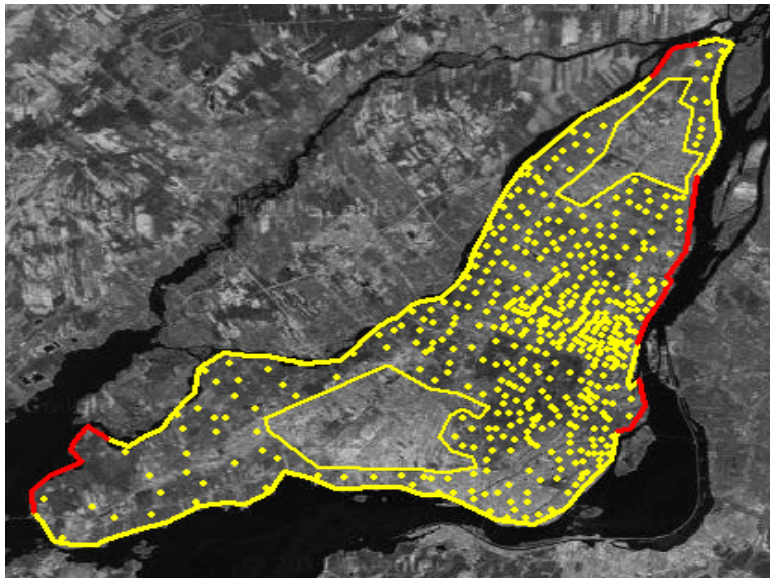
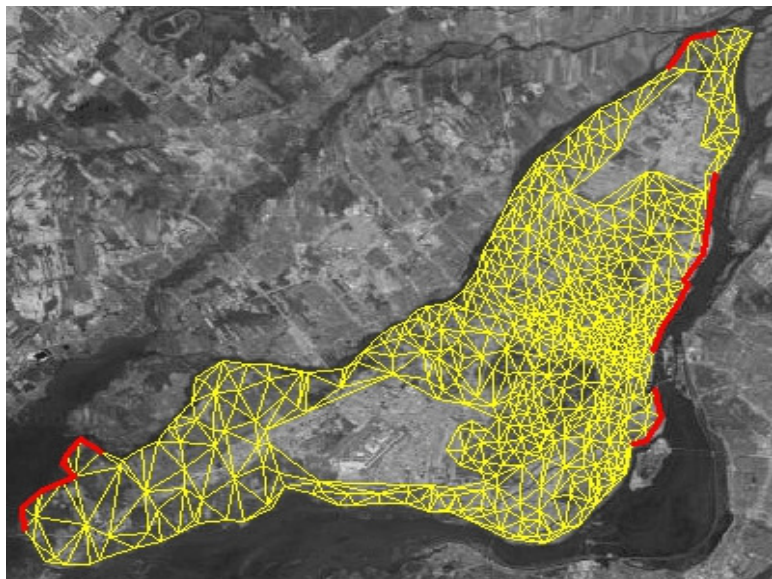


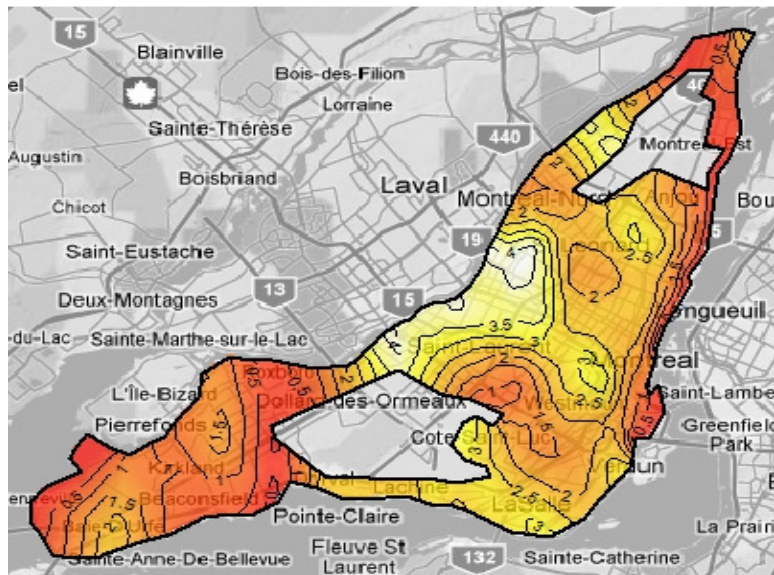
Multivariate and Functional Principal Components without Eigenanalysis

Jim Ramsay, McGill University
(Alois Kneip, University of Bonn)
International Society of Nonparametric Statistics
Cadiz, Spain
13 June 2014



Sangalli, L. M., Ramsay, J. O. and Ramsay, T. O. (2013) Spatial spline regression models. *Journal of the Royal Statistical Society, Series B*, **75**, 681-703.





PCA: The essential idea (Multivariate Case)

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

$$\mathbf{X} = \mathbf{F}\mathbf{A}$$

- where \mathbf{A} is a K by n matrix of principal component coefficients, with $K \ll n$
- and \mathbf{F} is a N by K matrix of principal component scores
- Usually $N \gg n$, 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 the optimal linear subspace \mathcal{R}^K , called a Grassmann manifold.

PCA: The essential idea (Functional Case)

- We have a N curves $x_i(t)$.
- We propose the reduced rank K bilinear model

$$\mathbf{x}(t) = \mathbf{F}\mathbf{a}'(t)$$

- where \mathbf{a} is a vector K principal component functions, and
- and \mathbf{F} is a N by K matrix of principal component scores.
- The fundamental goal of PCA is to identify the optimal linear subspace of functions \mathbf{a} .

Structural parameters

- 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.
- Principal loading matrix **A** is a structural parameter in multivariate PCA.

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.
- The principal component scores matrix \mathbf{F} contains 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, for example.

What we'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 \mathbf{F} as nuisance parameters and PC components in \mathbf{A} as structural parameters.
- Generalize PCA:
 - synthesize the treatment of multivariate and functional data
 - implement partial least squares: an approximation of an external vector \mathbf{y} via a K dimensional subspace \mathcal{R}^K
 - combine PCA with the registration of functional data

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

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

Why eigenanalysis gets in the way

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

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

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.

The parameter cascading strategy for multivariate PCA

- We add smoothness to the least squares criterion for \mathbf{F} given \mathbf{A} by attaching penalty terms:

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

- 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 \mathbf{A}

- This is defined in terms of only the PC coefficients \mathbf{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 \mathbf{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 : $\mathbf{F}\mathbf{W}\mathbf{W}^{-1}\mathbf{A}$ 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.

The parameter cascading strategy for functional PCA (functional case)

- The data are now N functions $x_i(t)$
- The principal coefficients are now functions $a_k(t), k = 1, \dots, K$.
- 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

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

where penalty matrix \mathbf{U} defines a roughness penalty for the a_k 's.

The PCA/PLS hybrid criterion

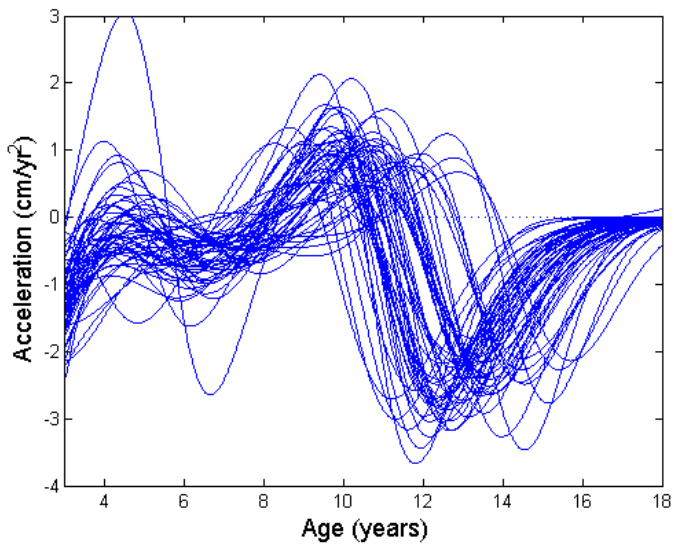
- Keeping to LS fitting for illustration, we now use fitting criterion

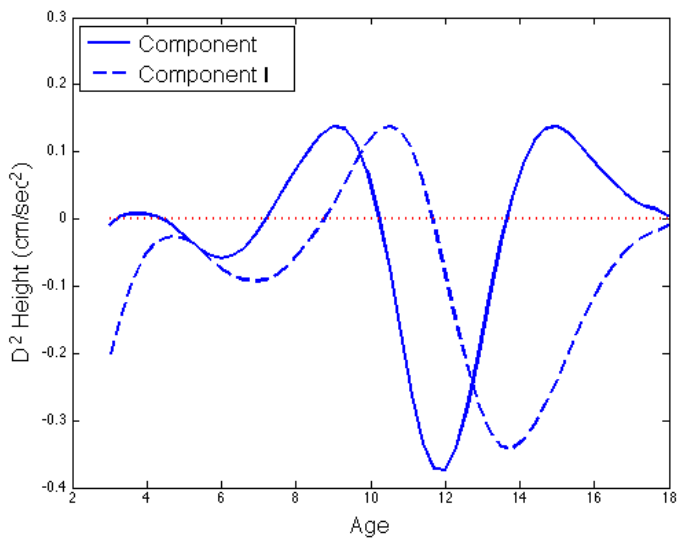
$$G(\mathbf{A}|\mathbf{X}, \mathbf{y}) = (1 - \gamma)\|\mathbf{X} - \mathbf{FA}\|^2 + \gamma\|\mathbf{y}'\mathbf{Q}(\mathbf{A})\mathbf{y}\|^2.$$

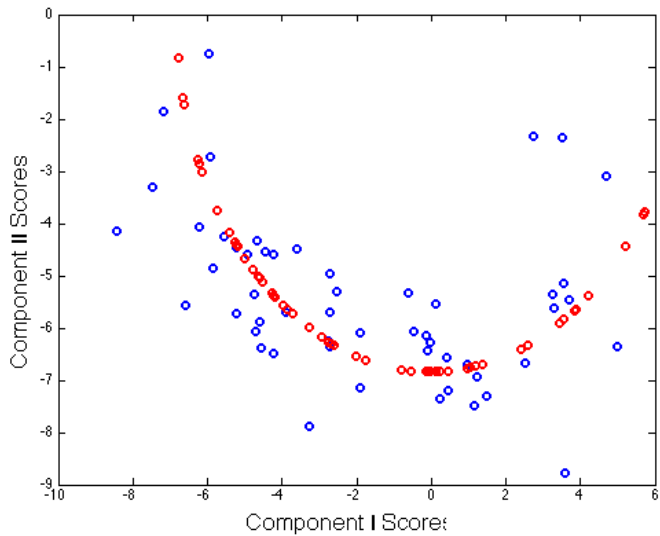
where the relaxation parameter $\gamma \in [0, 1]$ and

$$\mathbf{Q}(\mathbf{A}) = \mathbf{I} - \mathbf{F}(\mathbf{A})[\mathbf{F}(\mathbf{A})'\mathbf{F}(\mathbf{A})]^{-1}\mathbf{F}(\mathbf{A})'.$$

- The second term measures the extent to which external variable \mathbf{y} is unpredictable from within the subspace defined by the PC loadings in \mathbf{A} .
- The boundary conditions $\gamma = 0$ and $\gamma = 1$ correspond to pure PCA and pure partial least squares, respectively.
- The unregularized version criterion was first developed by de Jong and Kiers (1992).

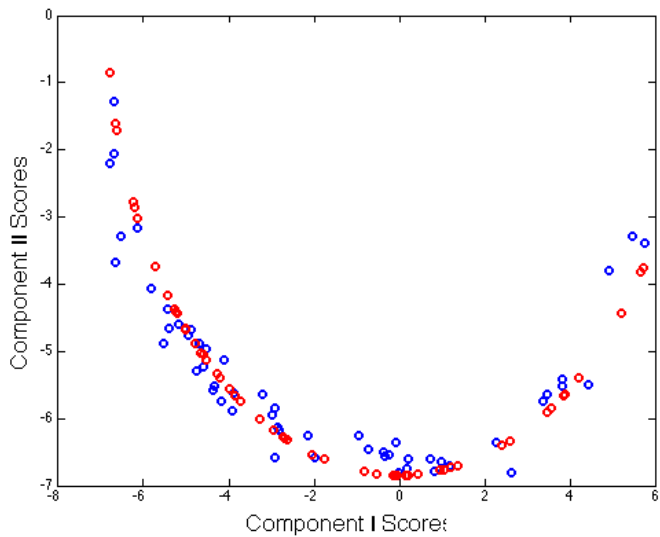




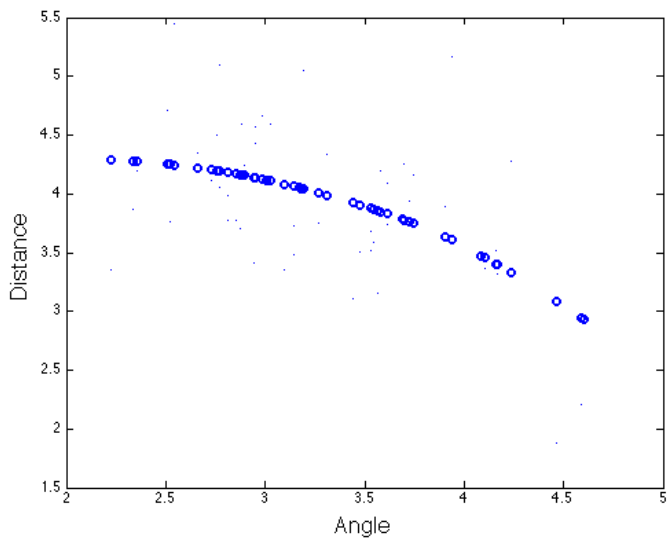


Regularized PCA of the children's acceleration curves

- The principal component scores in \mathbf{F} are close to being on a circle, indicated by the red dots.
- We would like to explore the use of scores that are required to be close to or on the circle.
- The penalty term $\lambda_2 \|\mathbf{F}\mathbf{P}_2\mathbf{F}'\|^2$, where projection matrix \mathbf{P} projects scores on to the circle of red dots, will serve that purpose.
- Here are the scores resulting from using $\lambda_2 = 1$.



- The unconstrained error sum of squares was 127.2 and the constrained value was 138.2, corresponding to a squared multiple correlation of 0.08.
- A heavier penalty puts the scores nearly exactly on the circle, corresponding to $R^2 = 0.12$.
- The angle associated with each pair of scores measures phase variation, which is how early or late the pubertal growth spurt is.
- But, we might have missed something ...



- The scores of the girls in the upper left are outside of the constant distance curve, and the girls on the bottom and lower right are inside.
- The upper left girls have earlier puberty, and also more intense spurts; the late puberty girls have milder growth spurts.
- Early puberty girls are compensated for losing out on a few years of growth by having more intense spurts.
- It looks like principal component scores for uncentered functional observations should be represented in hyperspherical coordinates!

Conclusions

- PCA via eigenanalysis restricts the extendability and versatility of PCA.
- Parameter cascading re-defines PCA as a much lower dimensional fitting problem,
- and greatly extends its capacity of represent data in a lower dimensional space.