Multivariate and Functional Principal Components without Eigenanalysis

Jim Ramsay, McGill University
Alois Kneip, University of Bonn
Department of Applied Mathematics
University of Colorado at Boulder
24 October 2014

PCA: The essential idea (Multivariate Case)

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

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

- where $\mathbf{A}$ is a K by n matrix of principal component coefficients, with $K << n$
- and $\mathbf{F}$ 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 the optimal linear subspace $\mathcal{R}^{\mathcal{K}}$, 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 **a** is a vector K principal component functions, and
- and **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 **a**.

PCA: The essence of PCA

- PCA identifies an optimal flat subspace.
- In principle, this task can be carried out in any Euclidean space, and does not require a rectilinear orthogonal coordinate system.
- Or, in fact, any coordinate system at all. In the Euclidean case, the subspace is called a *Grassmann manifold.*
- But we will assume a vector space structure in this talk.

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 $\mathbf{A}$ 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 **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.

More Generalizations of 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 $\mathcal{R}^{\mathcal{K}}$
- 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

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

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.

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 $\boldsymbol{\theta}$ requires the use of the Implicit Function Theorem:

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

- 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:

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

- 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: $\mathbf{F}\mathbf{W}\mathbf{W}^{-1}\mathbf{A}$ 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 $a_k(t), k = 1, \ldots, K$.
- 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.

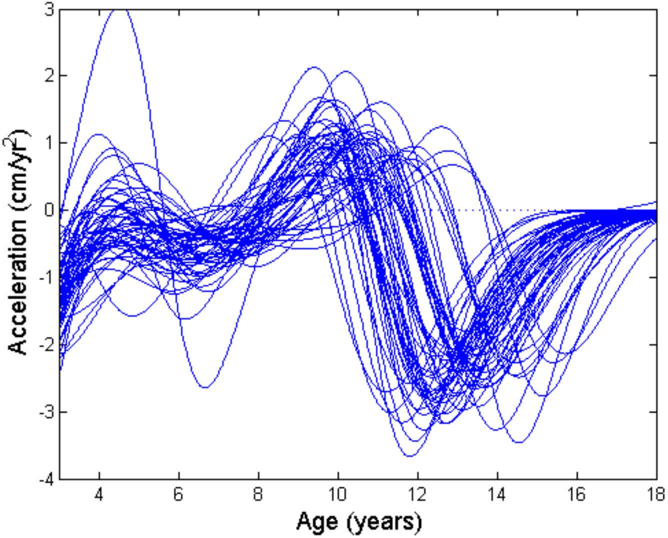
- The outer criterion could be

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

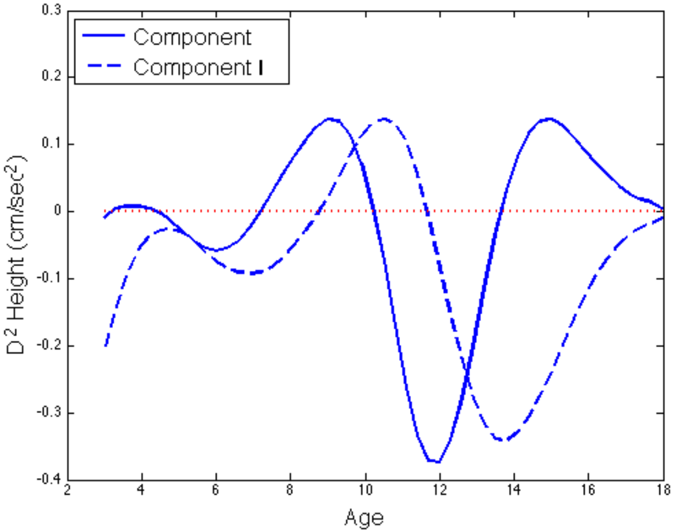
where penalty matrix **U** defines a roughness penalty for the a_k's.

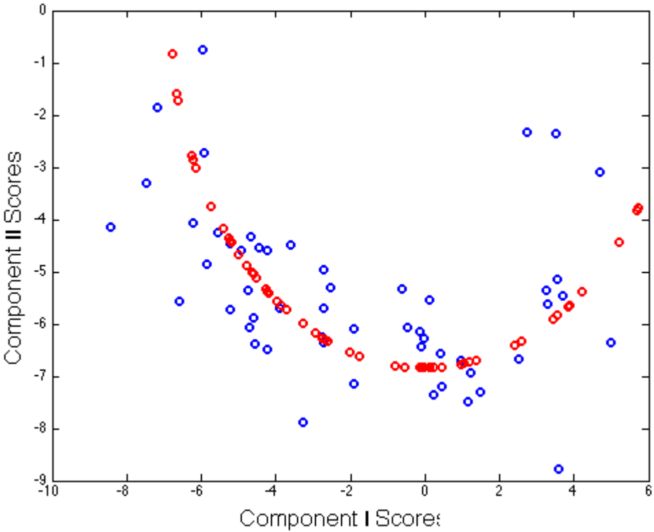
Example 1. PCA of female height acceleration curves

- The Berkeley Growth data contain heights of 56 girls at 31 unequally spaced ages.
- Nice estimates of height acceleration are possible using monotone smoothing methods.
- The principal component scores have a tightly curvilinear structure.



Acceleration (cm/yr^2)
Age (years)
3
2
1
0
-1
-2
-3
-4
4
6
8
10
12
14
16
18

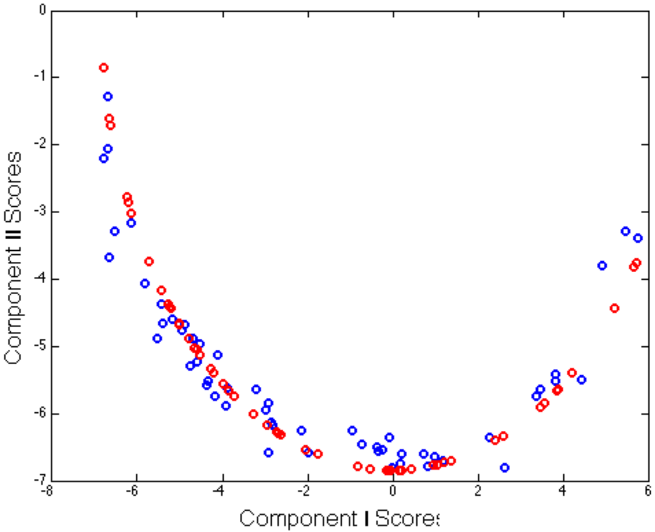

Component
Component I
D 2 Height (cm/sec2)
Age
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
2
4
6
8
10
12
14
16
18



Component II Scores
Component I Scores
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-8
-6
-4
-2
0
2
4
6

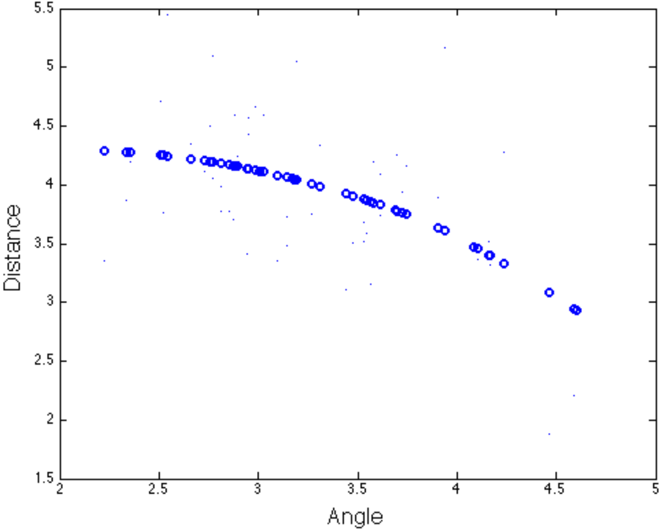
Regularized PCA of the children's acceleration curves

- The principal component scores in **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 **P** projects scores on to the circle of red dots, will serve that purpose.
- Here are the scores resulting from using $\lambda_2 = 1$.



Component II Scores
Component I Scores
0
-1
-2
-3
-4
-5
-6
-7
-8
-6
-4
-2
0
2
4
6

- 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 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 happens.
- But, we might have missed something ...



Distance
Angle
5.5
5
4.5
4
3.5
3
2.5
2
1.5
2
2.5
3
3.5
4
4.5
5

- 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 hyper-spherical coordinates!

The PCA/PLS hybrid criterion

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

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

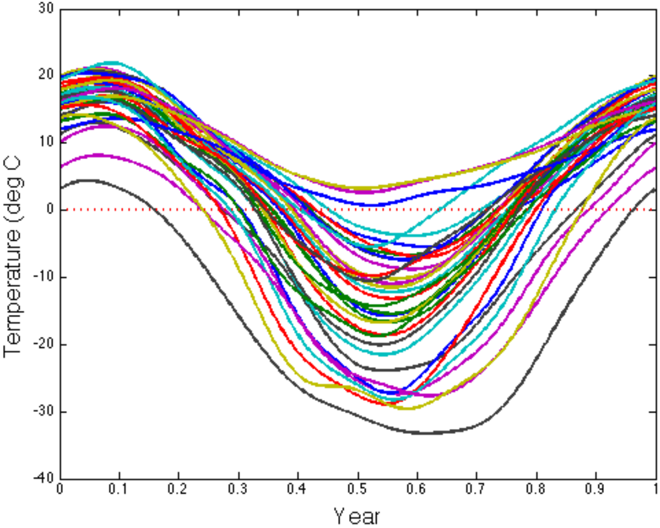
  where the relaxation parameter $\gamma \in [0,1]$ and

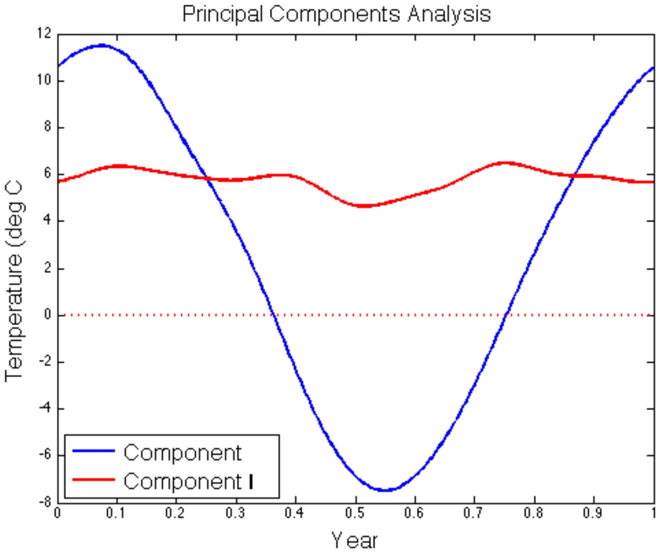
$$\mathbf{Q}(\mathbf{A}) = \mathbf{I} - \mathbf{F}(\mathbf{A})[\mathbf{F}(\mathbf{A})'\mathbf{F}(\mathbf{A})]^{-1}\mathbf{F}(\mathbf{A})'.$$

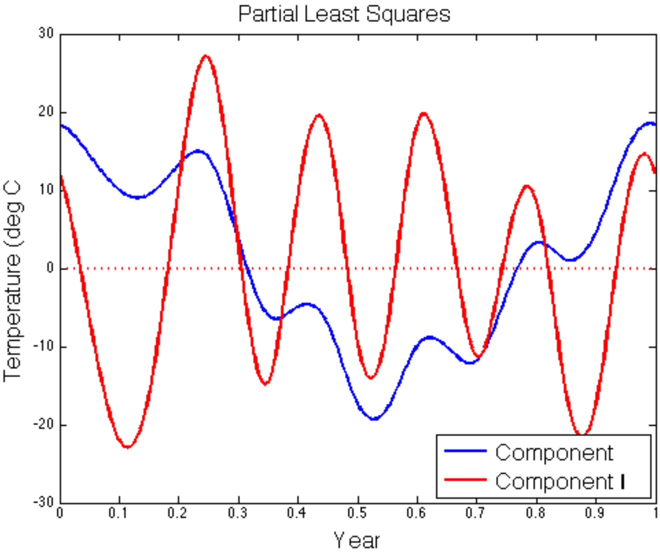
- The second term measures the extent to which external variable **y** is unpredictable from within the subspace defined by the PC loadings in **A**.
- The boundary conditions $\gamma = 0$ and $\gamma = 1$ correspond to pure PCA and pure partial least saquares, respectively.
- The unregularized solution was worked out by de Jong and Kiers (1992).

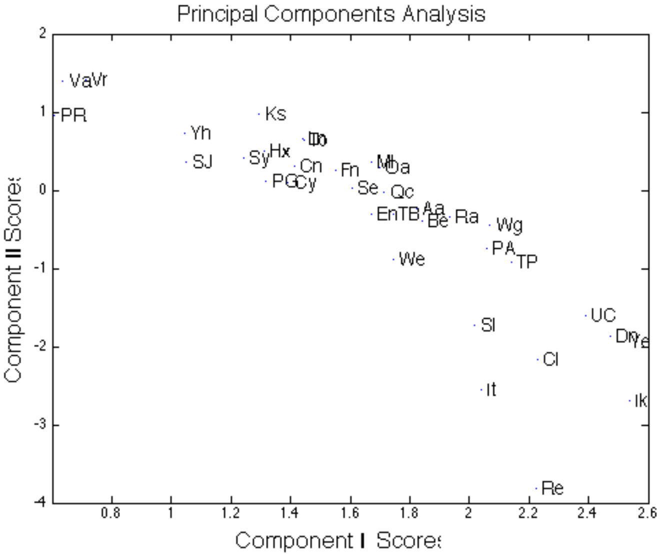
Example 2. PCA and PLS fits for daily average temperature and precipitation

- The Canadian weather data consist of daily temperature and precipitation data for 35 weather stations averaged over 34 years.
- We run the year from July 1st to June 30th in order to highlight winter variation.
- PCA of the temperature shows that two principal components can fit 97.3% of the temperature variation.
- How well can we fit annual precipitation averages from the principal component scores,
- and from two component scoress identified by PLS?

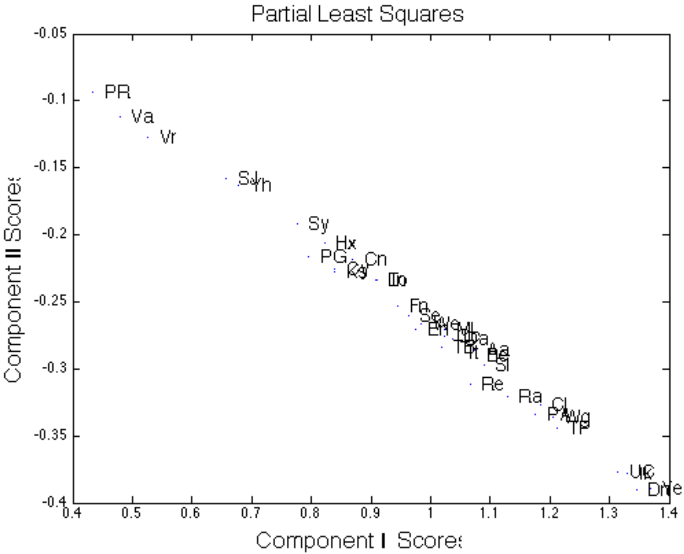

30
20
10
0
-10
-20
-30
-40
Temperature (deg C
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Year

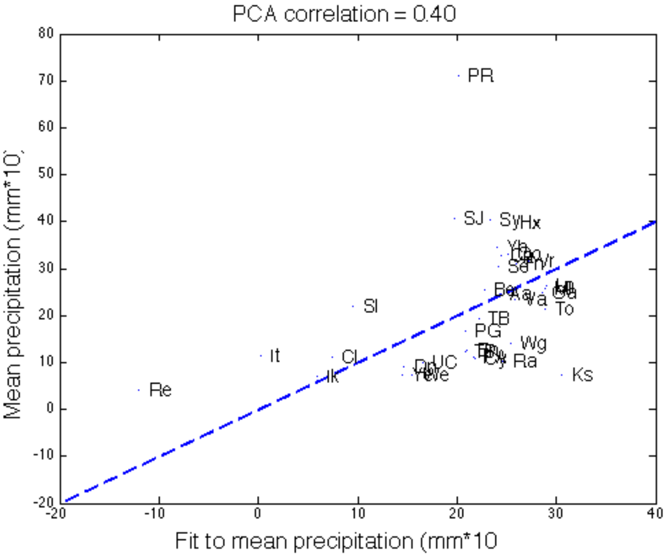

Principal Components Analysis
Temperature (deg C
Year
Component
Component I
12
10
8
6
4
2
0
-2
-4
-6
-8
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


Partial Least Squares
Temperature (deg C)
30
20
10
0
-10
-20
-30
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Year
Component
Component I

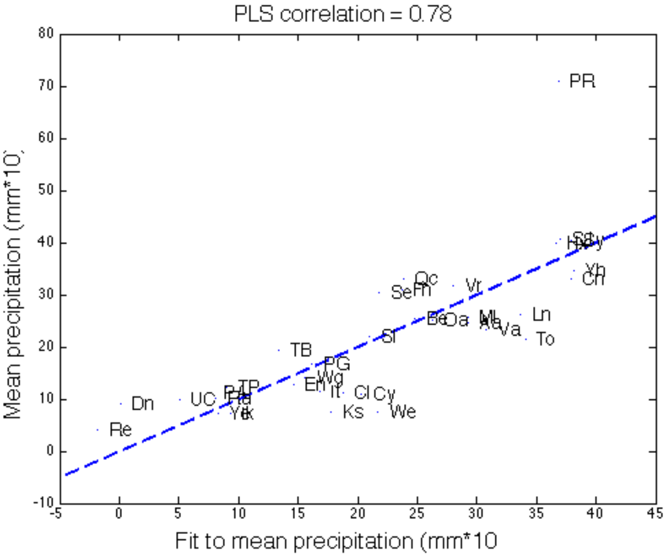


Principal Components Analysis
Component II Scores
Component I Scores
2
1
0
-1
-2
-3
-4
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
VaVr
PR
Ks
Yh
SJ
Sy
Hx
Cn
PG
Cy
Fn
Se
Qc
Oa
EnTB
Aa
Be
Ra
Wg
PA
TP
We
UC
Sl
Cl
It
Ik
Re


Partial Least Squares
Component II Score
Component I Score
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.35
-0.4
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
PR
Va
Vr
Sy
Hx
PG
Cn
Fo
Re
Ra
Cl



PCA correlation = 0.40
Mean precipitation (mm*10)
Fit to mean precipitation (mm*10
80
70
60
50
40
30
20
10
0
-10
-20
-20
-10
0
10
20
30
40
PR
SJ
Sy
Hx
SI
It
Cl
Ik
Re
TB
PG
Wg
Ra
Ks
To


PLS correlation = 0.78
Mean precipitation (mm*10)
Fit to mean precipitation (mm*10
80
70
60
50
40
30
20
10
0
-10
-5
0
5
10
15
20
25
30
35
40
45
PR
Ch
Vr
Qc
Se
Be
Oa
Ln
Va
To
Sl
TB
PG
Wg
It
Cl
Cy
Ks
We
Dn
UC
TP
Re

Example 2. Conclusions

- Fitting an external variable using principal component scores ("PCA regression") does achieve something.
- But optimizing the subspace for this task does much better.
- The canonical correlations between these two subspaces are 0.999 and 0.865, respectively.
- The two subspaces differ mainly in terms of the second component:
    - In PCA this straightforward annual level.
    - In PLS this is a 5-cycle sinusoid.
- The PLS fits group nicely into five tight clusters plus Prince Rupert on the upper west coast.
- In ascending order of amount of precipitation they are: (1) High Arctic, (2) Sub Arctic (3) Prairie, (4) Great Lakes, St. Lawrence and (5) coastal.

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 ability to represent data in a lower dimensional space.
- Roughness penalties or regularization can lead to simpler principal component structures.
- Partial least squares can do substantially better than PCA-regression in fitting external variables using high dimensional covariate spaces.