Data visualization

The first step in any data analysis is to get some overview of the data which involves descriptive techniques. Some questions of interest might for example be: Are there any variables spread out more than others? Are there variables indicating subgroups of the data? Are there outliers in the data?

Example 1. The Swiss bank data consists of 200 measurements on Swiss bank notes. The first half of the measurements are from genuine bank notes, the other half are from counterfeit bank notes. For each bank note we have data on

1) $X_1 =$ length of the bill
2) $X_2 =$ height of the bill (left)
3) $X_3 =$ height of the bill (right)
4) $X_4 =$ distance of the inner frame to the lower border
5) $X_5 =$ distance of the inner frame to the upper border
6) $X_6 =$ length of the diagonal of the central picture

Aim: study how these measurements may be used in determining whether a bill is genuine or counterfeit.

Boxplots are useful tools for comparing groups/batches, which enable to compare distributions of the data among different groups. It helps us see the location, skewness, spread, tail length and outlying points. The 5-number summary: lower extreme, lower quartile F_L, median, upper quartile F_U, upper extreme. Let $d_F = F_U - F_L$, then *outliers* (in the boxplot context) are defined as points *outside bars*:

$$F_L - 1.5d_F, \quad F_U + 1.5d_F.$$

Observe that "whiskers" from each end of the box are drawn to the most remote point that is not an outlier.

Study the parallel boxplot of the diagonal variable X_6. Do the same for the length of the bill, i.e. for X_1.

Histograms are density estimates and give an impression of the distribution of the data. They show possible multimodality of the data.

Scatter plots (bivariate or trivariate plots of variables against each other) help us understand relationships among the variables of a data set. Study the 2D scatterplot of X_5 vs X_6 and 3D plot of (X_4, X_5, X_6). To study interrelations between variables in multivariate data, we can also study **pairwise scatter diagrams** in the same plot (draftsman's or pairs plot). Scatter plots help in identifying separated points or sub-clusters, and they help in judging positive or negative dependencies.

Study also the covariance and correlation matrix for the bank data.

Ex.1. Car data.

1) Study the variable M (miles per gallon). Make the boxplots and find the 5 number summaries for American, Japanese and European cars. Which car group is most fuel efficient? Which car is most fuel efficient? Compare the worst Japanese car with the US cars. Can we say that the worst Japanese car is more fuel efficient than almost 50% of the US cars?

2) Draw a histogram for the mileage variable of the car data. Do the same for the 3 groups.

3) Study the correlation between mileage and weight. Make the scatter plot so that you can see the different groups. Comment on "Japanese cars generally have better mileage than the others."

Ex.2. Is the upper extreme always an outlier (in the boxplot context)? Is it possible for the mean or the median to lie outside of the lower and upper quartiles? Is it possible that all five numbers of the 5-number summary could be equal?

Spectral decomposition

For every square matrix $\mathbf{A} : n \times n$ a value λ and a nonzero vector $\mathbf{x}$ can be found such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad \Longleftrightarrow \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

Here λ is called an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is the eigenvector associated with λ. In order to obtain nontrivial solutions, we have to solve the characteristic equation

$$|\mathbf{A} - \lambda\mathbf{I}| = 0,$$

which has n roots, that is, $\mathbf{A}$ has n eigenvalues $\lambda_1, \ldots, \lambda_n$. After finding $\lambda_1, \ldots, \lambda_n$, the accompanying eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ can be found. Eigenvectors are unique only up to multiplication by a scalar: if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, then $k\mathbf{x}$, $k \in \mathbb{R}$, is also an eigenvector. Typically, $\mathbf{x}$ is scaled so that $\mathbf{x}'\mathbf{x} = 1$.

Consider now a **symmetric** (real) matrix $\mathbf{A} : n \times n$. Let $\mathbf{\Lambda}$ be the diagonal matrix of eigenvalues of $\mathbf{A}$ and let $\mathbf{V}$ be the matrix of normalized eigenvectors of $\mathbf{A}$ ($\mathbf{V}'\mathbf{V} = \mathbf{I}$). Then

$$\mathbf{A} = \mathbf{A}\mathbf{V}\mathbf{V}' = \mathbf{A}(\mathbf{v}_1, \dots, \mathbf{v}_n)\mathbf{V}'$$

$$= (\lambda_1\mathbf{v}_1, \lambda_2\mathbf{v}_2, \dots, \lambda_n\mathbf{v}_n)\mathbf{V}' = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'.$$

The expression $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$ for a symmetric matrix $\mathbf{A}$ in terms of its eigenvalues and eigenvectors is known as the **spectral decomposition** of $\mathbf{A}$.

Ex.3. Find the eigenvalues and eigenvectors for $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ -1 & 4 \end{pmatrix}$.

Ex.4. Let

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 2 & 0 \end{pmatrix}.$$

a) Find the eigenvalues and normalized eigenvectors of $\mathbf{A}$. Use the eigenvectors as columns in $\mathbf{V}$.
b) Show that $\mathbf{V}'\mathbf{A}\mathbf{V} = \mathbf{\Lambda}$, where $\mathbf{\Lambda}$ is diagonal with the eigenvalues of $\mathbf{A}$ on the diagonal.
c) Show that $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$.
d) Check that $tr(\mathbf{A}) = \sum_j \lambda_j$, $|\mathbf{A}| = \prod_j \lambda_j$.

Ex.5. Let

$$\mathbf{A} = \begin{pmatrix} 3 & 6 & -1 \\ 6 & 9 & 4 \\ -1 & 4 & 3 \end{pmatrix}.$$

a) Find the spectral decomposition of $\mathbf{A}$.
b) Find the spectral decomposition of $\mathbf{A}^2$ and show that $\mathbf{A}^2 = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}'$.
c) Find the spectral decomposition of $\mathbf{A}^{-1}$ and show that $\mathbf{A}^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}'$.

Singular value decomposition (SVD)

Any (real) matrix $\mathbf{A} : n \times p$ can be expressed in terms of eigenvalues and eigenvectors of $\mathbf{A}'\mathbf{A}$ and $\mathbf{A}\mathbf{A}'$. Let $rank(\mathbf{A}) = k$. The **singular value decomposition** of $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}',$$

where $\mathbf{U} : n \times k$, $\mathbf{D} : k \times k$, $\mathbf{V} : p \times k$. Here $\mathbf{D} = diag(\lambda_1, \ldots, \lambda_k)$, where $\lambda_1^2, \ldots, \lambda_k^2$ are the nonzero eigenvalues of $\mathbf{A}'\mathbf{A}$ or $\mathbf{A}\mathbf{A}'$. The values $\lambda_1, \ldots, \lambda_k$ are called the *singular values* of $\mathbf{A}$. The k columns of $\mathbf{U}$ are the normalized eigenvectors of $\mathbf{A}\mathbf{A}'$ corresponding to the eigenvalues $\lambda_1^2, \ldots, \lambda_k^2$. The k columns of $\mathbf{V}$ are the normalized eigenvectors of $\mathbf{A}'\mathbf{A}$ corresponding to the eigenvalues $\lambda_1^2, \ldots, \lambda_k^2$. Since the columns of $\mathbf{U}$ and $\mathbf{V}$ are normalized eigenvectors of symmetric matrices, they are mutually orthogonal, that is $\mathbf{U}'\mathbf{U} = \mathbf{V}'\mathbf{V} = \mathbf{I}$.

Ex.6. Find the singular value decomposition of

$$\mathbf{A} = \begin{pmatrix} 4 & -5 & -1 \\ 7 & -2 & 3 \\ -1 & 4 & -3 \\ 8 & 2 & 6 \end{pmatrix}.$$

Find also the eigenvalues and eigenvectors of $\mathbf{A}\mathbf{A}'$ and $\mathbf{A}'\mathbf{A}$ and compare with the results from SVD.

Matrix algebra in R: `http://www.statmethods.net/advstats/matrix.html`