ORIE 779 – Spring 2002

Functional Data Analysis

Student Presentation

Rotation of Principal Components and the VARIMAX Criterion

Presented by:

Trevor Park

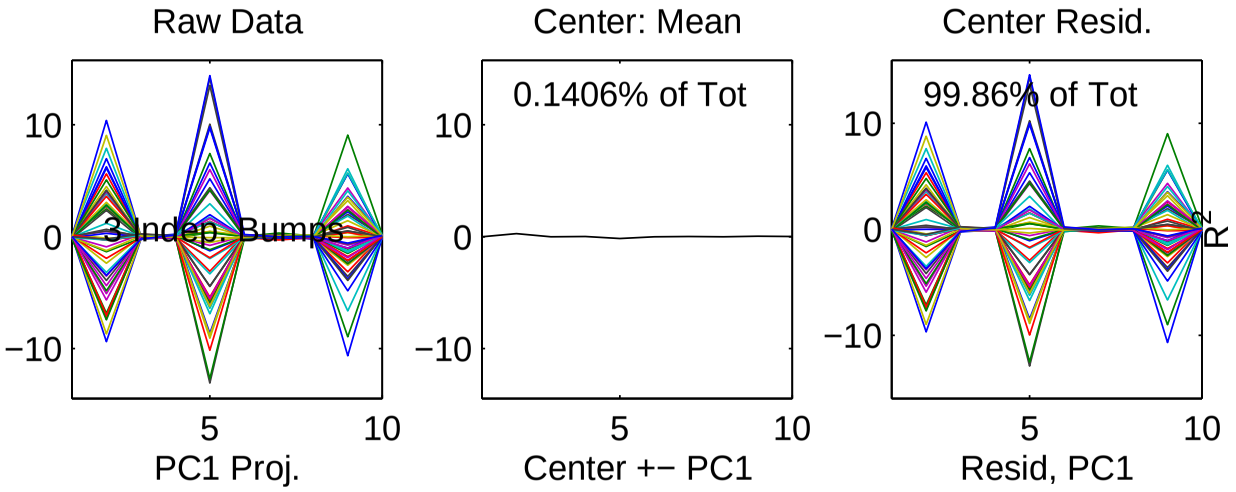
March 27, 2002

Motivation

Consider this "Three Independent Bump" toy example:

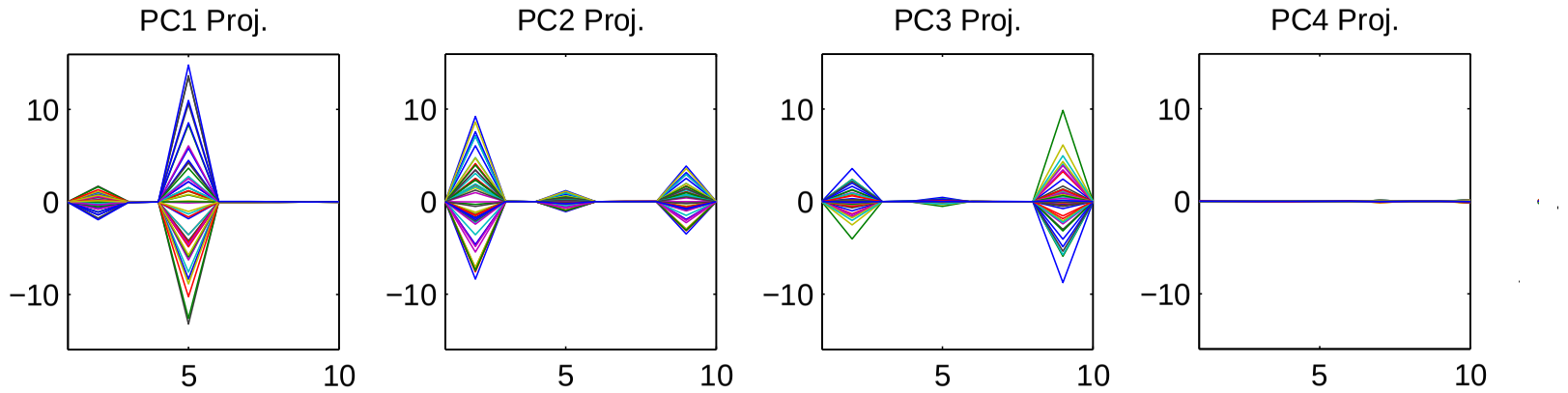
Simulate 50 10-point curves that vary mainly at three points. The variations at these three points are indep. Gaussians with standard deviations 1, 3/2, and 2/3.

We want to use principal components analysis to separate these "bumps," i.e. to show us that they are independent.



Motivation

However, the first few principal components do not separate the "bumps" very cleanly:



The first three PCs capture essentially all of the variation in the bumps, but they seem to be the "wrong" directions for the kind of analysis that we want.

Conceptually, the problem is that the principal component eigenvalues have very similar magnitudes, leading to ill-defined principal components.

Motivation

Since we want to separate the bumps, it would be nice if we could get components in directions that are more aligned with the coordinate axes (since the ideal "bump" represents variation along only one axis).

Still, we want to make use of the information contained in the principal components.

How can we do this?

A reasonable compromise: Take the principal components that are of interest, but *rotate* them so that they still define the same subspace (i.e. still account for the same sum-of-squares of the data) but are "more aligned" with the axes.

Background: Factor Analysis

The rotation idea has its origins in psychometrics, particularly in connection with the *factor analysis model*:

Assume a population of n individuals (*objects*), on each of which we have d separate numerical measurements (*features*). Organize into vectors for each individual:

$$\mathbf{X}_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{dj} \end{bmatrix}, \quad j = 1, \ldots, n$$

So X_{ij} represents the i^{th} measurement on the j^{th} individual.

Background: Factor Analysis

Assume the individual vectors follow a linear-regression-style model

$$\mathbf{X}_j = \boldsymbol{\mu} + \mathbf{A}\mathbf{f}_j + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, n$$

where

- $\boldsymbol{\mu}$ is an unknown $d \times 1$ common mean vector,
- $\mathbf{A}$ is an unknown $d \times m$ *loadings* matrix,
- $\mathbf{f}_j$ are i.i.d. $m \times 1$ random *factor* vectors such that

$$\mathrm{E}\,\mathbf{f}_j = \mathbf{0}, \quad \mathrm{Var}\,\mathbf{f}_j = \mathbf{I}_m,$$

- $\boldsymbol{\epsilon}_j$ are i.i.d. $d \times 1$ random "error" vectors with

$$\mathrm{E}\,\boldsymbol{\epsilon}_j = \mathbf{0}, \quad \mathrm{Var}\,\boldsymbol{\epsilon}_j = \boldsymbol{\Psi}, \quad \mathrm{Cov}(\boldsymbol{\epsilon}_j, \mathbf{f}_j) = \mathbf{0}$$

  where $\boldsymbol{\Psi}$ is a diagonal $d \times d$ matrix.

Background: Factor Analysis

Archetypal Example: Psychometrics

The individuals are people, and the measurements are their scores on a battery of standardized tests.

Apart from a common vector $\boldsymbol{\mu}$ of mean scores and some random variation $\boldsymbol{\epsilon}_j$, the differences in individual performance are explained by the terms

$$\mathbf{A}\mathbf{f}_j,$$

composed of

- the m-vectors $\mathbf{f}_j$ whose elements are *factors* numerically representing uncorrelated one-dimensional aspects of ability, normalized over the population
- the $d \times m$ *loadings* matrix $\mathbf{A}$ representing how much each factor influences performance on each test

Some examples of standardized tests: verbal skills, math skills, memory, dexterity, etc.

Note that the random variation ϵ_j is assumed to be just due to the variability that we might expect if a person takes the same kind of test many times.

Note that the factors are assumed to influence test performance *linearly*, as an approximation.

Background: Factor Analysis

Factor analysis was developed for exactly this type of problem.

It was originated by Charles Spearman in 1904, with earlier influences from Karl Pearson, as a way of objectively discovering the existence and nature of "general ability" or "general intelligence". Later on, factors discovered were assumed to be independent axes along which intelligence and ability varied.

General uses are:

- *Exploratory* — to find latent structure in the data, with hope of interpreting it
- *Confirmatory* — to determine whether assumed structure is actually present in the data

This viewpoint is not necessarily widely accepted in psychology today, and my summary of it is a bit of a caricature. Use of factor analysis in psychometrics has historically been quite controversial.

"Confirmatory" is used a bit loosely here; it doesn't necessarily imply a formal statistical test.

Background: Factor Analysis

Estimation in factor analysis focuses on the loadings $\mathbf{A}$, and is typically based on the population correlation matrix:

$$\text{Var}(\mathbf{X}_j) = \Sigma = \mathbf{A}\mathbf{A}^\mathsf{T} + \mathbf{\Psi}$$

which is estimated by the usual

$$\hat{\Sigma} = \mathbf{X}\left(\mathbf{I}_n - \mathbf{1}\cdot\mathbf{1}^\mathsf{T}/n\right)\mathbf{X}^\mathsf{T}$$

where $\mathbf{X} = [\,\mathbf{X}_1 \ \ldots \ \mathbf{X}_n\,]$.

The elements of $\mathbf{A}$ and Σ could only possibly be identified if the number of factors is small compared to the number of measurements ($m < d/2$).

There are many ways of estimating, including maximum likelihood based on Gaussian distributional assumptions. See Seber (1984).

Lead-in: How is factor analysis performed with so many unknowns?

Refer to the factor analysis model slide, if necessary, to motivate the expression for $\text{Var}(\mathbf{X}_j)$.

Think of $\mathbf{A}$ as a "thin" matrix.

Estimation methods tend to be computationally intensive.

Background: Factor Analysis

Even with a small number of factors, the loadings matrix is still not identified: If $\mathbf{R}$ is an orthogonal $m \times m$ matrix (a *rotation*) and

$$\mathbf{B} = \mathbf{AR},$$

and we define

$$\tilde{\mathbf{f}}_j = \mathbf{R}^\mathsf{T}\mathbf{f}_j$$

then

$$\mathbf{B}\tilde{\mathbf{f}}_j = \mathbf{ARR}^\mathsf{T}\mathbf{f}_j = \mathbf{Af}_j$$

and

$$\mathsf{E}\,\tilde{\mathbf{f}}_j = 0, \quad \mathsf{Var}\,\tilde{\mathbf{f}}_j = \mathbf{R}^\mathsf{T}\mathbf{I}_m\mathbf{R} = \mathbf{I}_m,$$
$$\mathsf{Cov}(\epsilon_j, \tilde{\mathbf{f}}_j) = \mathsf{Cov}(\epsilon_j, \mathbf{f}_j)\mathbf{R} = 0.$$

so $(\mathbf{A}, \mathbf{f}_j)$ and $(\mathbf{B}, \tilde{\mathbf{f}}_j)$ are equivalent loading/factor pairs.

This is essentially the maximal extent of the unidentifiability; assuming we have the correct number of factors, all possible valid loadings matrices are given by rotations of a single valid rotation matrix.

Note that these rotations preserve the subspace spanned by the columns of $\mathbf{A}$. So the same subspace is being represented by a different set of spanning vectors.

Background: Factor Analysis

Q: How should we choose the loadings matrix?
A: We want the loadings matrix that we can *most easily interpret*.

In the standardized test example:

We have some idea of what each test is intended to measure, so we should be able to interpret a factor according to which tests have its largest loadings.

So ideally we want, in each column of the loadings matrix $\mathbf{A}$,

- a few loadings of relatively large magnitude, and
- all other loadings very close to zero.

Background: Factor Analysis

Problem: Find a numerical criterion that evaluates how well a loadings matrix conforms to this ideal.

Given such a function ϕ defined on matrices, giving larger values for more conforming matrices, and an initial loadings matrix $\mathbf{A}$, we can then optimize

$$\max_{\text{orthogonal } \mathbf{R}} \phi(\mathbf{AR}).$$

In psychometrics, many formal criteria were proposed starting in the 1950's. The best known and most widely used criterion is VARIMAX, due to Kaiser (1958).

Initially, psychometricians did not use formal numerical criteria, but rather used intuition and ad hoc methods to find rotations that were subjectively appealing.

The VARIMAX Criterion

For a (loadings) matrix $\mathbf{A} = [\, a_{ik} \,]_{d \times m}$, Kaiser's VARIMAX is

$$\phi_V(\mathbf{A}) = \sum_k \left(d \sum_i a_{ik}^4 - \left(\sum_i a_{ik}^2 \right)^2 \right) / d^2.$$

In words,

1. for each column, take the variance of the *squares* of the elements, then
2. sum these variances over all columns.

The matrix that maximizes this criterion over some appropriately bounded set of matrices should be the "most interpretable" in that set.

(Note that rotations of a matrix $\mathbf{A}$ all have the same Fröbenius norm: $\mathrm{tr}(\mathbf{AR}(\mathbf{AR})^\mathsf{T}) = \mathrm{tr}(\mathbf{ARR}^\mathsf{T}\mathbf{A}^\mathsf{T}) = \mathrm{tr}(\mathbf{AA}^\mathsf{T})$.)

The version of VARIMAX defined here is what Kaiser called "raw" VARIMAX. Kaiser actually recommended that the rows of the loadings matrix be normalized before applying the criterion: "normal" VARIMAX. Though this has some application-specific appeal, it is not necessarily needed in rotation for principal components analysis.

The Fröbenius norm of a matrix $\mathbf{A}$ is actually $\sqrt{\text{tr}(\mathbf{A}\mathbf{A}^{\text{T}})}$.

The VARIMAX Criterion

Why does VARIMAX work? The intuition is easier for an earlier and simpler criterion called QUARTIMAX:

$$\phi_Q(\mathbf{A}) = \sum_k \sum_i a_{ik}^4.$$

How does this vary over the set of fixed-size matrices $\mathbf{A}$ of constant squared Fröbenius norm $\text{tr}(\mathbf{A}\mathbf{A}^\mathsf{T}) = \sum_k \sum_i a_{ik}^2$?

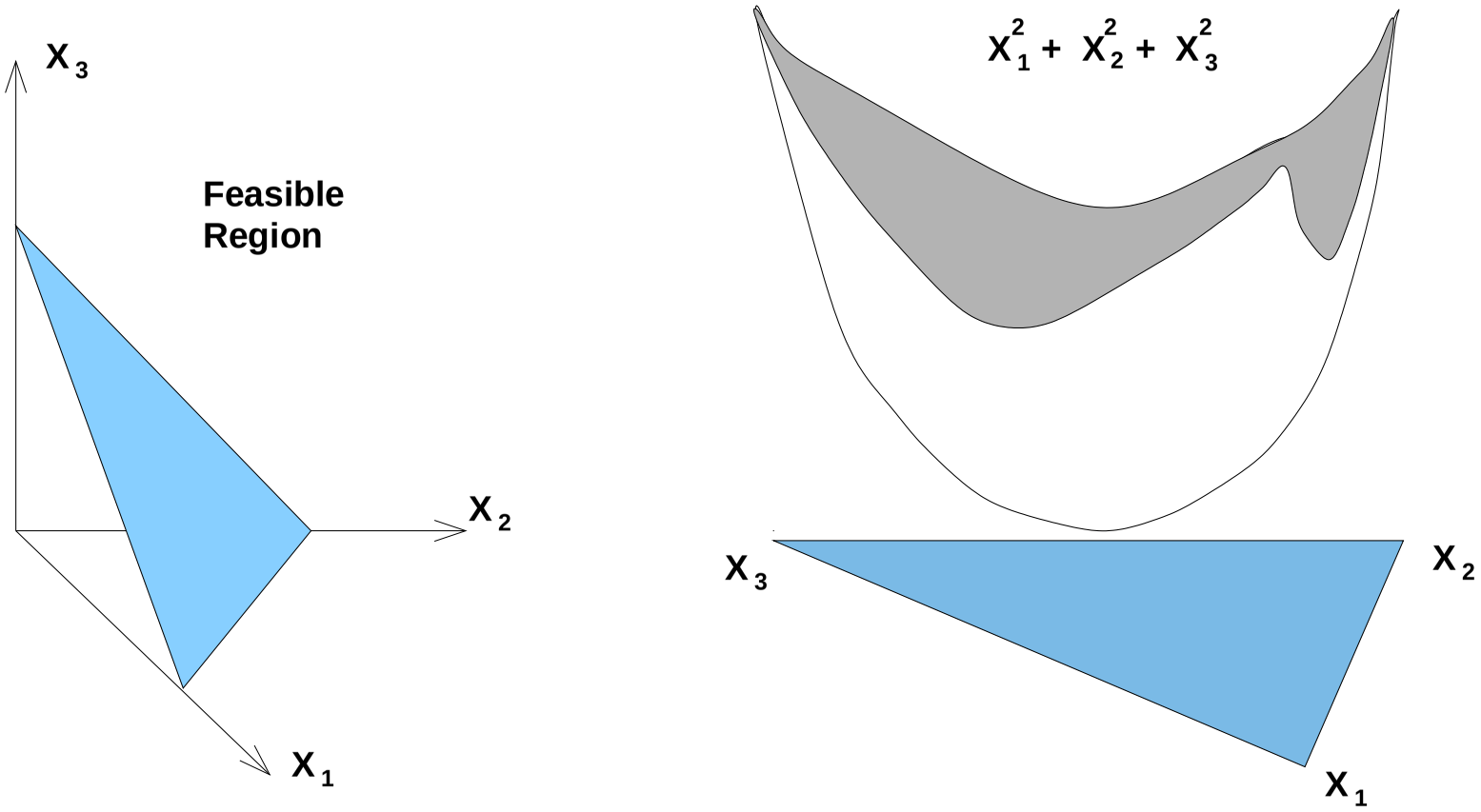
Consider the following problem:

$$\max_{\substack{x_1, x_2, x_3 \geq 0 \\ x_1 + x_2 + x_3 = c}} x_1^2 + x_2^2 + x_3^2.$$

Effectively, this is "maximize the variance of the positive values (x_1, x_2, x_3) subject to a constant mean."

Equate the x_j values with the a_{ik}^2 values. Note that we are only using the *squares* of the elements of the matrix, so signs do not matter. Only magnitudes matter.

The VARIMAX Criterion



(Schematic only.)

The sum of squares is maximized when one value is at its largest and the other two are zero.

This intuition remains true in higher dimensions.

When the feasible set is the set of rotations $\mathbf{AR}$ of a matrix $\mathbf{A}$, then in general the extreme points will not be feasible. However, we expect that maximizing the criterion will lead to some large magnitude elements and all other elements near zero, as desired.

The VARIMAX Criterion

VARIMAX works like QUARTIMAX, except that it disregards the column-to-column variation: For loadings matrix $\mathbf{A} = [\, a_{ik} \,]_{d \times m}$,

$$\frac{1}{m}\,\phi_V(\mathbf{A}) = \frac{1}{m}\sum_k \left(d \sum_i a_{ik}^4 - \left(\sum_i a_{ik}^2 \right)^2 \right) / d^2$$

$$= \frac{1}{md}\sum_{k,i} a_{ik}^4 \; - \; \frac{1}{(md)^2}\left(\sum_{k,i} a_{ik}^2 \right)^2$$

$$- \left[\frac{1}{m}\sum_k \left(\frac{1}{d}\sum_i a_{ik}^2 \right)^2 \; - \; \frac{1}{m^2}\left(\sum_k \left(\frac{1}{d}\sum_i a_{ik}^2 \right) \right)^2 \right]$$

$$= \text{Overall Variance of Squared Elements} -$$

$$\text{Variance of Squared Element Column Means}$$

Think of the VARIMAX criterion as the residual sum of squares left after accounting for the variation in the squared elements of $\mathbf{A}$ that can be attributed to the columns.

The VARIMAX Criterion

VARIMAX and QUARTIMAX are, in fact, related by

$$\phi_V(\mathbf{A}) = \frac{1}{d}\phi_Q(\mathbf{A}) - \frac{1}{d^2}\sum_k \Big(\sum_i a_{ik}^2\Big)^2.$$

If the columns of $\mathbf{A}$ are orthonormal, then

$$\phi_V(\mathbf{AR}) = \frac{1}{d}\phi_Q(\mathbf{AR}) - \frac{m}{d^2}$$

for all orthogonal matrices $\mathbf{R}$ (because $\mathbf{AR}$ still has orthonormal columns.)

So the two criteria are equivalent in this case, and Ramsay & Silverman (1997) actually uses the definition of QUARTIMAX as the definition of VARIMAX.

This is a useless observation in the context of factor analysis, as the loadings matrices generally will not have orthogonal columns, let alone orthonormal columns. However, it can be relevant in the context of principal components, as we will see ...

Application to Principal Components

Though introduced for factor analysis, VARIMAX has become popular for principal components analysis, as well.

Recent search of Science Citation Index —

190 articles where both "VARIMAX" and "principal components" appear, in many areas of applied science:

psychology and behavior research, climatology and environmental studies, geology and remote sensing, education research, criminal justice, dentistry, musicology, botany, dietetics, sports management, chemometrics, neuroscience, veterinary medicine, geriatrics, etc . . .

The articles date from as early as 1975!

Ramsay & Silverman illustrates VARIMAX in an example from climatology in Sec. 6.3.3.

Application to Principal Components

How does VARIMAX apply to principal components? Look at principal components in a different way:

We have a population of n objects represented by (feature) vectors

$$\mathbf{X}_j = \begin{bmatrix} X_{1j} \\ \vdots \\ X_{dj} \end{bmatrix}, \quad j = 1, \ldots, n,$$

which we can combine into a matrix $\mathbf{X} = [\mathbf{X}_1 \ldots \mathbf{X}_n]$.

Consider a data model of the form

$$\mathbf{X}_j = \boldsymbol{\mu} + \mathbf{B}\mathbf{D}^{\frac{1}{2}}\mathbf{f}_j, \quad j = 1, \ldots, n,$$

where

$\mathbf{B}_{d\times d}$ is orthogonal, $\mathbf{D}_{d\times d}$ is diagonal, $\mathbf{f}_j \sim$ indep.$(\mathbf{0}, \mathbf{I}_d)$.

Application to Principal Components

(Note: The "factors" in the vectors $\mathbf{f}_j$ are actually population-normalized versions of the "scores" or "projections" onto the principal component directions.)

Then, for $j = 1, \ldots, n$,

$$\mathsf{E}\,\mathbf{X}_j \;=\; \boldsymbol{\mu},$$

$$\mathsf{Var}\,\mathbf{X}_j \;=\; \mathbf{B}\mathbf{D}^{\frac{1}{2}}\,\mathsf{Var}(\mathbf{f}_j)\,\mathbf{D}^{\frac{1}{2}}\mathbf{B}^{\mathsf{T}} \;=\; \mathbf{B}\mathbf{D}\mathbf{B}^{\mathsf{T}}.$$

So $\mathbf{B}\mathbf{D}\mathbf{B}^{\mathsf{T}}$ is in fact the "theoretical" principal components decomposition for the population.

We estimate this, as usual, by finding the principal components of the empirical variance matrix:

$$\hat{\boldsymbol{\Sigma}} \;=\; \mathbf{X}\,(\mathbf{I}_n - \mathbf{1}\cdot\mathbf{1}^{\mathsf{T}}/n)\,\mathbf{X}^{\mathsf{T}} \;=\; \hat{\mathbf{B}}\hat{\mathbf{D}}\hat{\mathbf{B}}^{\mathsf{T}}$$

The elements of the vectors $\mathbf{f}_j$ are called "scores" both in the context of principal components and in the context of factor analysis.

Application to Principal Components

Principal Components

$$\mathbf{X}_j = \boldsymbol{\mu} + \mathbf{B}\mathbf{D}^{\frac{1}{2}}\mathbf{f}_j$$

- coordinate-free (invariant to orthogonal rotations)
- "factors" $\mathbf{f}_j$ have d components
- matrix $\mathbf{B}\mathbf{D}^{\frac{1}{2}}$ has orthogonal columns usually unique up to signs and order

Factor Analysis

$$\mathbf{X}_j = \boldsymbol{\mu} + \mathbf{A}\mathbf{f}_j + \boldsymbol{\epsilon}_j$$

- has "error" term $\boldsymbol{\epsilon}_j$ that depends on natural coordinates
- factors $\mathbf{f}_j$ have $m \ll d$ components
- columns of $\mathbf{A}$ not necessarily orthogonal and only unique up to right rotation

Application to Principal Components

Principal components analysis does not "need" any rotations: the principal components are essentially unique (except when there are equal eigenvalues, but this never occurs for empirical PCA).

Some reasons we might still want to rotate:

- because coordinate directions are easier to interpret than the principal components in some cases
- to detect or separate outliers, when only a small number of coordinates are affected
- because some principal components are ill-defined (have nearly equal eigenvalues); see Jolliffe (1989)

We might also decide to rotate if we have already chosen our coordinate directions carefully to match structure that we *a priori* suspect is in the data.

Application to Principal Components

The dilemma: what do we rotate in the context of principal components?

- If we rotate the matrix $\mathbf{B}\mathbf{D}^{\frac{1}{2}}$, we have

$$\mathbf{B}\mathbf{D}^{\frac{1}{2}}\mathbf{f}_j = (\mathbf{B}\mathbf{D}^{\frac{1}{2}}\mathbf{R})(\mathbf{R}^{\mathsf{T}}\mathbf{f}_j)$$

  - so we still have $\mathbf{R}^{\mathsf{T}}\mathbf{f}_j \sim (\mathbf{0}, \mathbf{I}_d)$, but
  - columns of $\mathbf{B}\mathbf{D}^{\frac{1}{2}}\mathbf{R}$ are not orthogonal.

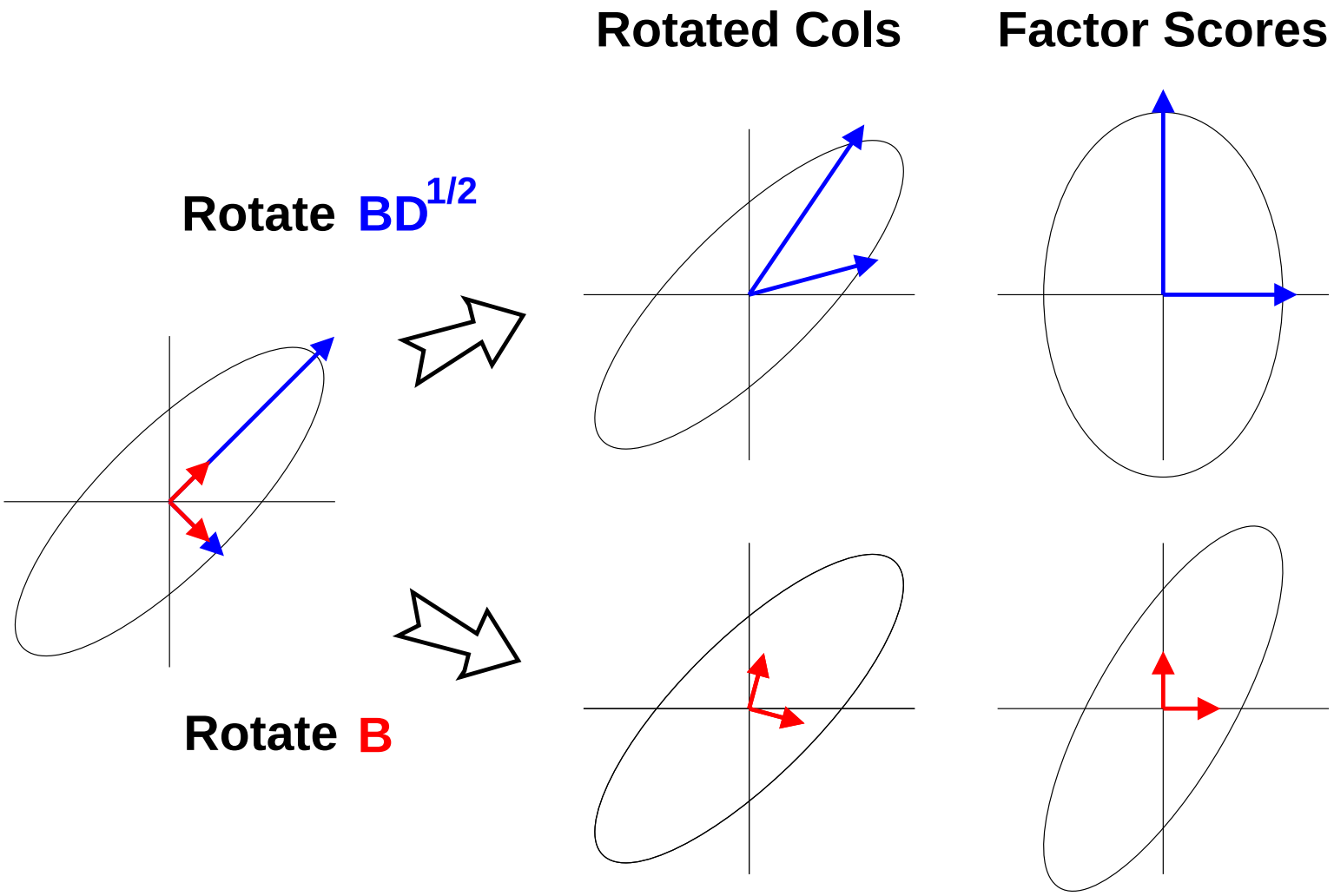
- If we rotate the matrix $\mathbf{B}$ only, we have

$$\mathbf{B}\mathbf{D}^{\frac{1}{2}}\mathbf{f}_j = (\mathbf{B}\mathbf{R})(\mathbf{R}^{\mathsf{T}}\mathbf{D}^{\frac{1}{2}}\mathbf{f}_j)$$

  - so columns of $\mathbf{B}\mathbf{R}$ are orthonormal, but
  - we have $\mathbf{R}^{\mathsf{T}}\mathbf{D}^{\frac{1}{2}}\mathbf{f}_j \sim (\mathbf{0}, \mathbf{R}^{\mathsf{T}}\mathbf{D}\mathbf{R})$, so "factors" are no longer uncorrelated.

In general, we will only be interested in rotating a *subset* of the principal components. Think about what would happen if we rotated *all* columns of an orthogonal matrix optimally under, say, VARIMAX; the result would just be a diagonal matrix, so not very interesting.

Application to Principal Components



The blue arrows represent the columns of $\mathbf{BD}^{\frac{1}{2}}$, while the red arrows represent the columns of $\mathbf{B}$.

The ellipses represent the structure of the data point cloud, either in the original space (left and center), or in terms of their factor score coordinates (right).

Application to Principal Components

Ramsay & Silverman only consider the case of rotating $\mathbf{B}$. As noted previously, in this case VARIMAX and QUARTIMAX are equivalent, since $\mathbf{B}$ is orthogonal.

If we rotate $\mathbf{B}\mathbf{D}^{\frac{1}{2}}$ instead, then VARIMAX and QUARTIMAX are *not* equivalent rotation criteria, and will generally give different results.

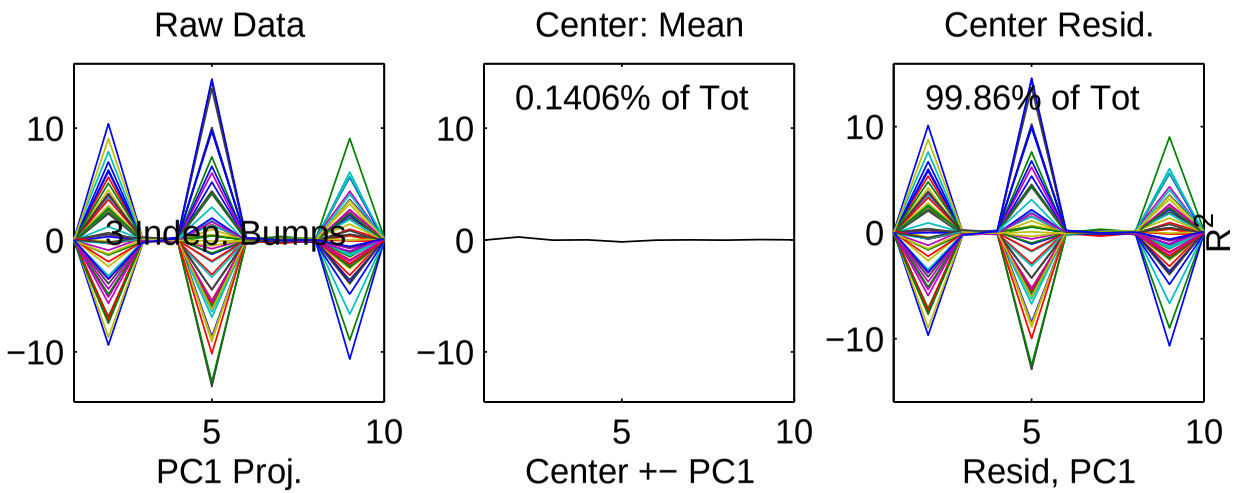
Note that we can mitigate this problem by only rotating subsets of principal components that have nearly equal eigenvalues (as suggested by Jolliffe (1989)).

Finally, note that, after rotation using VARIMAX or QUARTIMAX, the resulting analysis is no longer coordinate-free; the optimal rotations are tied to the coordinate system being used.

The moral of the story: be careful what you rotate, what criterion you use for rotation, and what coordinate system you choose to work with.

Examples

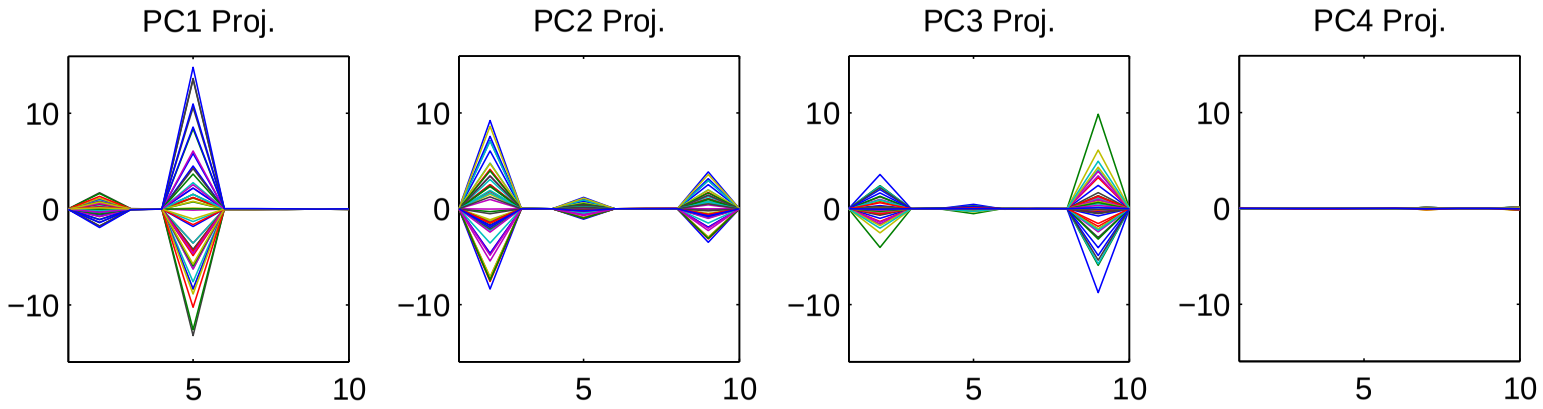
Back to the "Three Independent Bump" example ...



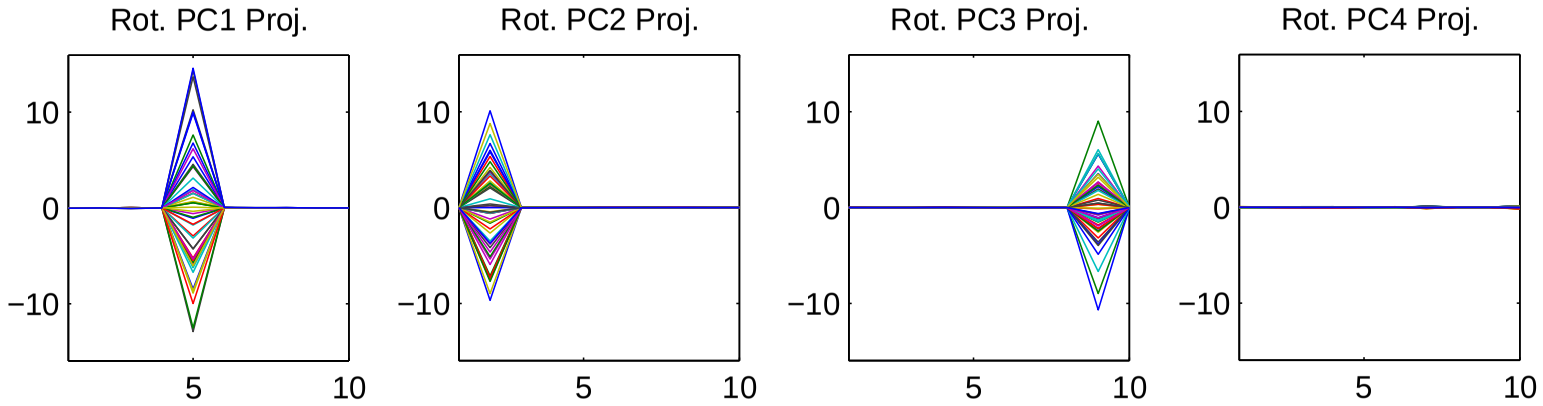
Suppose we rotate the first few principal component *directions* according to the VARIMAX (or equivalently QUARTIMAX) criterion. (So we are rotating the first few columns of the $\mathbf{B}$ matrix.)

Examples

The original principal components ...



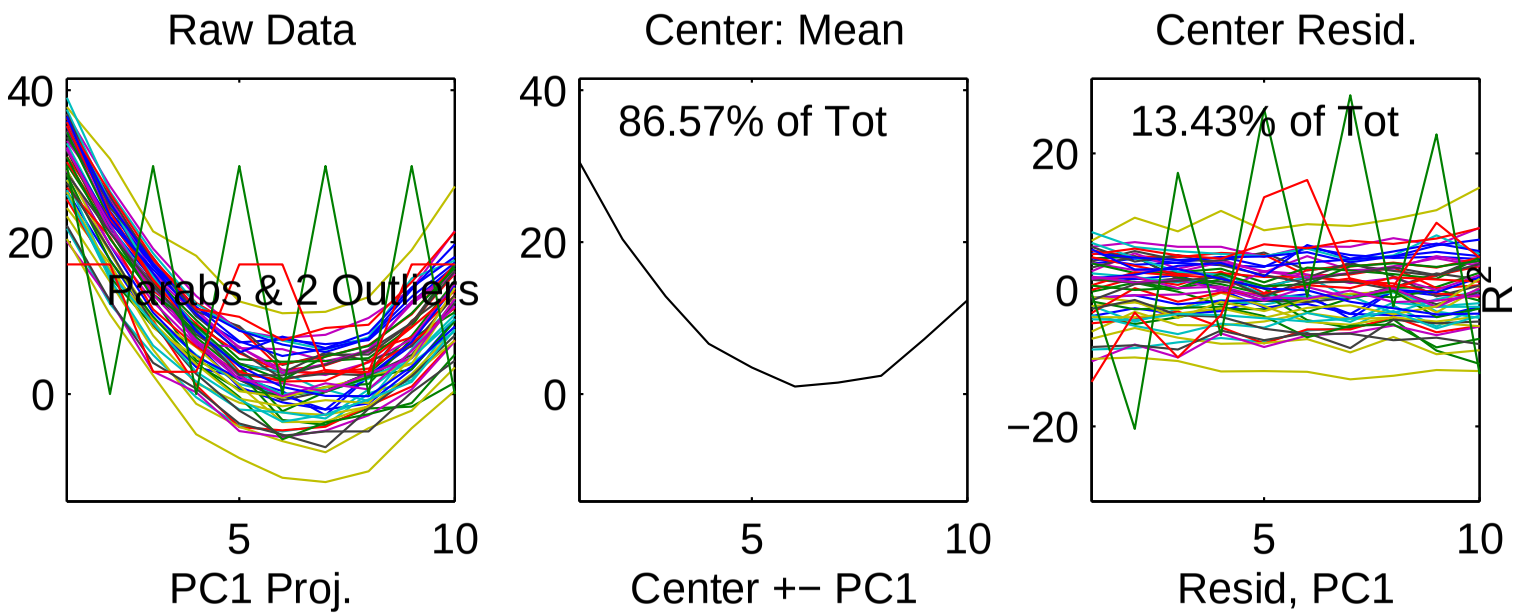
... and the VARIMAX-rotated principal components:



The VARIMAX rotation has apparently very cleanly separated the three bumps.

Examples

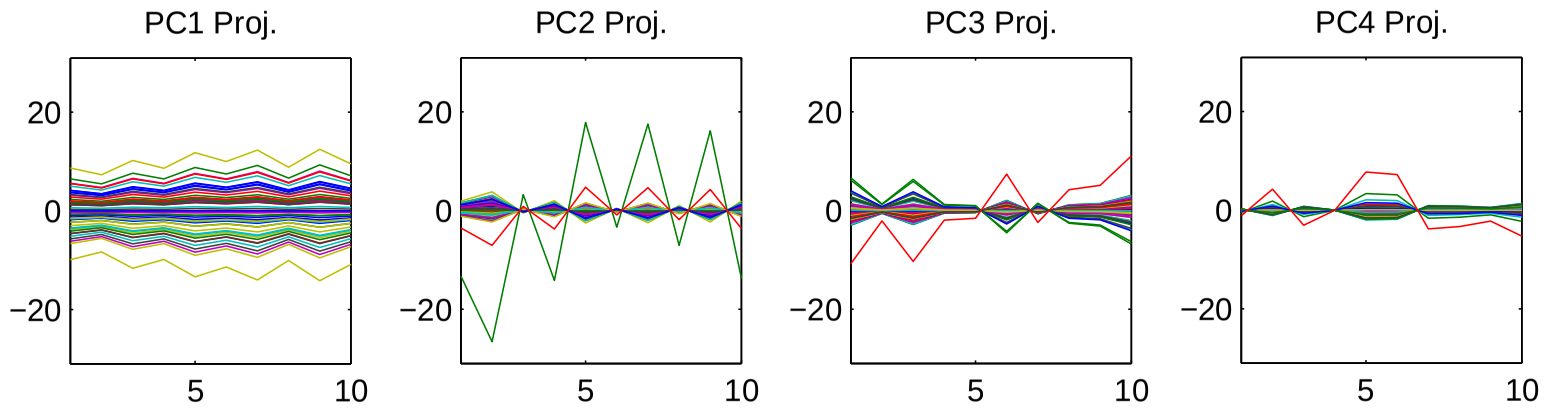
Here is the "Parabolas with Two Outliers" example ...



We want our components to separate the outliers from the main body of the parabola data so that the hidden structure of the parabola data is more apparent.

Examples

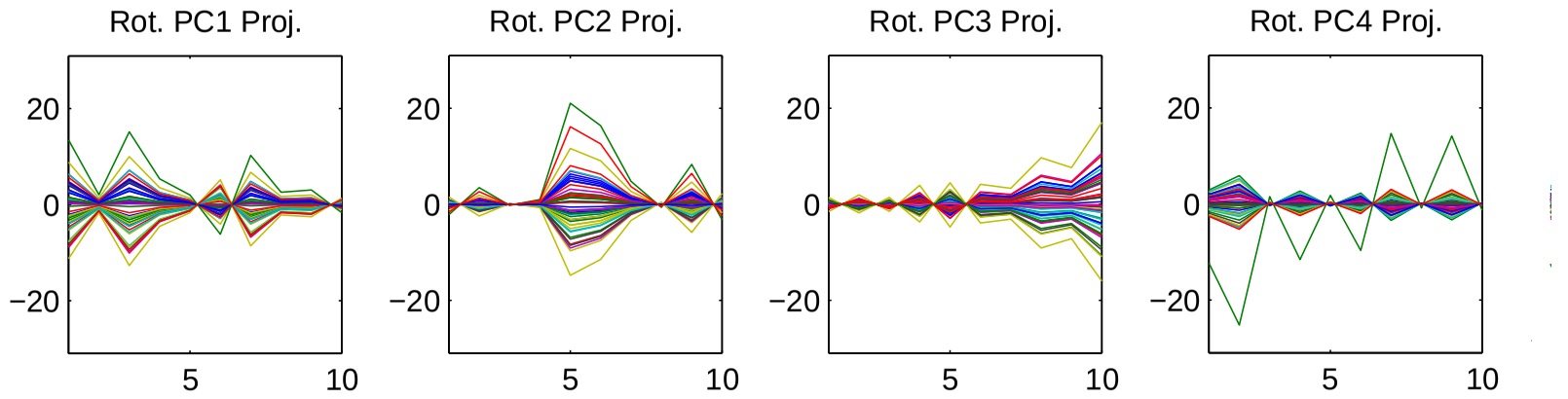
However, the first few principal components seem to mix up the main data and the outliers very severely:



The vertical shift component has "wobbles," and the tilt component is not to be found. PC2 captures the first outlier, but the second is spread over PCs 2, 3, and 4.

Examples

Applying VARIMAX rotation to the principal component directions with respect to the natural basis just makes things worse:



The problem is that neither the outliers nor the structure in the parabola curves is very much like any of the natural basis directions: They are spread over all of the points of the curve.

Examples

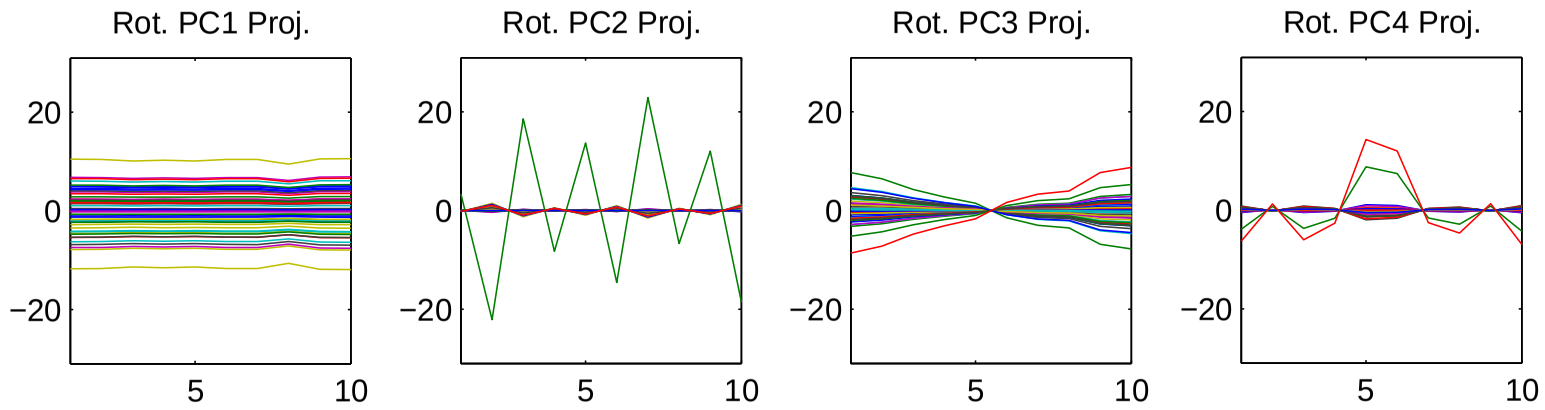
It would be nice if we used a basis that more clearly separates the parabolas and the outliers.

Observation: The outliers seem to be at a higher *frequency* than the parabola structure.

So perhaps we can separate them with something like a Fourier basis. For convenience, and to keep the computations real, we will use a Fourier-like basis called a *discrete cosine basis*. Basically, it is a bunch of cosine-like curves that vary from low frequency to high frequency.

Examples

Applying VARIMAX rotation with respect to the discrete cosine basis yields the results we want:



The outliers are nicely separated from the parabola data structure, and the hidden tilt component is now visible. (Interestingly, the analysis has failed to separate the outliers from *each other*.)

References

Harman, H.H. (1976), *Modern Factor Analysis* (3^{rd} ed.), University of Chicago Press: Chicago.

Jolliffe, I.T. (1989), "Rotation of Ill-defined Principal Components," *Applied Statistics*, 38, 139–147.

Kaiser, H.F. (1958), "The Varimax Criterion for Analytic Rotation in Factor Analysis," *Psychometrika*, 23, 187–200.

Ramsay, J.O., and Silverman, B.W. (1997), *Functional Data Analysis*, New York: Springer.

Seber, G.A.F. (1984), *Multivariate Observations*, New York: Wiley.