Linear Dimensionality Reduction

Introduction

When faced with situations involving high-dimensional data, it is natural to consider the possibility of projecting those data onto a lower-dimensional subspace without losing important information regarding some characteristic of the original variables.

One way of accomplishing this reduction of dimensionality is through variable selection, also called **feature selection**.

Another way is by creating a reduced set of linear or nonlinear transformations of the input variables. The creation of such composite variables (or features) by projection methods is often referred to as **feature extraction**. Usually, we wish to find those low-dimensional projections of the input data that enjoy some sort of optimality properties.

Early examples of projection methods were linear methods such as principal component analysis (PCA) and canonical variate and correlation analysis (CVA or CCA), and these have become two of the most popular dimensionality-reducing techniques in use today.

Both PCA and CVA are, at heart, eigenvalue-eigenvector problems. Furthermore, both can be viewed as special cases of multivariate reduced-rank regression. This latter connection to regression is fortuitous.

Whereas PCA and CVA were once regarded as isolated statistical tools, their now being part of such a well-traveled tool as regression means that we should be able to carry out feature selection and extraction, as well as outlier detection within an integrated framework.

Principal Component Analysis

Principal component analysis (PCA) was introduced as a technique for deriving a reduced set of orthogonal linear projections of a single collection of correlated variables, $\mathbf{X} = (X_1, \ldots, X_r)^T$, where the projections are ordered by decreasing variances.

Variance is a second-order property of a random variable and is an important measurement of the amount of information in that variable.

PCA has also been referred to as a method for decorrelating $\mathbf{X}$; as a result, the technique has been independently rediscovered by many different fields, with alternative names such as **Karhunen-Loève transform** and **empirical orthogonal functions**, which are used in communications theory and atmospheric sciences, respectively.

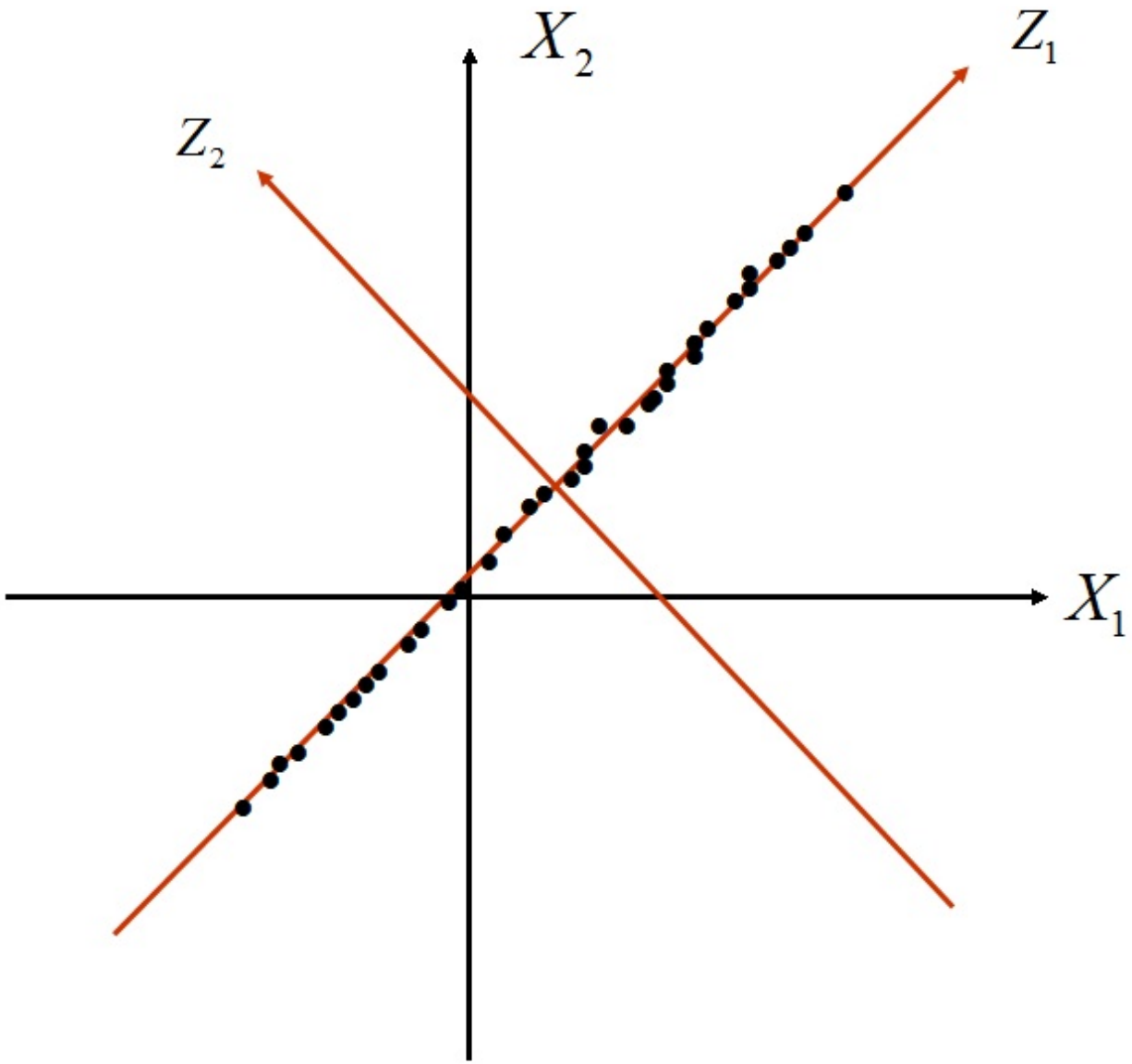
PCA is used primarily as a dimensionality-reduction technique.

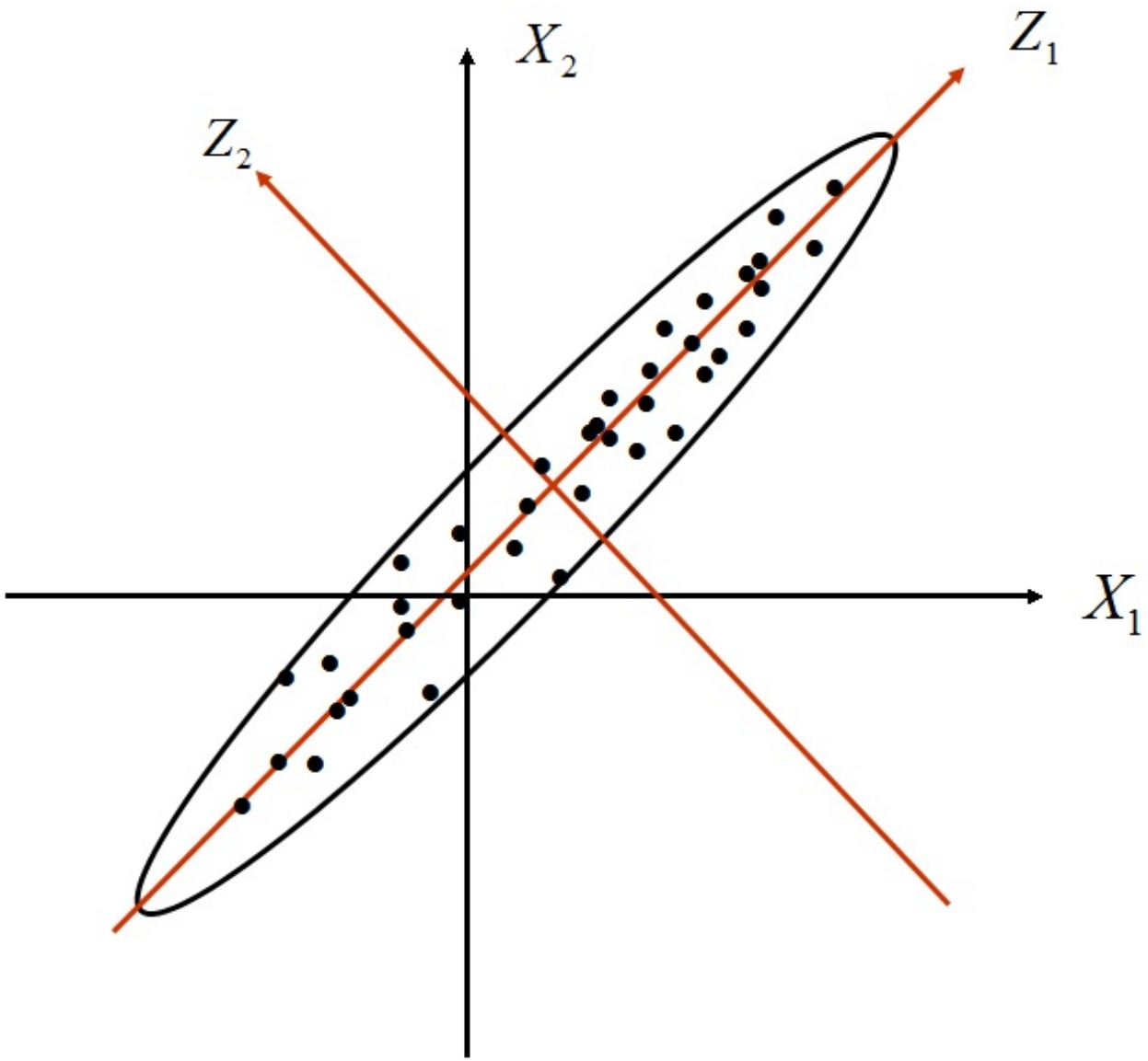
In this role, PCA is used, for example, in lossy data compression, pattern recognition, and image analysis.

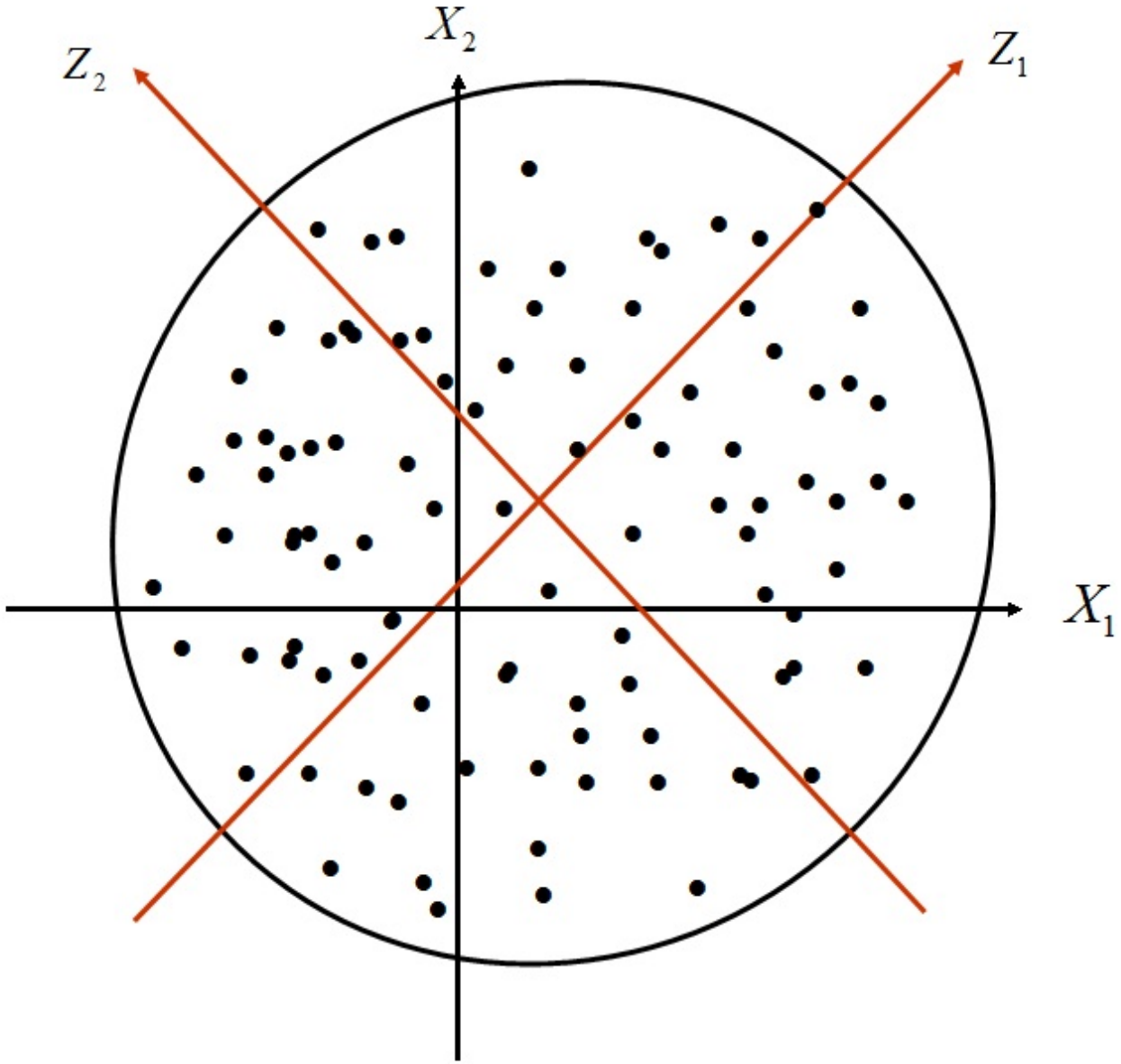
In addition to reducing dimensionality, PCA can be used to discover important features of the data. Discovery in PCA takes the form of graphical displays of the principal component scores.

The first few principal component scores can reveal whether most of the data actually live on a linear subspace of $\mathcal{R}^r$ and can be used to identify outliers, distributional peculiarities, and clusters of points.

The last few principal component scores show those linear projections of $\mathbf{X}$ that have smallest variance; any principal component with zero or near-zero variance is virtually constant, and, hence, can be used to detect collinearity, as well as outliers that pop up and alter the perceived dimensionality of the data.


X_2
Z_1
Z_2
X_1


X_2
Z_1
Z_2
X_1


X_2
Z_2
Z_1
X_1

Example: The Nutritional Value of Food

See the textbook.

Population Principal Components

Assume that the random r-vector $\mathbf{X} = (X_1, \ldots, X_r)^T$ has mean $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_{XX}$.

PCA seeks to replace the set of r (unordered and correlated) input variables, $X_1, X_2, \ldots, X_r$, by a (potentially smaller) set of t (ordered and uncorrelated) linear projections, $\xi_1, \ldots, \xi_t$ $(t \leq r)$, of the input variables,

$$\xi_j = \mathbf{b}_j^T \mathbf{X} = b_{j1} X_1 + \cdots + b_{jr} X_r, \quad j = 1, \ldots, t,$$

where we minimize the loss of information due to replacement.

In PCA, information is interpreted as the total variation of the original input variables,

$$\sum_{j=1}^{r} \text{Var}(X_j) = \text{tr}(\boldsymbol{\Sigma}_{XX}).$$

From the spectral decomposition theorem, we can write

$$\mathbf{\Sigma}_{XX} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}_r,$$

where the diagonal matrix $\mathbf{\Lambda}$ has diagonal elements the eigenvalues, $\{\lambda_j\}$, of $\mathbf{\Sigma}_{XX}$, and the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{\Sigma}_{XX}$.

Thus, the total variation is

$$\text{tr}(\mathbf{\Sigma}_{XX}) = \text{tr}(\mathbf{\Lambda}) = \sum_{j=1}^{r} \lambda_j.$$

The jth coefficient vector, $\mathbf{b}_j = (b_{j1}, \ldots, b_{jr})^T$, is chosen so that:

- The first t linear projections ξ_j, $j = 1, \ldots, t$, of $\mathbf{X}$ are ranked in importance through their variances $\{\text{Var}(\xi_j)\}$, which are listed in decreasing order of magnitude: $\text{Var}(\xi_1) \geq \text{Var}(\xi_2) \geq \cdots \geq \text{Var}(\xi_t)$.

- ξ_j is uncorrelated with all ξ_k, $k < j$.

The linear projections are then known as the first t principal components of $\mathbf{X}$.

There are two popular derivations of the set of principal components of X:

- PCA can be derived using a least-squares optimality criterion;
- it can be derived as a variance-maximizing technique.

Least-Squares Optimality of PCA

Let $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_t)^T$ be a $t \times r$-matrix of weights ($t \leq r$).

The linear projection can be written as a t-vector,

$$\boldsymbol{\xi} = \mathbf{BX}.$$

- $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_t)^T$.

We want to find an r-vector $\boldsymbol{\mu}$ and an $r \times t$ matrix $\mathbf{A}$ such that the projections $\boldsymbol{\xi}$ have the property that $\mathbf{X} \approx \boldsymbol{\mu} + \mathbf{A}\boldsymbol{\xi}$ in some least-square sense.

We use the least-squares error criterion,

$$\mathrm{E}\{(\mathbf{X} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\xi})^T(\mathbf{X} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\xi})\},$$

as our measure of how well we can reconstruct $\mathbf{X}$ by the linear projection $\boldsymbol{\xi}$.

We can write the criterion in a more transparent manner by substituting $\mathbf{BX}$ for ξ.

The criterion is now a function of an $r \times t$-matrix $\mathbf{A}$ and a $t \times r$-matrix $\mathbf{B}$ (both of full rank t), and an r-vector μ.

The goal is to choose $\mathbf{A}$, $\mathbf{B}$, and μ to minimize

$$\mathrm{E}\big\{(\mathbf{X} - \mu - \mathbf{ABX})^T(\mathbf{X} - \mu - \mathbf{ABX})\big\}.$$

For example, when $t = 1$, we can write the least-squares problem,

$$\min_{\boldsymbol{\mu},\mathbf{A},\mathbf{B}} \mathrm{E} \sum_{j=1}^{r} (X_j - \mu_j - a_{j1}\mathbf{b}_1^T\mathbf{X})^2.$$

- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_r)^T$;
- $\mathbf{A} = \mathbf{a}_1 = (a_{11}, \ldots, a_{r1})^T$;
- $\mathbf{B} = \mathbf{b}_1 = (b_{11}, \ldots, b_{1t})^T$.

The criterion can be minimized by the reduced-rank regression solution,

$$\mathbf{A}^{(t)} = (\mathbf{v}_1, \ldots, \mathbf{v}_t) = \mathbf{B}^{(t)T}, \quad \boldsymbol{\mu}^{(t)} = (\mathbf{I}_r - \mathbf{A}^{(t)}\mathbf{B}^{(t)})\boldsymbol{\mu}_X.$$

- $\mathbf{v}_j = \mathbf{v}_j(\boldsymbol{\Sigma}_{XX})$ is the eigenvector associated with the jth largest eigenvalue, λ_j, of $\boldsymbol{\Sigma}_{XX}$.

Thus, our best rank-t approximation to the original $\mathbf{X}$ is given by

$$\hat{\mathbf{X}}^{(t)} = \boldsymbol{\mu}^{(t)} + \mathbf{C}^{(t)}\mathbf{X} = \boldsymbol{\mu}_X + \mathbf{C}^{(t)}(\mathbf{X} - \boldsymbol{\mu}_X),$$

- $\mathbf{C}^{(t)} = \mathbf{A}^{(t)}\mathbf{B}^{(t)} = \sum_{j=1}^{t} \mathbf{v}_j \mathbf{v}_j^T$ is the reduced-rank regression coefficient matrix with rank t for the principal components case.

The minimum is given by $\sum_{j=t+1}^{r} \lambda_j$, the sum of the smallest $r-t$ eigenvalues of $\boldsymbol{\Sigma}_{XX}$.

It may be helpful to think of these results in the following way.

Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$ be the $r \times r$-matrix whose columns are the complete set of r ordered eigenvectors of Σ_{XX}. It can be shown that the most accurate rank-t least-squares reconstruction of $\mathbf{X}$ can be obtained by using the composition of two linear maps $L' \circ L$.

The first map $L : \mathcal{R}^r \to \mathcal{R}^t$ takes the first t columns of $\mathbf{V}$ to form t linear projections of $\mathbf{X}$, and then the second $L' : \mathcal{R}^t \to \mathcal{R}^r$ uses those same t columns of $\mathbf{V}$ to carry out a linear reconstruction of $\mathbf{X}$ from those projections.

The first t principal components of $\mathbf{X}$ are given by the linear projections, $\xi_1, \ldots, \xi_t$, where

$$\xi_j = \mathbf{v}_j^T \mathbf{X}, \quad j = 1, 2, \ldots, t.$$

The covariance between ξ_i and ξ_j is

$$\text{cov}(\xi_i, \xi_j) = \text{cov}(\mathbf{v}_i^T \mathbf{X}, \mathbf{v}_j^T \mathbf{X}) = \mathbf{v}_i^T \Sigma_{XX} \mathbf{v}_j = \lambda_j \mathbf{v}_i^T \mathbf{v}_j = \delta_{ij} \lambda_j.$$

- δ_{ij} is the Kronecker delta, which equals 1 if $i = j$ and zero otherwise.

Thus,

- λ_1, the largest eigenvalue of $\mathbf{\Sigma}_{XX}$, is the $\text{Var}(\xi_1)$;
- λ_2, the second-largest eigenvalue of $\mathbf{\Sigma}_{XX}$, is the $\text{Var}(\xi_2)$;
- and so on, while all pairs of derived variables are uncorrelated, $\text{cov}(\xi_i, \xi_j) = 0$, $i \neq j$.

A goodness-of-fit measure of how well the first t principal components represent the r original variables in the lower-dimensional space is given by the ratio

$$\frac{\lambda_{t+1} + \cdots + \lambda_r}{\lambda_1 + \cdots + \lambda_r}$$

which is the proportion of the total variation in the input variables that is explained by the last $r - t$ principal components.

If the first t principal components explain a large proportion of the total variation in $\mathbf{X}$, then the ratio should be small.

PCA as a Variance-Maximization Technique

In the original derivation of principal components, the coefficient vectors,

$$\mathbf{b}_j = (b_{j1}, b_{j2}, \ldots, b_{jr})^T, \quad j = 1, \ldots, t,$$

were chosen in a sequential manner so that

- the variances of the derived variables, $\mathrm{Var}(\xi_j) = \mathbf{b}_j^T \boldsymbol{\Sigma}_{XX} \mathbf{b}_j$, are arranged in descending order subject to the normalizations $\mathbf{b}_j^T \mathbf{b}_j = 1$, $j = 1, \ldots, t$;
- they are uncorrelated with previously chosen derived variables, i.e., $\mathrm{cov}(\xi_i, \xi_j) = \mathbf{b}_i^T \boldsymbol{\Sigma}_{XX} \mathbf{b}_j = 0$, $i < j$.

The first principal component, ξ_1, is obtained by choosing the r coefficients, $\mathbf{b}_1$, for the linear projection ξ_1, so that the variance of ξ_1 is a maximum.

A unique choice of $\{\xi_j\}$ is obtained through the normalization constraint $\mathbf{b}_j^T\mathbf{b}_j = 1$, for all $j = 1, 2, \ldots, t$.

The objective function is

$$f(\mathbf{b}_1) = \mathbf{b}_1^T\boldsymbol{\Sigma}_{XX}\mathbf{b}_1 - \lambda_1(1 - \mathbf{b}_1^T\mathbf{b}_1).$$

- λ_1 is a Lagrangian multiplier.

Differentiating $f(\mathbf{b}_1)$ with respect to $\mathbf{b}_1$ and setting the result equal to zero for a maximum yields

$$\frac{\partial f(\mathbf{b}_1)}{\partial \mathbf{b}_1} = 2(\mathbf{\Sigma}_{XX} - \lambda_1 \mathbf{I}_r)\mathbf{b}_1 = \mathbf{0}.$$

This is a set of r simultaneous equations.

If $\mathbf{b}_1 \neq \mathbf{0}$, then λ_1 must be chosen to satisfy the determinantal equation

$$|\mathbf{\Sigma}_{XX} - \lambda_1 \mathbf{I}| = 0.$$

Thus, λ_1 has to be the largest eigenvalue of $\mathbf{\Sigma}_{XX}$, and $\mathbf{b}_1$ be the eigenvector associated with λ_1.

Question

Why has to be the largest?

The second principal component, ξ_2, is then obtained by choosing a second set of coefficients, $\mathbf{b}_2$, for the next linear projection, ξ_2, so that the variance of ξ_2 is largest among all linear projections of $\mathbf{X}$ that are also uncorrelated with ξ_1 above.

The variance of ξ_2 is $\text{Var}(\xi_2) = \mathbf{b}_2^T \mathbf{\Sigma}_{XX} \mathbf{b}_2$, and this has to be maximized subject to the normalization constraint $\mathbf{b}_2^T \mathbf{b}_2 = 1$ and orthogonality constraint $\mathbf{b}_1^T \mathbf{b}_2 = 0$.

From the function

$$f(\mathbf{b}_1) = \mathbf{b}_2^T \mathbf{\Sigma}_{XX} \mathbf{b}_2 - \lambda_2(1 - \mathbf{b}_2^T \mathbf{b}_2) + \mu \mathbf{b}_1^T \mathbf{b}_2.$$

- λ_2 and μ are the Lagrangian multipliers.

Differentiating $f(\mathbf{b}_2)$ with respect to $\mathbf{b}_2$ and setting the result equal to zero for a maximum yields

$$\frac{\partial f(\mathbf{b}_2)}{\partial \mathbf{b}_2} = 2(\mathbf{\Sigma}_{XX} - \lambda_2 \mathbf{I}_r)\mathbf{b}_2 + \mu \mathbf{b}_1 = \mathbf{0}.$$

Premultiplying this derivative by $\mathbf{b}_1^T$ and using the orthogonality and normalization constraints, we have that $2\mathbf{b}_1^T\mathbf{\Sigma}_{XX}\mathbf{b}_2 + \mu = 0$.

Premultiplying the equation $(\mathbf{\Sigma}_{XX} - \lambda_1\mathbf{I}_r)\mathbf{b}_1 = 0$ by $\mathbf{b}_2^T$ yields $\mathbf{b}_2^T\mathbf{\Sigma}_{XX}\mathbf{b}_1 = 0$, whence $\mu = 0$.

This means that λ_2 is the second largest eigenvalue of $\mathbf{\Sigma}_{XX}$, and the coefficient vector $\mathbf{b}_2$ for the second principal component is the eigenvector associated with λ_2.

In this sequential manner, we obtain the remaining sets of coefficients for the principal components $\xi_3, \xi_4, \ldots, \xi_r$, where the ith principal component ξ_i is obtained by choosing the set of coefficients, $\mathbf{b}_i$, for the linear projection ξ_i so that ξ_i has the largest variance among all linear projections of $\mathbf{X}$ that are also uncorrelated with $\xi_1, \xi_2, \ldots, \xi_{i-1}$.

The coefficients of these linear projections are given by the ordered sequence of eigenvectors $\{\mathbf{b}_j\}$ associated with the jth largest eigenvalue λ_j, of $\mathbf{\Sigma}_{XX}$.

Sample Principal Components

In practice, we estimate the principal components using n independent observations, $\{\mathbf{X}_i,\ i = 1, 2, \ldots, n\}$, on $\mathbf{X}$.

We estimate $\boldsymbol{\mu}_X$ by

$$\hat{\boldsymbol{\mu}}_X = \bar{\mathbf{X}} = n^{-1} \sum_{i=1}^{n} \mathbf{X}_i.$$

As before, let $\mathbf{X}_{ci} = \mathbf{X}_i - \bar{\mathbf{X}}$, $i = 1, 2, \ldots, n$, and set $\mathcal{X}_c = (\mathbf{X}_{c1}, \ldots, \mathbf{X}_{cn})$ to be an $r \times n$ matrix.

We estimate $\mathbf{\Sigma}_{XX}$ by the sample covariance matrix,

$$\hat{\mathbf{\Sigma}}_{XX} = n^{-1}\mathbf{S} = n^{-1}\mathcal{X}_c\mathcal{X}_c^T.$$

The ordered eigenvalues of $\hat{\mathbf{\Sigma}}_{XX}$ are denoted by $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_r \geq 0$, and the eigenvector associated with the jth largest sample eigenvalue $\hat{\lambda}_j$ is the jth sample eigenvector $\hat{\mathbf{v}}_j$, $j = 1, \ldots, r$.

We estimate $\mathbf{A}^{(t)}$ and $\mathbf{B}^{(t)}$ by

$$\mathbf{A}^{(t)} = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_t) = \hat{\mathbf{B}}^{(t)T}.$$

- $\hat{\mathbf{v}}_j$ is the jth sample eigenvector of $\hat{\Sigma}_{XX}$, $j = 1, 2, \ldots, t$ $(t \leq r)$.

The best rank-t reconstruction of $\mathbf{X}$ is given by

$$\hat{\mathbf{X}}^{(t)} = \bar{\mathbf{X}} + \hat{\mathbf{C}}^{(t)}(\mathbf{X} - \bar{\mathbf{X}}),$$

where

$$\hat{\mathbf{C}}^{(t)} = \hat{\mathbf{A}}^{(t)}\hat{\mathbf{B}}^{(t)} = \sum_{j=1}^{t} \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^T$$

is the reduced-rank regression coefficient matrix corresponding to the principal components case.

The jth sample PC score of $\mathbf{X}$ is given by

$$\hat{\xi}_j = \hat{\mathbf{v}}_j^T \mathbf{X}_c, \qquad \mathbf{X}_c = \mathbf{X} - \bar{\mathbf{X}}.$$

The variance, λ_j, of the jth principal component is estimated by the sample variance $\hat{\lambda}_j$, $j = 1, 2, \ldots, t$.

A sample estimate of the measure of how well the first t principal components represent the r original variables is given by the statistic

$$\frac{\hat{\lambda}_{t+1} + \cdots + \hat{\lambda}_r}{\hat{\lambda}_1 + \cdots + \hat{\lambda}_r},$$

which is the proportion of the total sample variation that is explained by the last $r - t$ sample principal components.

It is hoped that the sample variances of the first few sample PCs will be large, whereas the rest will be small enough for the corresponding set of sample PCs to be omitted.

A variable that does not change much (relative to other variables) in independent measurements may be treated approximately as a constant, and so omitting such low-variance sample PCs and putting all attention on high-variance sample PCs is, therefore, a convenient way of reducing the dimensionality of the data set.

How Many Principal Components to Retain?

Probably the main question asked while carrying out a PCA is how many principal components to retain.

Because the criterion for a good projection in PCA is a high variance for that projection, we should only retain those principal components with large variances.

The question, therefore, boils down to one involving the magnitudes of the eigenvalues of Σ_{XX}: How small can an eigenvalue be while still regarding the corresponding principal component as significant?

Scree Plot

The sample eigenvalues from a PCA are ordered from largest to smallest.

It is usual to plot the ordered sample eigenvalues against their order number; such a display is called a scree plot, after the break between a mountainside and a collection of boulders usually found at its base.

If the largest few sample eigenvalues dominate in magnitude, with the remaining sample eigenvalues very small, then the scree plot will exhibit an elbow in the plot corresponding to the division into large and small values of the sample eigenvalues.

The order number at which the elbow occurs can be used to determine how many principal components to retain.

It is usually recommended to retain those PCs up to the elbow and also the first PC following the elbow.

A related popular criterion for use when an elbow may not be present in the scree plot is to use a cutoff point of 90% of total variance.

PC Rank Trace

The problem of deciding how many principal components to retain is equivalent to obtaining a useful estimate of the rank of the regression coefficient matrix $\mathbf{C}$ in the principal components case.

So, if we can obtain a good estimate of the rank, we should have a solution to this problem.

In the principal components case, the expressions for the points in the rank trace simplify greatly and are very simple to compute.

It is not difficult to show that

$$\Delta\hat{\mathbf{C}}^{(t)} = \left(1 - \frac{t}{r}\right)^{1/2},$$

$$\Delta\hat{\boldsymbol{\Sigma}}_{\varepsilon\varepsilon}^{(t)} = \left(\frac{\hat{\lambda}_{t+1}^2 + \cdots + \hat{\lambda}_r^2}{\hat{\lambda}_1^2 + \cdots + \hat{\lambda}_r^2}\right)^{1/2}.$$

A plot of $\Delta\hat{\boldsymbol{\Sigma}}_{\varepsilon\varepsilon}^{(t)}$ against $\Delta\hat{\mathbf{C}}^{(t)}$ is called a PC rank trace plot.

All the information regarding the dimensionality of the regression is, therefore, contained in the residual covariance matrices and not in the regression coefficients. Furthermore, the $r+1$ plotted points decrease monotonically from $(1,1)$ to $(0,0)$.

We assess the rank t of $\mathbf{C}$ by $\hat{t}$, the smallest integer value between 1 and r at which an elbow can be detected in the PC rank trace plot.

Kaiser's Rule

When dealing with the PCA of a sample correlation matrix, Kaiser (1960) suggested (in the context of exploratory factor analysis) that only those principal components be retained whose eigenvalues exceed unity.

This decision guideline is based upon the argument that because the total variation of all r standardized variables is equal to r, it follows that a principal component should account for at least the average variation of a single standardized variable.

This rule is popular but controversial; there is evidence that the cutoff value of 1 is too high.

A modified rule retains all PCs whose eigenvalues of the sample correlation matrix exceed 0.7.

Graphical Displays

For diagnostic and data analytic purposes, it is usual to plot the first sample PC scores against the second sample PC scores,

$$(\hat{\xi}_{i1}, \hat{\xi}_{i2}), \quad i = 1, 2, \ldots, n.$$

- $\hat{\xi}_{ij} = \hat{\mathbf{v}}_j^T \mathbf{X}_i$, $i = 1, 2, \ldots, n$, $j = 1, 2$.

A more general graphical tool for displaying the sample PC scores associated with the largest few sample eigenvalues (variances) is the scatter plot matrix, in which all possible pairs of variables are plotted in two dimensions.

A three-dimensional scatterplot of the first three sample PC scores is also strongly recommended, especially if a brush and spin feature is available.

Example: Face Recognition Using Eigenfaces

See the textbook.

Invariance and Scaling

A shortcoming of PCA is that the principal components are not invariant under rescalings of the initial variables.

In other words, a PCA is sensitive to the units of measurement of the different input variables.

Standardizing (centering and then scaling) the **X**-variables,

$$\mathbf{Z} \leftarrow (\text{diag}\{\hat{\Sigma}_{XX}\})^{-1/2}(\mathbf{X} - \hat{\mu}_X),$$

is equivalent to carrying out PCA using the correlation (rather than the covariance) matrix.

When using the correlation matrix, the total variation of the standardized variables is r, the trace of the correlation matrix.

The lack of scale invariance implies that a PCA using the correlation matrix may be very different from a similar analysis using the corresponding covariance matrix, and no simple relationship exists between the two sets of results.

In the initial formulation and application of PCA, we note that Hotelling (1933) who was dealing with a battery of test scores, extracted principal components from the correlation matrix of the data.

Standardization in the PCA context has its advantages.

In some fields, standardization is customary.

In heterogeneous situations, where the units of measurement of the input variables are not commensurate or the ranges of values of the variables differ considerably, standardization is especially relevant.

If the variables have heterogeneous variances, it is a good idea to standardize the variables before carrying out PCA because the variables with the greatest variances will tend to overwhelm the leading principal components with the remaining variables contributing very little.

On statistical inference grounds, standardization is usually regarded as a nuisance because it complicates the distributional theory.

Indeed, the asymptotic distribution theory for the eigenvalues and eigenvectors of a sample correlation matrix turns out to be extremely difficult to derive.

Furthermore, certain simplifications, such as pretending that the sample correlation matrix has the same distributional properties as the sample covariance matrix, tend not to work and, hence, lead to incorrect inference results for principal components.

What Can Be Gained from Using PCA?

The short answer is that it depends on what we are trying to accomplish and the nature of the application in question.

PCA is a linear technique built for several purposes: it enables us,

- to decorrelate the original variables in the study, regardless of whether $r < n$ or $n < r$;

- to carry out data compression, where we pay decreasing attention to the numerical accuracy by which we encode the sequence of principal components;

- to reconstruct the original input data using a reduced number of variables according to a least-squares criterion;
- to identify potential clusters in the data.

In certain applications, PCA can be misleading.

PCA is heavily influenced when there are outliers in the data (e.g., in computer vision, images can be corrupted by noisy pixels), and such considerations have led to the construction of robust PCA.

In other situations, the linearity of PCA may be an obstacle to successful data reduction and compression, we may further consider nonlinear versions of PCA.

Canonical Variate and Correlation Analysis

Canonical variate and correlation analysis (CVA or CCA) is a method for studying linear relationships between two vector variates, which we denote by $\mathbf{X} = (X_1, \ldots, X_r)^T$ and $\mathbf{Y} = (Y_1, \ldots, Y_s)^T$.

As such, it has been used to solve theoretical and applied problems in econometrics, business (primarily, finance and marketing), psychometrics, geography, education, ecology, and atmospheric sciences (e.g., weather prediction).

Hotelling applied CVA to the relationship between a set of two reading test scores ($X_1 =$ reading speed, $X_2 =$ reading power) and a set of two arithmetic test scores ($Y_1 =$ arithmetic speed, $Y_2 =$ arithmetic power) obtained from 140 fourth-grade children, so that $r = s = 2$.

Canonical Variates and Canonical Correlations

We assume that

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$$

is a collection of $r+s$ variables partitioned into two disjoint subcollections, where $\mathbf{X}$ and $\mathbf{Y}$ are jointly distributed with mean vector and covariance matrix given by

$$\mathsf{E}\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{pmatrix}, \quad \mathsf{E}\left\{ \begin{pmatrix} \mathbf{X}-\boldsymbol{\mu}_X \\ \mathbf{Y}-\boldsymbol{\mu}_Y \end{pmatrix} \begin{pmatrix} \mathbf{X}-\boldsymbol{\mu}_X \\ \mathbf{Y}-\boldsymbol{\mu}_Y \end{pmatrix}^T \right\} = \begin{pmatrix} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & \boldsymbol{\Sigma}_{YY} \end{pmatrix}.$$

CVA seeks to replace the two sets of correlated variables, $\mathbf{X}$ and $\mathbf{Y}$, by t pairs of new variables,

$$(\xi_i, \omega_i), \quad i = 1, 2, \ldots, t, \quad t \leq \min(r, s),$$

where

$$\begin{cases} \xi_j = \mathbf{g}_j^T \mathbf{X} = g_{1j} X_1 + g_{2j} X_2 + \cdots + g_{rj} X_r \\ \\ \omega_j = \mathbf{h}_j^T \mathbf{Y} = h_{1j} Y_1 + h_{2j} Y_2 + \cdots + h_{sj} Y_s \end{cases}$$

are linear projections of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

The jth pair of coefficient vectors, $\mathbf{g}_j = (g_{1j}, \ldots, g_{rj})^T$ and $\mathbf{h}_j = (h_{1j}, \ldots, h_{sj})^T$, are chosen so that

- the pairs $\{(\xi_j, \omega_j)\}$ are ranked in importance through their correlations,

$$\rho_j = \text{corr}\{\xi_j, \omega_j\} = \frac{\mathbf{g}_j^T \boldsymbol{\Sigma}_{XY} \mathbf{h}_j}{(\mathbf{g}_j^T \boldsymbol{\Sigma}_{XX} \mathbf{g}_j)^{1/2} (\mathbf{h}_j^T \boldsymbol{\Sigma}_{YY} \mathbf{h}_j)^{1/2}},$$

  which are listed in descending order of magnitude: $\rho_1 \geq \rho_2 \geq \cdots \geq \rho_t$.

- ξ_j is uncorrelated with all previously derived ξ_k:

$$\text{cov}\{\xi_j, \xi_k\} = \mathbf{g}_j^T \mathbf{\Sigma}_{XX} \mathbf{g}_k = 0, \quad k < j.$$

- ω_j is uncorrelated with all previously derived ω_k:

$$\text{cov}\{\omega_j, \omega_k\} = \mathbf{h}_j^T \mathbf{\Sigma}_{YY} \mathbf{h}_k = 0, \quad k < j.$$

The pairs are known as the first t pairs of canonical variates of $\mathbf{X}$ and $\mathbf{Y}$ and their correlations as the t largest canonical correlations.

The CVA technique ensures that every bit of correlation is wrung out of the original $\mathbf{X}$ and $\mathbf{Y}$ variables and deposited in an orderly fashion into pairs of new variables, (ξ_j, ω_j), $j = 1, 2, \ldots, t$, which have a special correlation structure.

If the notion of correlation is regarded as the primary determinant of information in the system of variables, then CVA is a major tool for reducing the dimensionality of the original two sets of variables.

Example: COMBO-17 Galaxy Photometric Catalogue

See the textbook.

Least-Squares Optimality of CVA

See the textbook.

CVA as a Correlation-Maximization Technique

Hotelling's approach to CVA maximized correlations between linear combinations of $\mathbf{X}$ and of $\mathbf{Y}$.

Consider, again, the arbitrary linear projections $\xi = \mathbf{g}^T\mathbf{X}$ and $\omega = \mathbf{h}^T\mathbf{Y}$, where, for the sake of convenience and with no loss of generality, we assume that $\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}_X = \mathbf{0}$ and $\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}_Y = \mathbf{0}$. Then, both ξ and ω have zero means.

We further assume that they both have unit variances; that is, $\mathbf{g}^T\boldsymbol{\Sigma}_{XX}\mathbf{g} = 1$ and $\mathbf{h}^T\boldsymbol{\Sigma}_{YY}\mathbf{h} = 1$.

The first step is to find the vectors $\mathbf{g}$ and $\mathbf{h}$ such that the random variables ξ and ω have maximal correlation,

$$\text{corr}(\xi, \omega) = \mathbf{g}^T \mathbf{\Sigma}_{XY} \mathbf{h},$$

among all such linear functions of $\mathbf{X}$ and $\mathbf{Y}$.

We set

$$f(\mathbf{g}, \mathbf{h}) = \mathbf{g}^T \boldsymbol{\Sigma}_{XY} \mathbf{h} - \frac{1}{2}\lambda(\mathbf{g}^T \boldsymbol{\Sigma}_{XX} \mathbf{g} - 1) - \frac{1}{2}\mu(\mathbf{h}^T \boldsymbol{\Sigma}_{YY} \mathbf{h} - 1),$$

where λ and μ are Lagrangian multipliers.

Differentiate $f(\mathbf{g}, \mathbf{h})$ with respect to $\mathbf{g}$ and $\mathbf{h}$, and then set both partial derivatives equal to zero:

$$\frac{\partial f}{\partial \mathbf{g}} = \mathbf{\Sigma}_{XY}\mathbf{h} - \lambda \mathbf{\Sigma}_{XX}\mathbf{g} = \mathbf{0},$$

$$\frac{\partial f}{\partial \mathbf{h}} = \mathbf{\Sigma}_{YX}\mathbf{g} - \mu \mathbf{\Sigma}_{YY}\mathbf{h} = \mathbf{0}.$$

Multiplying the above equations on the left by $\mathbf{g}^T$ and $\mathbf{h}^T$, respectively, we obtain

$$\mathbf{g}^T\boldsymbol{\Sigma}_{XY}\mathbf{h} - \lambda\mathbf{g}^T\boldsymbol{\Sigma}_{XX}\mathbf{g} = \mathbf{0},$$

$$\mathbf{h}^T\boldsymbol{\Sigma}_{YX}\mathbf{g} - \mu\mathbf{h}^T\boldsymbol{\Sigma}_{YY}\mathbf{h} = \mathbf{0}.$$

Then we have

$$\mathbf{g}^T\boldsymbol{\Sigma}_{XY}\mathbf{h} = \lambda = \mu.$$

Rearranging terms and by $\lambda = \mu$, we get that

$$-\lambda \mathbf{\Sigma}_{XX}\mathbf{g} + \mathbf{\Sigma}_{XY}\mathbf{h} = \mathbf{0},$$

$$\mathbf{\Sigma}_{YX}\mathbf{g} - \lambda \mathbf{\Sigma}_{YY}\mathbf{h} = \mathbf{0}.$$

Premultiplying the first equation by $\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}$, then substituting the second equation into the results, and rearranging terms gives

$$(\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY} - \lambda^2\mathbf{\Sigma}_{YY})\mathbf{h} = \mathbf{0}.$$

It is equivalent to

$$(\mathbf{\Sigma}_{YY}^{-1/2}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1/2} - \lambda^2\mathbf{I}_s)\mathbf{h} = \mathbf{0}.$$

For there to be a nontrivial solution to this equation, the following determinant has to be zero:

$$\left|\boldsymbol{\Sigma}_{YY}^{-1/2}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2} - \lambda^2\mathbf{I}_s\right| = 0.$$

It can be shown that the determinant is a polynomial in λ^2 of degree s, having s real roots, $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \lambda_s^2 \geq 0$, which are the eigenvalues of

$$\mathbf{R} = \boldsymbol{\Sigma}_{YY}^{-1/2} \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-1/2}$$

with associated eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$.

The maximal correlation between ξ and ω would, therefore, be achieved if we took $\lambda = \lambda_1$, the largest eigenvalue of $\mathbf{R}$.

The resultant choice of coefficients $\mathbf{g}$ and $\mathbf{h}$ of ξ and ω, respectively, are given by the vectors

$$\mathbf{g}_1 = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{v}_1, \quad \mathbf{h}_1 = \boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{v}_1.$$

In other words, the first pair of canonical variates is given by (ξ_1, ω_1), where $\xi_1 = \mathbf{g}_1^T\mathbf{X}$ and $\omega_1 = \mathbf{h}_1^T\mathbf{Y}$, and their correlation is $\text{corr}(\xi_1, \omega_1) = \mathbf{g}_1^T\boldsymbol{\Sigma}_{XY}\mathbf{h}_1 = \lambda_1$.

Given (ξ_1, ω_1), let $\xi = \mathbf{g}^T\mathbf{X}$ and $\omega = \mathbf{h}^T\mathbf{Y}$ denote a second pair of arbitrary linear projections with unit variances.

We require (ξ, ω) to have maximal correlation among all such linear combinations of $\mathbf{X}$ and $\mathbf{Y}$, which are also uncorrelated with (ξ_1, ω_1).

This last condition translates into

$$\mathbf{g}^T\Sigma_{XX}\mathbf{g}_1 = \mathbf{h}^T\Sigma_{YY}\mathbf{h}_1 = 0.$$

Furthermore, we require

$$\text{corr}(\xi, \omega_1) = \mathbf{g}^T \boldsymbol{\Sigma}_{XY} \mathbf{h}_1 = \lambda_1 \mathbf{g}^T \boldsymbol{\Sigma}_{XX} \mathbf{g}_1 = 0,$$

$$\text{corr}(\omega, \xi_1) = \mathbf{h}^T \boldsymbol{\Sigma}_{YX} \mathbf{g}_1 = \lambda_1 \mathbf{h}^T \boldsymbol{\Sigma}_{YY} \mathbf{h}_1 = 0.$$

We set

$$f(\mathbf{g},\mathbf{h}) = \mathbf{g}^T\boldsymbol{\Sigma}_{XY}\mathbf{h} - \frac{1}{2}\lambda(\mathbf{g}^T\boldsymbol{\Sigma}_{XX}\mathbf{g}-1) - \frac{1}{2}\mu(\mathbf{h}^T\boldsymbol{\Sigma}_{YY}\mathbf{h}-1)$$

$$+ \eta\mathbf{g}^T\boldsymbol{\Sigma}_{XX}\mathbf{g}_1 + \nu\mathbf{h}^T\boldsymbol{\Sigma}_{YY}\mathbf{h}_1.$$

Differentiate $f(\mathbf{g}, \mathbf{h})$ with respect to $\mathbf{g}$ and $\mathbf{h}$, and then set both partial derivatives equal to zero:

$$\frac{\partial f}{\partial \mathbf{g}} = \mathbf{\Sigma}_{XY}\mathbf{h} - \lambda \mathbf{\Sigma}_{XX}\mathbf{g} + \eta \mathbf{\Sigma}_{XX}\mathbf{g}_1 = \mathbf{0},$$

$$\frac{\partial f}{\partial \mathbf{h}} = \mathbf{\Sigma}_{YX}\mathbf{g} - \mu \mathbf{\Sigma}_{YY}\mathbf{h} + \nu \mathbf{\Sigma}_{YY}\mathbf{h}_1 = \mathbf{0}.$$

After simple algebra, we, therefore, take the second pair of canonical variates to be (ξ_2, ω_2), where

$$\mathbf{g}_2 = \boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{v}_2, \quad \mathbf{h}_2 = \boldsymbol{\Sigma}_{YY}^{-1/2}\mathbf{v}_2,$$

and their correlation is $\text{corr}(\xi_2, \omega_2) = \mathbf{g}_2^T\boldsymbol{\Sigma}_{XY}\mathbf{h}_2 = \lambda_2$.

We continue this sequential procedure, deriving eigenvalues and eigenvectors, until no further solutions can be found.

This gives us sets of coefficients for the pairs of canonical variates,

$$(\xi_1, \omega_1), (\xi_2, \omega_2), \ldots, (\xi_k, \omega_k), \quad k = \min(r, s),$$

where the ith pair of canonical variates (ξ_i, ω_i) is obtained by choosing the coefficients $\mathbf{g}_i$ and $\mathbf{h}_i$ such that (ξ_i, ω_i) has the largest correlation among all pairs of linear combinations of $\mathbf{X}$ and $\mathbf{Y}$ that are also uncorrelated with all previously derived pairs, (ξ_j, ω_j), $j = 1, 2, \ldots, i-1$.

Sample Estimates

Thus, $\mathbf{G}$ and $\mathbf{H}$ are estimated by

$$\hat{\mathbf{G}}^{(t)} = \begin{pmatrix} \hat{\mathbf{v}}_1^T \\ \vdots \\ \hat{\mathbf{v}}_t^T \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{YY}^{-1/2} \hat{\boldsymbol{\Sigma}}_{YX} \hat{\boldsymbol{\Sigma}}_{XX}^{-1} = \begin{pmatrix} \hat{\lambda}_1 \hat{\mathbf{u}}_1^T \\ \vdots \\ \hat{\lambda}_t \hat{\mathbf{u}}_t^T \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{XX}^{-1/2}, \quad \hat{\mathbf{H}}^{(t)} = \begin{pmatrix} \hat{\mathbf{v}}_1^T \\ \vdots \\ \hat{\mathbf{v}}_t^T \end{pmatrix} \hat{\boldsymbol{\Sigma}}_{YY}^{-1/2},$$

where $\hat{\mathbf{u}}_j$ is the eigenvector associated with the jth largest eigenvalue $\hat{\lambda}_j^2$ of the $(r \times r)$ symmetric matrix

$$\hat{\mathbf{R}}^* = \hat{\boldsymbol{\Sigma}}_{XX}^{-1/2} \hat{\boldsymbol{\Sigma}}_{XY} \hat{\boldsymbol{\Sigma}}_{YY}^{-1} \hat{\boldsymbol{\Sigma}}_{YX} \hat{\boldsymbol{\Sigma}}_{XX}^{-1/2}, \quad j = 1, \ldots, t.$$

$\hat{\mathbf{v}}_j$ is the eigenvector associated with the jth largest eigenvalue $\hat{\lambda}_j^2$ of the $(r \times r)$ symmetric matrix

$$\hat{\mathbf{R}} = \hat{\mathbf{\Sigma}}_{YY}^{-1/2} \hat{\mathbf{\Sigma}}_{YX} \hat{\mathbf{\Sigma}}_{XX}^{-1} \hat{\mathbf{\Sigma}}_{XY} \hat{\mathbf{\Sigma}}_{YY}^{-1/2}, \quad j = 1, \ldots, t.$$

The jth row of $\hat{\boldsymbol{\xi}} = \hat{\mathbf{G}}^{(t)}\mathbf{X}$ and the jth row of $\boldsymbol{\omega} = \mathbf{H}^{(t)}\mathbf{Y}$ together form the jth pair of sample canonical variates $(\hat{\xi}_j, \hat{\omega}_j)$ given by

$$\hat{\xi}_j = \mathbf{g}_j^T\mathbf{X}, \quad \hat{\omega}_j = \hat{\mathbf{h}}_j^T\mathbf{Y}.$$

The values (or canonical variate scores) of $\hat{\xi}_j$ and $\hat{\omega}_j$ are

$$\hat{\xi}_{ij} = \hat{\mathbf{g}}_j^T \mathbf{X}_i, \quad \hat{\omega}_{ij} = \hat{\mathbf{h}}_j^T \mathbf{Y}_i, \quad i = 1, 2, \ldots, n,$$

where

$$\hat{\mathbf{g}}_j^T = \hat{\mathbf{v}}_j^T \hat{\mathbf{\Sigma}}_{YY}^{-1/2} \hat{\mathbf{\Sigma}}_{YX} \hat{\mathbf{\Sigma}}_{XX}^{-1} = \hat{\lambda}_j \hat{\mathbf{u}}_j^T \hat{\mathbf{\Sigma}}_{XX}^{-1/2}, \quad \hat{\mathbf{h}}_j^T = \mathbf{v}_j^T \mathbf{\Sigma}_{YY}^{-1/2}$$

are the jth rows of $\hat{\mathbf{G}} = \hat{\mathbf{G}}^{(t)}$ and $\hat{\mathbf{H}} = \hat{\mathbf{H}}^{(t)}$, respectively.

The sample canonical correlation coefficient for the jth pair of sample canonical variates, $(\hat{\xi}_j, \hat{\omega}_j)$, is given by

$$\hat{\rho}_j = \hat{\lambda}_j = \frac{\hat{\mathbf{g}}_j^T \hat{\mathbf{\Sigma}}_{XY} \hat{\mathbf{h}}_j}{(\hat{\mathbf{g}}_j^T \hat{\mathbf{\Sigma}}_{XX} \hat{\mathbf{g}}_j)^{1/2} (\hat{\mathbf{h}}_j^T \hat{\mathbf{\Sigma}}_{YY} \hat{\mathbf{h}}_j)^{1/2}}, \quad j = 1, 2, \ldots, t.$$

It is usually hoped that the first t pairs of sample canonical variates will be the most important, exhibiting a major proportion of the correlation present in the data, whereas the remainder can be neglected without losing too much information concerning the correlational structure of the data.

Thus, only those pairs of canonical variates with high canonical correlations should be retained for further analysis.

Invariance

Unlike principal component analysis, canonical correlations are invariant under simultaneous nonsingular linear transformation of the random vectors $\mathbf{X}$ and $\mathbf{Y}$.

Suppose we consider linear transformations of $\mathbf{X}$ and $\mathbf{Y}$:

$$\mathbf{X} \to \mathbf{DX}, \quad \mathbf{Y} \to \mathbf{FY},$$

where the $(r \times r)$-matrix $\mathbf{D}$ and the $(s \times s)$-matrix $\mathbf{F}$ are nonsingular.

Then, the canonical correlations of $\mathbf{DX}$ and $\mathbf{FY}$ are identical to those of $\mathbf{X}$ and $\mathbf{Y}$.

How Many Pairs of Canonical Variates to Retain?

Because the question of how many pairs of canonical variates to retain is equivalent to determining the rank t of the regression coefficient matrix $\mathbf{C}^{(t)}$ in a reduced-rank regression for CVA, we approach this problem as a rank determination problem.

In the case of the rank trace, no reductions of the expressions for the coordinates of the plotted points can be obtained for the CV case as we were able to do for the PC case.

The CV rank trace can have points plotted on the exterior to the unit square, and the sequence of points may not be monotonically decreasing; we can, however, introduce a regularization parameter into the rank-trace computations to keep the plotted points within the unit square.