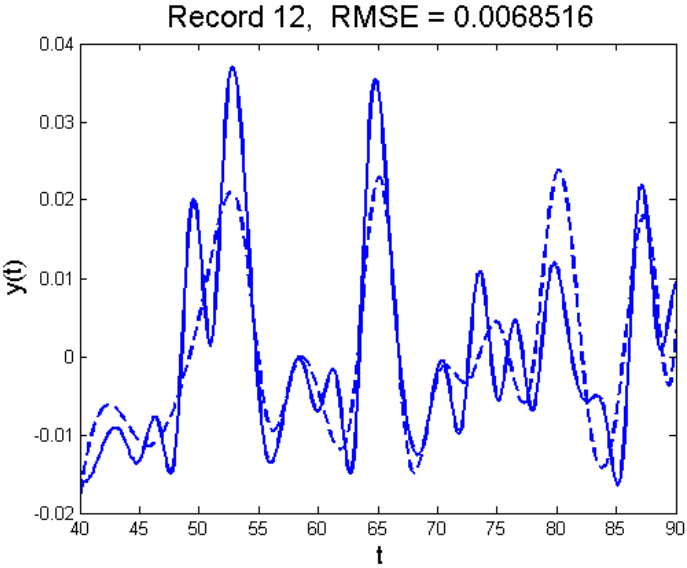

Record 7, RMSE = 0.0038948
y(t)
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
40
45
50
55
60
65
70
75
80
85
90
t

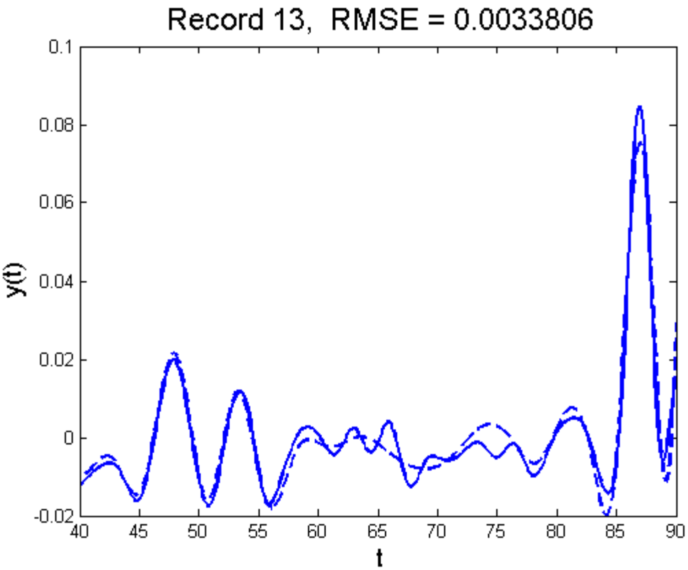

Record 8, RMSE = 0.0056099
y(t)
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
40
45
50
55
60
65
70
75
80
85
90
t

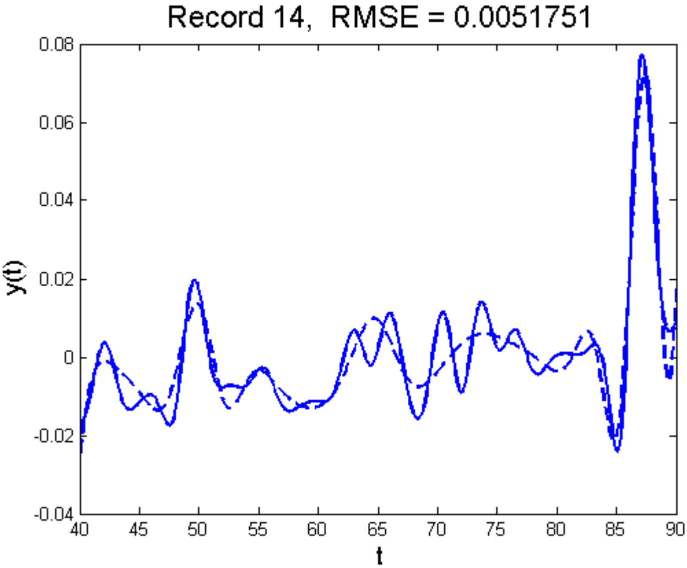

Record 9, RMSE = 0.0071025
0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
y(t)
40
45
50
55
60
65
70
75
80
85
90
t

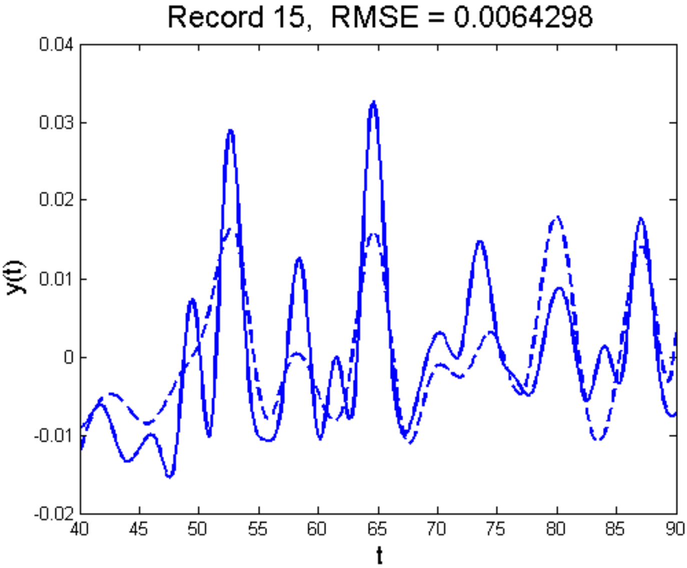

Record 10, RMSE = 0.0044603
y(t)
t
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
40
45
50
55
60
65
70
75
80
85
90


Record 11, RMSE = 0.005269
y(t)
0.05
0.04
0.03
0.02
0.01
0
-0.01
-0.02
-0.03
40
45
50
55
60
65
70
75
80
85
90
t


Record 12, RMSE = 0.0068516
y(t)
t
0.04
0.03
0.02
0.01
0
-0.01
-0.02
40
45
50
55
60
65
70
75
80
85
90


Record 13, RMSE = 0.0033806
y(t)
t
0.1
0.08
0.06
0.04
0.02
0
-0.02
40
45
50
55
60
65
70
75
80
85
90


Record 14, RMSE = 0.0051751
y(t)
0.08
0.06
0.04
0.02
0
-0.02
-0.04
40
45
50
55
60
65
70
75
80
85
90
t


Record 15, RMSE = 0.0064298
y(t)
t
0.04
0.03
0.02
0.01
0
-0.01
-0.02
40
45
50
55
60
65
70
75
80
85
90

Conclusions

- PCA via eigenanalysis restricts the extendability and versatility of PCA.
- Parameter cascading not only regularizes the estimation of nuisance parameters in **F**,
- It re-defines PCA as a much lower dimensional fitting problem.
- Other variations of PCA are being investigated, including