8 Principal Components Analysis (PCA)

8.1 Eigenvalues and Diagonal Forms

Recall that for a $p \times p$ matrix B, $\lambda_1, \ldots, \lambda_p$ and $e_1, \ldots, e_p$, $|e_j| \neq 0$ are called eigenvalues and eigenvectors of B, respectively, if $Be_j = \lambda_j e_j$. Note that λ_i is an eigenvalue of B iff $\lambda_1, \ldots, \lambda_p$ are the roots of the pth order polynomial $|B - \lambda I|$ in λ.

For simplicity of calculations, we often normalise the eigenvectors such that e_j so that $|e_j| = 1$. The set of eigenvalues is unique, the eigenvectors are not unique.

Some properties that we have already discussed and that are useful in the context of PCA:

1. if B is symmetric, eigenvalues and eigenvectors are real (otherwise they can be complex)
2. if B is symmetric, then the eigenvectors corresponding to different eigenvalues are orthogonal, and the remaining eigenvectors (those corresponding to identical eigenvalues) can be chosen to be orthogonal. That is, we can select an orthogonal basis in R^p that consists of eigenvectors of B.
3. $|B| = \prod_{i=1}^{p} \lambda_i$
4. $tr(B) = \sum_{i=1}^{p} \lambda_i$
5. the eigenvalues of a diagonal matrix are the diagonal elements themselves;
6. the number of nonzero eigenvalues is equal to the rank of B (recall definition of rank);
7. a symmetric matrix B is positive definite iff all its eigenvalues are positive; it is positive semidefinite if all its eigenvalues are nonnegative;
8. the (non-zero) eigenvalues of B and B^T are identical, but the eigenvectors are not in general identical;
9. if B^{-1} exists then it has eigenvalues $\lambda_1^{-1}, \ldots, \lambda_p^{-1}$;

Given an $n \times p$ matrix X let $Y = XA$ for some orthogonal $p \times p$ matrix A, i.e. $A^T A = I$. Find A such that $Y^T Y$ is diagonal, i.e. such that

$$A^T X^T X A = L = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Note that since A is orthogonal, $A^T X^T X A = L$ is equivalent to $X^T X A = AL$, i.e. to

$$X^T X(\mathbf{a}_1, \dots, \mathbf{a}_p) = (\mathbf{a}_1, \dots, \mathbf{a}_p) \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix},$$

where a_j is the jth column of A. Hence we need $X^T X a_j = \lambda_j \mathbf{a}_j$, $j = 1, \dots, p$. By definition, this means that λ_j is an eigenvalue of $X^T X$ and a_j is its associated eigenvector. Since $X^T X$ is symmetric, we can choose the eigenvectors in such a way that $\mathbf{a}_i^T \mathbf{a}_j = 0$ if $i \neq j$ and $\mathbf{a}_i^T \mathbf{a}_i = 1$ so that A is indeed orthogonal. Therefore, we have for the columns of Y, given as $y_i = X\mathbf{a}_i$,

$$y_i^T y_i = \lambda_i, \qquad y_i^T y_j = 0, \quad i \neq j.$$

Interpretation. If X is a centered data matrix, then $Y = XA$ is a linear transformation of the data. This linear transformation re-defines the original variables (columns of X) to new ones (columns of Y) such that these new variables are **uncorrelated**.

Spectral Decomposition Since $A^T A = I$ and $A^T X^T X A = L$, we also have $X^T X = ALA^T$. This can be written as

$$X^T X = ALA^T = (\mathbf{a}_1, \dots, \mathbf{a}_p) \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_p \end{bmatrix} \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{bmatrix} = \sum_{i=1}^{p} \lambda_i \mathbf{a}_i \mathbf{a}_i^T,$$

which is known as the **spectral decomposition** of a matrix $X^T X$. Note that $= \sum_{i=1}^{p} \lambda_i \mathbf{a}_i \mathbf{a}_i^T$ is the sum of p matrices, each of rank 1 (why?), and each multiplied by a scalar λ_i. If some of the

eigenvalues λ_i are 'close to zero', the matrix can be approximated by the sum of fewer rank-one matrices.

Any symmetric matrix B can always be written in the form $B = V\Lambda V^T$ where Λ is diagonal with the eigenvalues of B and orthogonal matrix V consisting of orthonormal eigenvectors of B. Vice versa, whenever V^TBV is a diagonal matrix for some orthogonal matrix V, then the diagonal elements of V^TBV are the eigenvalues of B. (Exercise: Prove it.)

Inverse If the matrix X^TX of sums of squares and products (SSP) is of full rank (all eigenvalues are positive) then the inverse must exist. In particular, L^{-1} exists and is given as $diag(\lambda_1^{-1}, \ldots, \lambda_p^{-1})$. We can derive the inverse $(X^TX)^{-1}$ as follows:

$$(X^TX)^{-1} = (ALA^T)^{-1} = \sum_{i=1}^{p} \lambda_i^{-1} \mathbf{a}_i \mathbf{a}_i^T.$$

Note, if some eigenvalues λ of X^TX are relatively close to zero (relative to the measurement scale) then $(X^TX)^{-1}$ is almost **singular** and even might be **ill conditioned** (numerically non-invertible).

Repeated Eigenvalues It is possible that some roots of the polynomial $|B - \lambda I|$ have multiplicity greater than 1 (i.e. algebraic multiplicity greater than 1), this means that several eigenvalues have the same value. However, if B is symmetric (as in the case of covariance matrices), dimension of an eigenspace (i.e. geometric multiplicity) corresponding to an eigenvalue λ_i with an algebraic multiplicity κ_i equals to κ_i. That is, we can find κ_i linearly independent eigenvectors corresponding to λ_i. By the Gramm-Schmidt procedure, we can select these eigenvectors to be orthonormal.

9 The Method of Principal Components Analysis

We consider here the situation of multivariate data sets where typically all variables are quantitative (and preferably continuous). As before, X is the centered data matrix, i.e. $(I - 11^T/n)X_0$. It is often plausible that the data do not vary equally in all p dimensions, but that most variation happens in some lower dimensional subspace. Principal components (PC) are such axes, so that the first PC reflects most variation, the second PC reflects the second most variation and is orthogonal to the first etc. (Recall our discussion on SVD!)

A new variable, or new axis, $\mathbf{y}$ is a linear combination of the original p variables x_j , $j = 1, \ldots, p$

$$\mathbf{y} = x_1 a_1 + x_2 a_2 + \ldots + x_p a_p, \quad or \;\; \mathbf{y} = X\mathbf{a}.$$

The **first principal component** is defined as the linear combination, i.e. choice of $p \times 1$-vector $\mathbf{a}$ that maximises the (sample) variance s_y^2 of y. Here $s_y^2 = y^t y/(n-1)$ where

$$y^T y/(n-1) = \frac{(X\mathbf{a})^T X\mathbf{a}}{n-1} = \mathbf{a}^T S\mathbf{a}.$$

This can obviously be arbitrarily large by scaling a, hence we constrain the maximisation to $||a|| = 1$. To maximise a function subject to a constraint we can use a Lagrange multiplier λ to maximise with respect to $\mathbf{a}$:

$$f(a) = \mathbf{a}^T S\mathbf{a} - \lambda(a^T a - 1).$$

Differentiation yields that we need $\mathbf{a}$ such that

$$(S - \lambda I)a = 0$$

which requires for a nontrivial solution that $|(S - \lambda I)| = 0$. This is exactly the definition of eigenvalues and eigenvectors, i.e. λ has to be an eigenvalue and $\mathbf{a}$ be the associated eigenvector of the sample variance matrix S.

Let us work out which ones to choose so as to have maximal variance. I.e., we want $\mathbf{a}^T S\mathbf{a}$ maximal and we know that $S\mathbf{a} = \lambda\mathbf{a}$, hence $\mathbf{y}^T\mathbf{y}/(n-1) = \lambda$, which will be maximal if we choose $\mathbf{a}$ to be the eigenvector associated with the largest eigenvalue.

The above shows how to obtain the **first principal component**. The second, third etc. principal components are defined as those linear combinations with second, third etc. largest variance, such that they are orthogonal to all previous principal components. They are calculated analogously to the first one. Let X be the centered data matrix and S the sample variance-covariance matrix. Let $(\lambda_i, \ldots, \lambda_p)$ be the eigenvalues of S in descending order and $(\mathbf{a}_1, \ldots, \mathbf{a}_p)$ be the associated (normalised) eigenvectors. The principal components can be calculated as:

$$\mathbf{y}_1 = X\mathbf{a}_1, \ldots, \mathbf{y}_p = X\mathbf{a}_p.$$

Note that the eigenvectors $\mathbf{a}_i$ are often called **loadings** or **latent** vectors; and y_{ji} is often called the **score** of the ith principal component on the jth observation/individual.

Properties:

1. The sample variances of the PCs are $s^2_{y_i} = \hat{\lambda}_i = \mathbf{y}_i^T \mathbf{y}_i/(n-1)$ and hence $s^2_{y_1} \geq s^2_{y_1} \ldots \geq s^2_{y_p}$

2. Since the eigenvalues λ_i of S are equal to sample variances $s^2_{y_i}$, they must be nonnegative. Hence, S (and SSP) is positive semidefinite.

3. The principal components are orthogonal and centered, and hence are uncorrelated, i.e. $\mathbf{y}_i^T \mathbf{y}_j = 0$, $i \neq j$.

4. The empirical variance-covariance matrix of the transformed data set $Y = (\mathbf{y}_1, \ldots, \mathbf{y}_p)$ is

$$S_y = Y^T Y/(n-1) = L = diag(\lambda_1, \ldots, \lambda_p).$$

Example. Consider data on concentrations of 13 different chemicals in wines grown in the same region in Italy that are derived from three different cultivars. There are 178 samples of wine, with one row per wine sample and one column per concentration (out of the 13 different chemicals in that sample) (Multivariate Analysis, A. Coghlan).

Let us start from creating a matrix scatterplot where each pair of variables plotted against each other and diagonal cells show histograms of each of the variables.

```
> library("car")
> scatterplotMatrix(wine)
```

We now look at variance of each of the 13 components:

```
        V2           V3           V4           V5           V6           V7
6.590623e-01 1.248015e+00 7.526464e-02 1.115269e+01 2.039893e+02  3.916895e-01

        V8           V9          V10          V11          V12          V13
9.977187e-01 1.548863e-02 3.275947e-01 5.374449e+00 5.224496e-02 5.040864e-01

        V14
9.916672e+04
```

Scaling and Standardisation Principal components are scale and unit dependent. In the wine example, we find that the concentrations of the 13 chemicals in the wine samples show a wide range of variances, from 0.01548862 for V9 to 99166.72 for V14. This is a range of approximately $99166.72/0.01548862 = 6,402,554$-fold in the variances. Hence, the resulting PCs would be very different because the first principal component would be dominated by the variables which show the largest variances, such as V14.

Also, if we get different measurement units, we also observe different PCs. Indeed, suppose researcher A measured weight, height and age of people in lb, ft, and years yielding data matrix X, while researcher B measured the same variables on the same people in kg, m, and months yielding data matrix $\tilde{X}$. The relation between the two data matrices is $\tilde{X} = XK$, where $K = diag(1/2.2, 1/3.28, 12)$. The relation between the two sample covariance matrices is $\tilde{S} = KSK$ (since note $K^T = K$).

In consequence, it is important for the interpretation of a principal component analysis that the original variables are measured on comparable scales. It is not meaningful to look at linear combinations of disparate measurements like length, numbers of attributes, age, weight. If the variables are not comparable, we could standardise them to zero mean and unit variance, i.e. use the correlation matrix instead of the variance matrix for the PCA. Remember that we obtain the matrix of correlations R as

$$R = D^{-\frac{1}{2}} S D^{-\frac{1}{2}},$$

where D is the diagonal matrix containing the diagonal elements of S, i.e. the sample variances.

The eigenvalues and eigenvectors of R can be interpreted similar to principal component for S, but they reflect the relative patterns among the variables rather than absolute values. A PCA on R can also be useful even when the measurements are all in the same unit but there are large differences in the variances (possibly indicating that they are again not really comparable despite being in the same unit).

In summary, when we decide whether to apply PCA to S or R, ask yourself if the variables are measured on comparable scales. If not, then using S makes no sense, and using R might make some sense. Also think about whether an analysis of a subset of the variables, e.g. only those that are on comparable scales might help. If all original variables are measured on comparable scales, then decide whether you are interested in the relative patterns (use R) or the absolute patterns (use S).

You easily can standardise variables in R as follows:

```
> standardisedconcentrations <- as.data.frame(scale(wine[2:14]))

####### reality check ##########
> sapply(standardisedconcentrations, var)
 V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14
  1   1   1   1   1   1   1   1   1   1   1   1   1

> sapply(standardisedconcentrations[1:6], mean)
           V2            V3            V4            V5            V6            V7
-8.591766e-16 -6.776446e-17  8.045176e-16 -7.720494e-17 -4.073935e-17 -1.395560e-17

> sapply(standardisedconcentrations[7:13], mean)
           V8            V9           V10           V11           V12           V13
 6.958263e-17 -1.042186e-16 -1.221369e-16  3.649376e-17  2.093741e-16  3.003459e-16

           V14
 -1.034429e-16
```

We can now a principal components analysis on the standardised concentrations:

```
> wine.pca <- prcomp(standardisedconcentrations)
> summary(wine.pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     2.169 1.5802 1.2025 0.95863 0.92370 0.80103 0.74231

                          PC8     PC9     PC10    PC11    PC12    PC13
                       0.59034 0.53748 0.5009 0.47517 0.41082 0.32152
```

```
Proportion of Variance 0.362 0.1921 0.1112 0.07069 0.06563 0.04936 0.04239
                       0.02681 0.02222 0.0193 0.01737 0.01298 0.00795

Cumulative Proportion  0.362 0.5541 0.6653 0.73599 0.80162 0.85098 0.89337
                       0.92018 0.94240 0.9617 0.97907 0.99205 1.00000
```

we can also check the total variance explained by the components, i.e. the sum of the variances of the components:

```
> sum((wine.pca$sdev)^2)
[1] 13
```

This is not surprising. Why?

Now, compare to the unscaled PCA:

```
> wine.pcaUNSCALED <- prcomp(wine[2:14])
> summary(wine.pcaUNSCALED)
Importance of components:
                            PC1      PC2     PC3     PC4     PC5     PC6    PC7
Standard deviation     314.9632 13.13527 3.07215 2.23409 1.10853 0.91710 0.5282
                            PC8      PC9     PC10    PC11    PC12    PC13
                         0.3891  0.3348   0.2678  0.1938   0.1452  0.09057
Proportion of Variance   0.9981  0.00174 0.00009 0.00005 0.00001 0.00001 0.0000
                         0.0000 0.0000 0.0000 0.0000 0.0000 0.00000
Cumulative Proportion    0.9981  0.99983 0.99992 0.99997 0.99998 0.99999 1.0000
                         1.0000 1.0000 1.00000 1.0000 1.0000 1.00000
```

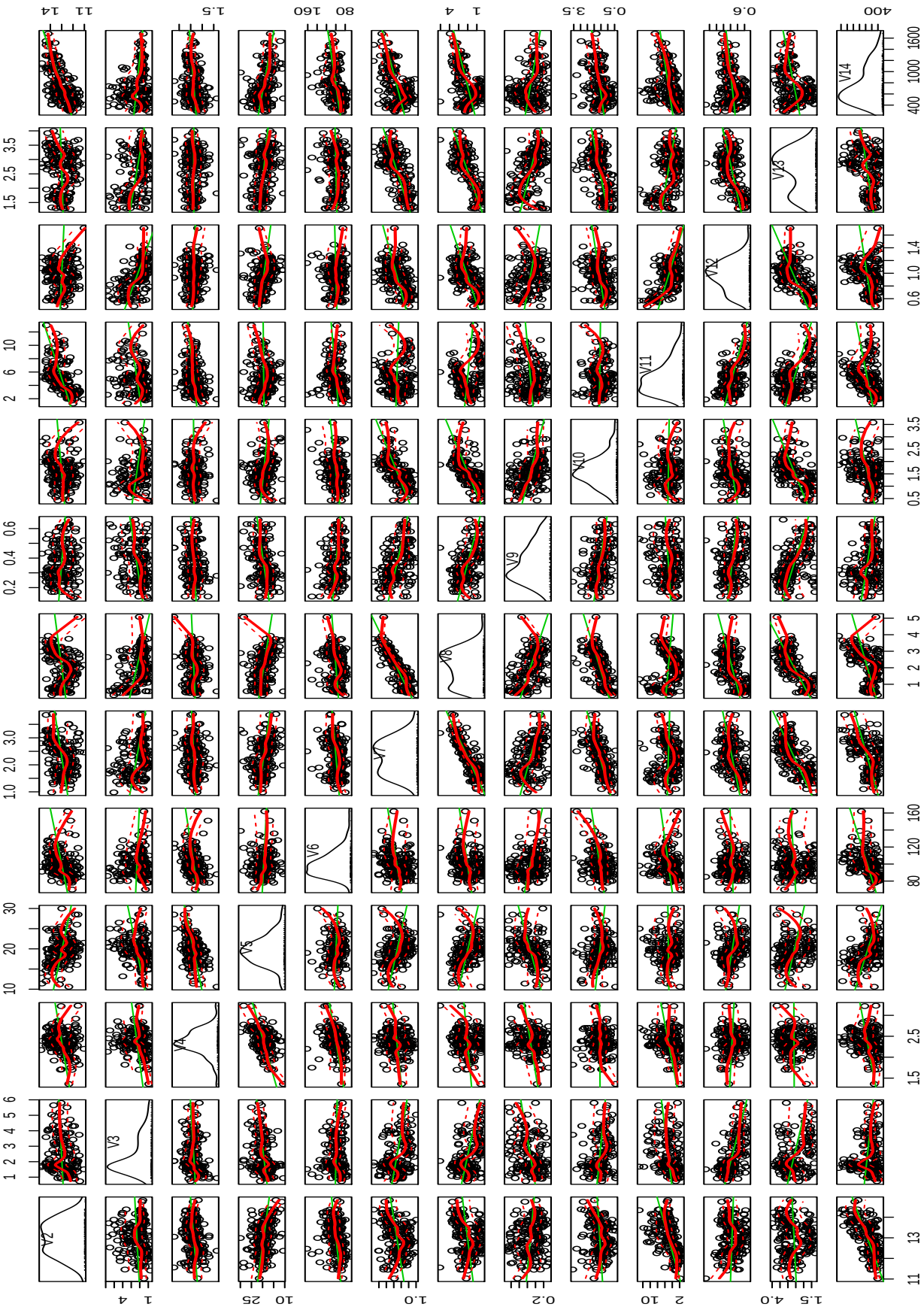

Figure 9.1: Matrix scatter plot for the wine data.