Principal Component Analysis and Factor Analysis

Scenario

We have a $n \times p$ numerical data matrix. Assume $n \geqslant p$ (or transpose it). We want to do one of

- show the relationships between the row (*cases*) in a low-dimensional plot of derived variables.
- show the relationships between the columns (*variables*) in a low-dimensional plot of linear combinations of cases,
- show both rows and columns on a two-dimensional plot.

The most commonly used tool for the first two is *principal component analysis* (in so-called 'R' and 'Q' modes respectively), whereas the third is tackled by *biplots*.

Note *principal* **not** *principle*.

Principal Component Analysis

PCA has several properties, most of which could be used to define it.

1. Consider all projections of the p-dimensional space onto 1 dimension. The first *principal component* (PC1) is the projection with the largest variance. A projection forms a linear combination of the variables with coefficient vector of (geometric) length one. Since variance does not depend on sign, the sign of PC1 is arbitrary.
2. Subsequent PCs are defined as the projected variable which is uncorrelated with the earlier PCs and has maximal variance, again with arbitrary sign.
3. The first $k < p$ principal components from a set of k dimensions in which the variables are uncorrelated (and if zero-mean, orthogonal), with the largest variance matrix (in any one of several senses).

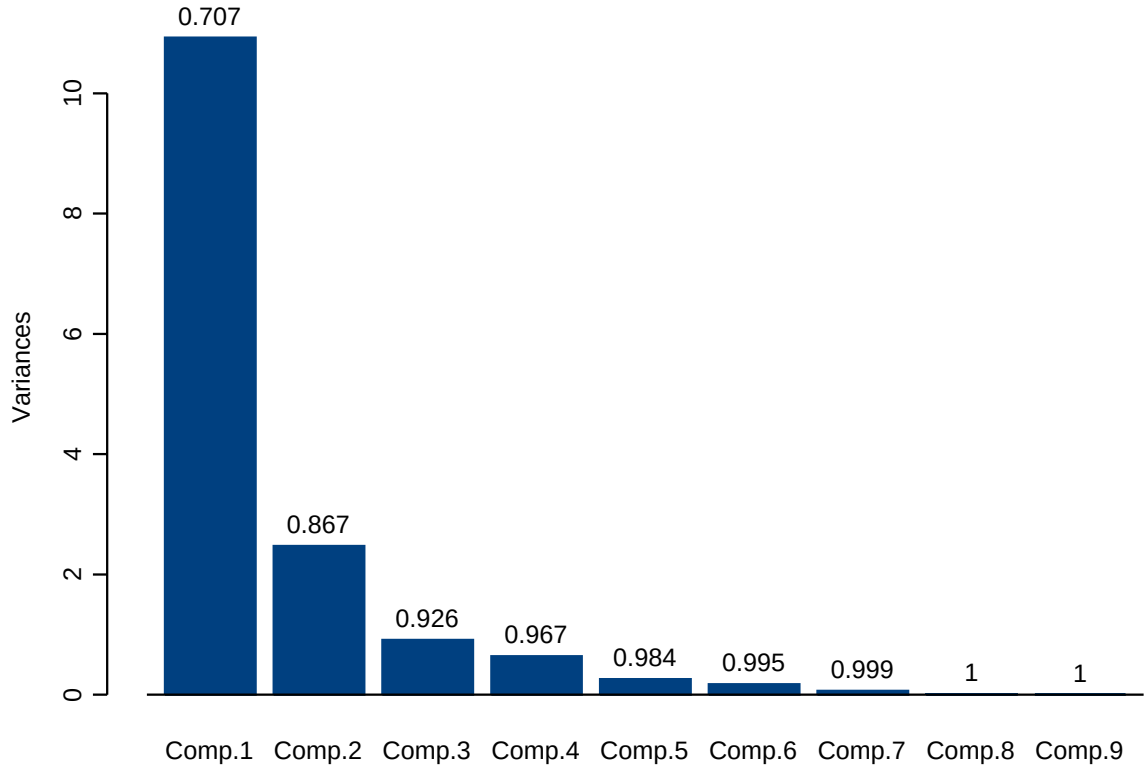
4. Consider the projection onto the space spanned by the first $k < p$ principal components. This is the k–diml space which minimizes the sum of squared distances from the cases to the space.
5. The same projection maximizes the sum of squared distances between all pairs of projected points.

The first three properties are of the (co)variances of unit-length linear combinations of variables and so can be defined from the covariance matrix of the data. This allows the use of weighted or robust covariance matrices, and also to think about PCs for multivariate distributions.

Although interest is usually in the most variable PCs, occasionally the least variable PCs are used to explore which linear combinations of the data are (nearly) constant.

Screeplots

The p PCs have decreasing variances. A screeplot is a plot of those variances. In the S-PLUS form (but not in R) the cumulative proportions are shown.



This is the forensic glass data which appears to be about 6-dimensional.

Scaling Issues

Since PCA measures variances, it is determined by the scaling of the variables, and really only makes sense if the variables are on comparable scales.

This might be the case if they are measurements in the same units or logs have been taken, but it should be normal practice to scale the variables to unit variance, or equivalently to replace the covariance matrix by the correlation matrix. (That was deliberately not done for the screeplot example.)

Sphering

Transforming the data to the PCs makes the new variables uncorrelated but of different variances. If we now rescale the PCs to unit variance, the data have unit variance in each direction, and are said to be *sphered*.

Care is needed with small variances. Clearly PCs with zero variances cannot be sphered, and those whose variability is entirely due to numerical rounding errors (in calculating them, or in data collection) should not be.

SVD and Eigendecompositions

How do we compute PCs? The best way is to use the *singular value decomposition*. Let X be the data matrix, normally centred so each column has mean zero. Then

$$X = UDV^T$$

where D is diagonal with non-negative and decreasing values and U and V are orthonormal. Then (see handout), the principal components are given by the columns of V, in order.

Now consider the covariance matrix of X,

$$S = n^{-1}X^TX = n^{-1}VDU^TUDV^T = n^{-1}VD^2V^T$$

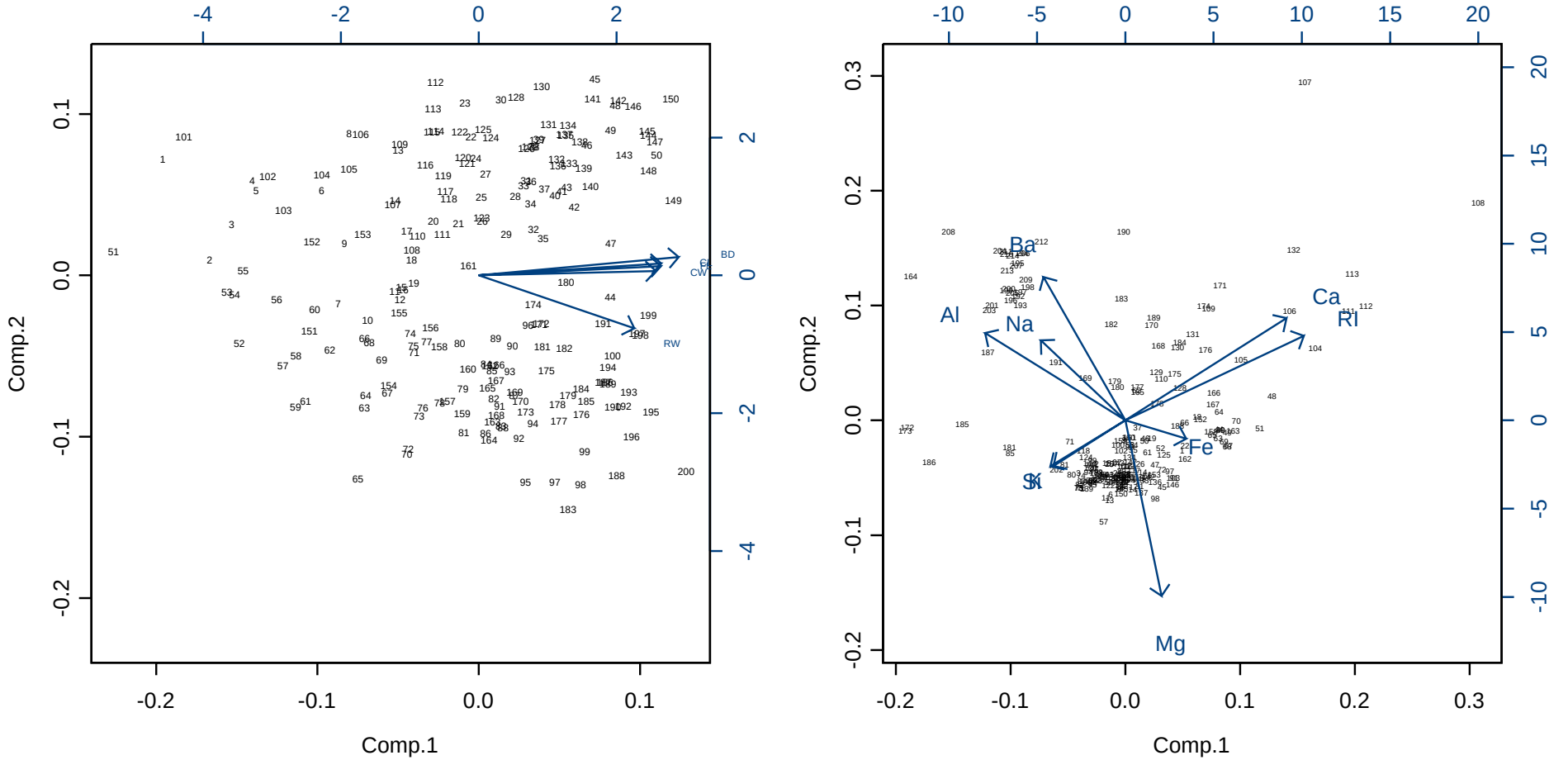
so we have an eigendecomposition of the variance matrix X, which is an alternative derivation (and not as good computationally). But this can be applied to any covariance matrix, not just an empirical one.

Biplots

Biplots at their most general are two-dimensional plots showing a set of data points and a set of axes.

Biplots were originally named by K. R. Gabriel, who used them with principal components. The simplest biplot is to show the first two PCs together with the projections of the axes of the original variables. We should take care to have equal scaling on the axes (package MASS function `eqscplot`).

So for the crabs and forensic glass data we might have:



PC biplots for (left) logged crabs data and (right) forensic glass data with correlation scaling.

There are lots of variations, and this is not the default in S-PLUS, for example.

Suppose X is centred (variables have mean subtracted). Projecting onto the first two PCs puts the data into a two-dimensional subspace—let the projected data on the original variables be $\widetilde{X}$. Since this has rank 2, it can be written

$$\widetilde{X} = GH^T \text{ where } G \text{ is } n \times 2 \text{ , } H \text{ is } p \times 2$$

A biplot plots the rows of G as the points and the rows of H as the axes (shown in R as arrows).

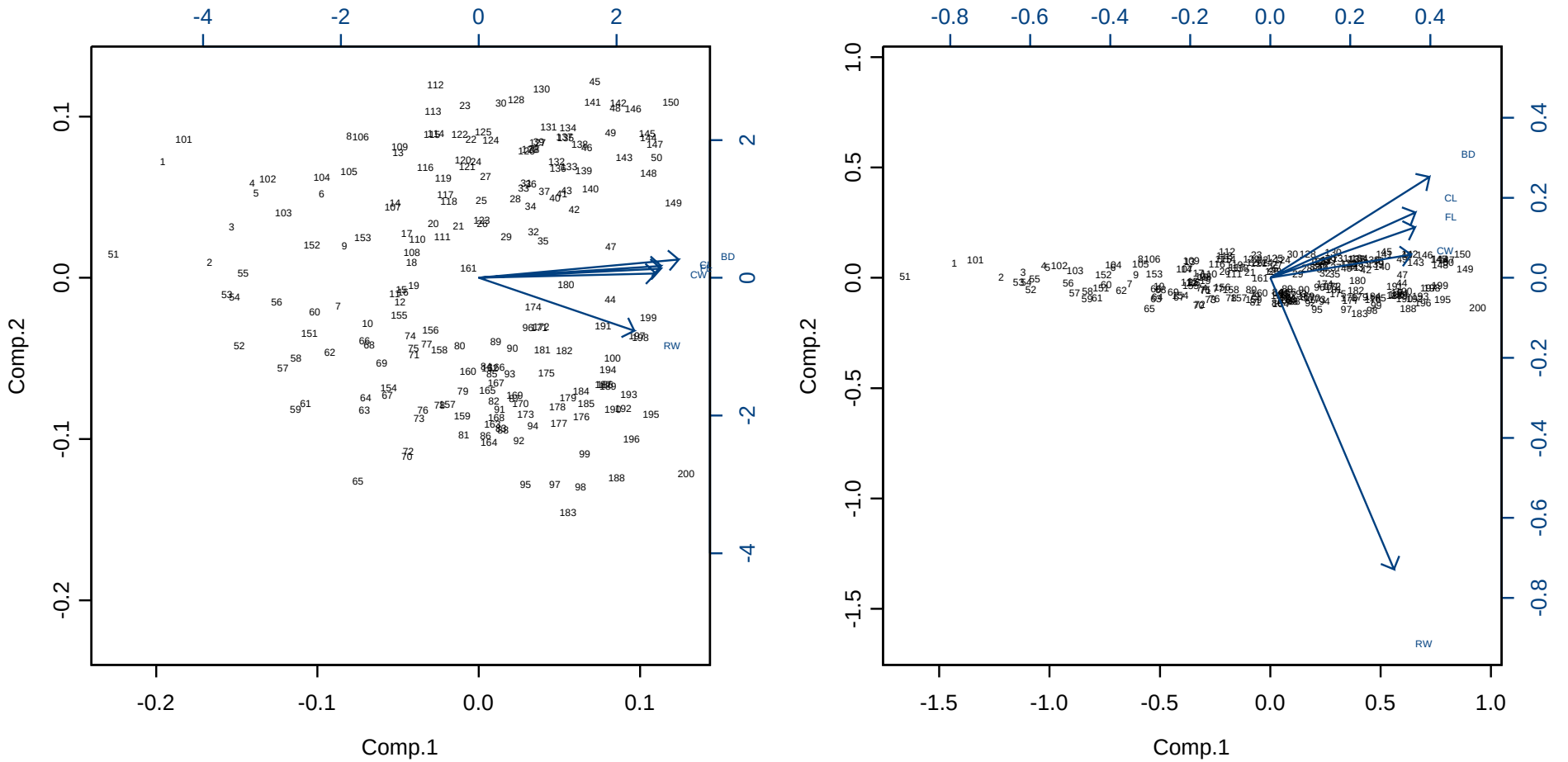
If $X = UDV^T$ is the singular-value decomposition (SVD) of X, then $X = UD_2V^T$ where D_2 is the diagonal matrix D with all but the first two diagonal elements set to zero. So we can take

$$G = UD_2^{1-\lambda}, H = VD_2^{\lambda}$$

The issue is the relative scaling of the x and y axes within which the biplot is displayed.

Gabriel chose $\lambda = 1$ for his PC biplot. Then the data points are shown on the first two *sphered* PCs and the Euclidean distances between points represents Mahalanobis distance and inner products between arrows represent covariances.

The other less common choice is $\lambda = 0$, when the points are represented by the first two PCs and the arrows are the projected axes.



Biplots for the logged crabs data with $\lambda = 1$ on the left and $\lambda = 0$ on the right.

Relationship to Factor Analysis

Principal component analysis looks for linear combinations of the data matrix X that are uncorrelated and of high variance. We can write the data columns as linear combinations of the PCs.

Independent component analysis seeks to explain the data as linear combinations of independent factors.

Factor analysis seeks linear combinations of variables, called *factors*, that represent underlying fundamental quantities of which the observed variables are expressions. More precisely, the *manifest* variables are linear combinations of the factors, plus *unique* (or *specific*) factors.

Factor analysis and PCA are often confused, and indeed SPSS has PCA as a method of factor analysis.

The factor analysis model for $k < p$ common factors $\boldsymbol{f}$ is

$$\boldsymbol{x} = \boldsymbol{\mu} + \Lambda \boldsymbol{f} + \boldsymbol{u}$$

where the components of $\boldsymbol{f}$ have unit variance and are uncorrelated, $\boldsymbol{u}$ has mean zero and unknown *diagonal* covariance matrix Ψ, and $\boldsymbol{f}$ and $\boldsymbol{u}$ are taken to be uncorrelated. (It is common to assume that no $\Psi_{ii} = 0$ or we could take one of the observed variables as a factor.)

Note that *all* the correlations amongst the variables in $\boldsymbol{x}$ must be explained by the common factors; if we assume joint normality the observed variables $\boldsymbol{x}$ will be conditionally independent given $\boldsymbol{f}$.

Since factor analysis allows an arbitrary diagonal covariance matrix Ψ, its measure of fit of the $\boldsymbol{u}_i$ depends on the problem and should be independent of the units of measurement of the observed variables. (Changing the units of measurement of the observations does not change the common factors if the loadings and unique factors are re-expressed in the new units.)

So factor analysis is really a model for the covariance matrix Σ of the data as

$$\Sigma = \Lambda\Lambda^T + \Psi$$

and such a decomposition with a $p \times k$ matrix Λ gives rise to a k factor model. Since it is scale independent, we can further view it as model of the *correlations* of the data matrix.

We have some choice about how to fit such a model. The most satisfactory is to assume joint normality and use maximum likelihood, but this is a hard optimization problem with often many local maxima.

Principal component analysis also seeks a linear subspace like $\Lambda \boldsymbol{f}$ to explain the data, but measures the lack of fit by the sum of squares of the $\boldsymbol{u}_i$. So again we have

$$\boldsymbol{x} = \boldsymbol{\mu} + \Lambda \boldsymbol{f} + \boldsymbol{u}$$

and the covariance matrix is

$$\Sigma = \Lambda\Lambda^T + \Xi$$

where now the error matrix Ξ will not be diagonal but will be small (in sums of squares of elements).

Fitting is just a question of minimizing sums of squares and reduces to matrix decompositions (singular value decomposition of the data matrix or eigendecomposition of the covariance matrix).

‘Principal Factor Analysis’

An early method to do factor analysis was *via* PCA, the so-called ‘principal factor analysis’. This should not be used!

Suppose we know Ψ. We can then write

$$\Sigma - \Psi = \Lambda\Lambda^T$$

where Λ is $p \times k$ and so can be found as the first k PCs of the LHS. If we don’t know Ψ we start with a guess $\widehat{\Psi}$ at the specific variances and iterate

$$\Sigma - \widehat{\Psi} = \widehat{\Lambda}\widehat{\Lambda}^T, \qquad \widehat{\Psi} = \text{diag}(\Sigma - \widehat{\Lambda}\widehat{\Lambda}^T)$$

This need not converge, but it usually does (albeit slowly). The starting point will influence the values found, and they will fit S exactly on the diagonal but not (usually) off the diagonal.

Note that this method of estimation is not scale-independent, so best done on the correlations. It minimizes the sum of squares of the elements of $\Sigma - \Lambda\Lambda^T - \widehat{\Psi}$

We can get an initial guess from the residual variances of the regressions of each variable on all the others.

To sum up

- PCA is a model for the covariance structure which is expanded in an ordered set of p components of decreasing variance.
- Factor analysis is a model for the correlation structure in terms of $k < p$ underlying unordered factors plus measurement errors.
- PCA is often used to approximate data in lower dimensions, in which case the criterion is that the sum of squares of the approximation errors is small.
- Factor analysis also is used to approximate data in lower dimensions. The criterion is that the measurement errors are independent, and projection to the factor scores is controversial.
- There is a way to pick (in order) the principal components, but factor analysis only defines a k-diml subspace.