§1: Multidimensional Data

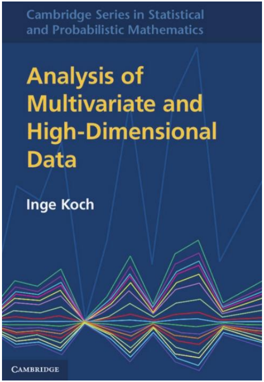
106-1 高維度資料分析

開課班級: 統碩 1, 2，統計系 4, 巨資學士學程

授課教師: 吳漢銘 (國立臺北大學統計學系)

教學網站: http://www.hmwu.idv.tw

上課用書: Inge Koch, 2013, Analysis of Multivariate and High-Dimensional Data, Cambridge University Press; 1 edition.



系級: ____________ 學號: ____________ 姓名: __________________

1.1 Multivariate and High-Dimensional Problems

1. Scientists (Pearson, 1901; Hotelling, 1933; Fisher, 1936) developed methods for analysing multivariate data in order to
   1. understand the _structure_ in the data,
   2. _summarise_ it in simpler ways,
   3. understand the _relationship_ of one part of the data to another part,
   4. make _decisions and inferences_ based on the data.
2. Linear methods: Principal Component Analysis (_PCA_), Canonical Correlation Analysis (_CCA_), Linear Discriminant Analysis (_LDA_).
3. Renewed requirements for linear methods have arisen to handle very _large_ and _high-dimensional_ data.

1. The data structure can often be obscured by <u>noise</u>:
   1. <u>reduce</u> the original data in such a way that <u>informative and interesting</u> structure in the data is preserved.
   2. <u>remove</u> noisy, irrelevant or purely random variables, dimensions or features.
2. PCA has become indispensable as a dimension reduction tool and is often used as a <u>first step</u> in a more comprehensive analysis.
3. Traditionally one assumes that the <u>dimension d</u> is small compared to the <u>sample size n</u>.
4. For the asymptotic theory, n increases while the dimension remains <u>constant</u>.

1.1 Multivariate and High-Dimensional Problems

1. Now we encounter:
    1. data whose dimension is comparable to the sample size, and both are large;
    2. high-dimension low sample size (HDLSS) data whose dimension d vastly exceeds the sample size n, so $d >> n$; and
    3. functional data whose observations are functions.
2. High-dimensional and functional data pose special challenges, and their theoretical and asymptotic treatment is an active area of research.
3. Gaussian assumptions will often not be useful for high-dimensional data.

1.1 Multivariate and High-Dimensional Problems

1. A deviation from <u>normality</u> does not affect the applicability of PCA or CCA.
2. Exercise care when making <u>inferences</u> based on Gaussian assumptions or when we want to exploit the normal asymptotic theory.
3. A number of topics that are needed in subsequent chapters.

   §1.2 displaying or visualising data,

   §1.3 introduces random vectors and data,

   §1.4 discusses Gaussian random vectors and summarises results,

   §1.5 deal with matrices, including the spectral decomposition.

1.2 Visualisation

1. Before we analyse a set of data, it is important to __look at it__.
2. We get __useful clues__ such as skewness, bi- or multi-modality, outliers, or distinct groupings; these influence or direct our analysis.
3. Graphical displays are __exploratory data analysis__ tools, which, if appropriately used, can enhance our understanding of data.
4. The insight obtained from graphical displays is more __subjective__ than quantitative.
   1. __Visual__ cues are easier to understand and interpret than numbers alone.
   2. The knowledge gained from graphical displays can __complement__ more quantitative answers.

1.2.1 Three-Dimensional Visualisation

1. 2D scatterplots are a natural way of looking at data with <u>three or more variables</u>.
2. As the number of variables increases, sequences of 2D scatterplots become less feasible to interpret, but <u>rotating the data</u> can better reveal the structure of the data.
3. The scatterplots in Figure 1.1 (Example 2.4 of Section 2.3) display the 10,000 observations and the three variables CD3, CD8 and CD4 of the five-dimensional HIV^{+} and HIV^{-} data sets, which contain measurements of blood cells relevant to HIV.

1.2.1 Three-Dimensional Visualisation

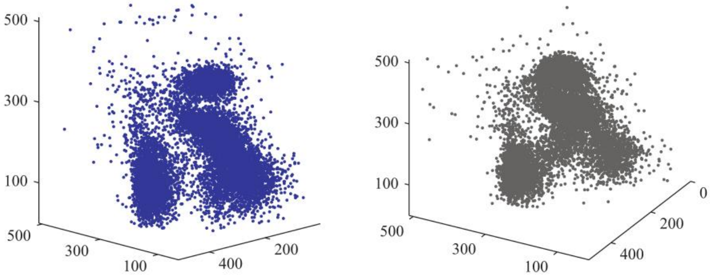


Figure 1.1 HIV$^+$ data (*left*) and HIV$^-$ data (*right*) of Example 2.4 with variables *CD3*, *CD8* and *CD4*.

1. Compare two figures by presenting the data in the form of __movies__ or combine a series of different views of the same data.
2. Other possibilities: __project__ the five-dimensional data onto a smaller number of orthogonal directions and displaying the lower-dimensional projected data (Figure 1.2.)

1.2.1 Three-Dimensional Visualisation

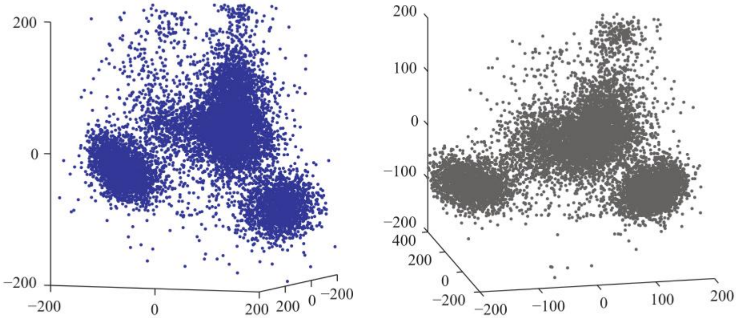


Figure 1.2 Orthogonal projections of the five-dimensional HIV$^+$ data (*left*) and the HIV$^-$ data (*right*) of Example 2.4.

1. We can see a smaller <u>fourth cluster</u> in the top right corner of the HIV$^-$ data, which seems to have almost disappeared in the HIV$^+$ data in the left panel. (see Section 2.4, how to find informative projections.)

1.2.1 Three-Dimensional Visualisation

1. Representing low-dimensional data in a number of 3D scatterplots (Figure 1.3) - which make use of <u>colour</u> and different plotting <u>symbols</u> to enhance interpretation.
2. Display the four variables of Fisher's iris data - sepal length, sepal width, petal length and petal width - in <u>a sequence of</u> 3D scatterplots. The data consist of three species: Setosa (red), Versicolor (green) and Virginica (black).
3. We can see that the red observations are well separated from the other two species for all combinations of variables, whereas the green and black species are not as easily <u>separable</u>. (more detail in Example 4.1 of Section 4.3.2.)

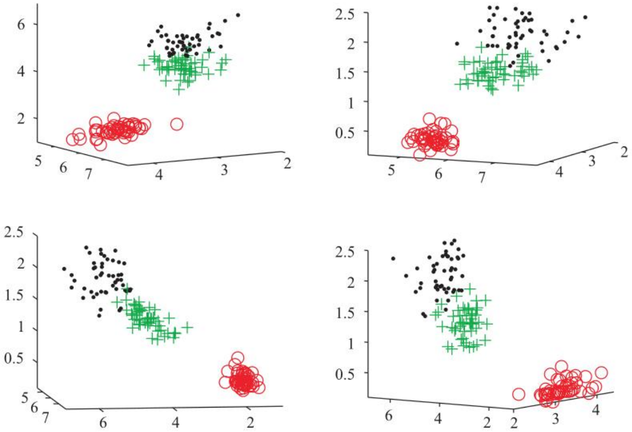


Figure 1.3 Three species of the iris data: dimensions 1, 2 and 3 (*top left*), dimensions 1, 2 and 4 (*top right*), dimensions 1, 3 and 4 (*bottom left*) and dimensions 2, 3 and 4 (*bottom right*).

1.2.2 Parallel Coordinate Plots (Inselberg, 1985)

1. As the dimension grows, 3D scatterplots become less relevant, unless we know that only some variables are important.
2. The idea of PCP is to present the data as __two-dimensional__ graphs:
   1. The variable numbers are represented as values on the y-axis.
   2. For a vector __$X = [X_1, \cdots, X_d]^T$__ we represent the first variable X_1 by the point $(X_1, 1)$ and the jth variable X_j by (X_j, j).
   3. Connect the d points by a __line__ which goes from $(X_1, 1)$ to $(X_2, 2)$ and so on to (X_d, d).
   4. Apply the same rule to the next d-dimensional datum.

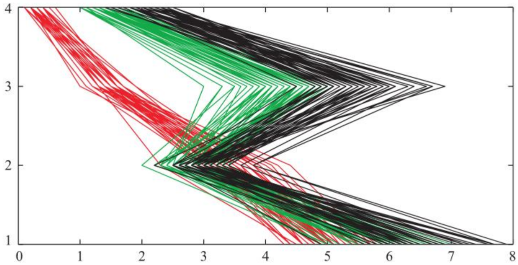


Figure 1.4 Iris data with variables represented on the y-axis and separate colours for the three species as in Figure 1.3.

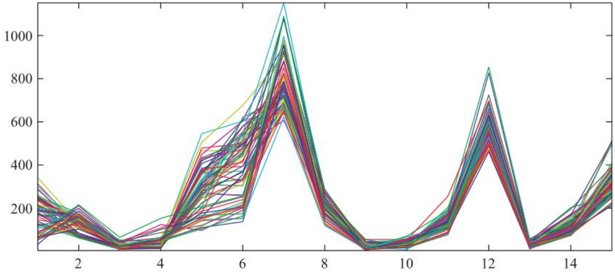


Figure 1.5 Parallel coordinate view of the illicit drug market data of Example 2.14.

1.2.2 Parallel Coordinate Plots (Inselberg, 1985)

1. A vertical PCP for Fisher's iris data (Figure 1.4): the data fall into two distinct groups, dimension 3 separates the two groups most strongly.
2. In a horizontal PCP (the 66 monthly observations on 15 features or variables of the illicit drug market data, Example 2.14), the x-axis represents the variable numbers $1, \ldots, d$.
   1. Two variables are excluded, as these have much higher values and would obscure the values of the remaining variables.
   2. Looking at variable 5, heroin overdose, the questionarises whether there could be two groups of observations corresponding to the high and low values of this variable.
3. Interactive graphical displays and movies are valuable visualisation tools.

1.3 Multivariate Random Vectors and Data

1. Random vectors are vector-valued functions defined on a <u>sample space</u>.
2. For a single random vector we assume that there is a <u>model</u> such as the first few moments or the distribution, or we might assume that the random vector satisfies a <u>"signal plus noise"</u> model.
3. We are then interested in deriving <u>properties</u> of the random vector under the model.
4. This scenario is called the <u>**population case**</u>.

1.3 Multivariate Random Vectors and Data

1. For a collection of random vectors, we assume the vectors to be __independent and identically distributed__ and to come from the same model.
2. Typically we do not know the true moments.
3. We use the __collection__ to construct estimators for the moments, and we derive __properties__ of the estimators.
4. Such properties may include how "good" an estimator is as the number of vectors in the collection grows, or we may want to draw __inferences__ about the appropriateness of the model.
5. This scenario is called the __**sample case**__.

1.3 Multivariate Random Vectors and Data

1. Refer to the collection of random vectors as the __data__ or the __(random) sample__.
2. In applications, specific values are measured for each of the random vectors in the collection. We call these values the __realised__ or __observed__ values of the data or simply the observed data.
3. The observed values are no longer random.
4. The distinction between the two scenarios is important, as we typically have to switch from the __population parameters__, such as the mean, to the __sample parameters__, in this case the sample mean.
5. The definitions for the population and the data are similar but not the same.

1.3.1 The Population Case

1. Let $\mathbf{X} = [X_1, \cdots, X_d]^T$ be a random vector from a distribution $F : \mathcal{R}^d \to [0, 1]$.
2. The individual X_j, with $j \leqslant d$, are random variables, (the components or entries) of $\mathbf{X}$, and $\mathbf{X}$ is d-dimensional or d-variate.
3. Assume that $\mathbf{X}$ has a finite d-dimensional mean or expected value $E(\mathbf{X})$ and a finite $d \times d$ covariance matrix $\mathrm{var}(\mathbf{X})$.
4. Write $\boldsymbol{\mu} = E(\mathbf{X})$ and $\Sigma = \mathrm{var}(\mathbf{X}) = E\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right]$.
5. The entries of $\boldsymbol{\mu}$ and Σ are

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \text{ and } \Sigma = \begin{Bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{Bmatrix}$$

where $\sigma_j^2 = \mathrm{var}(X_j)$ and $\sigma_{jk} = \mathrm{cov}(X_j, X_k)$.

1.3.1 The Population Case

1. Write $\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma)$ for a random vector $\mathbf{X}$ which has mean $\boldsymbol{\mu}$ and covariance matrix Σ.
2. If $\mathbf{X}$ is a d-dimensional random vector and A is a $d \times k$ matrix, for some $k \geqslant 1$, then $A^T\mathbf{X}$ is a k-dimensional random vector.
3. **Result 1.1** Let $\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma)$ be a d-variate random vector. Let A and B be matrices of size $d \times k$ and $d \times l$, respectively.
    1. The mean and covariance matrix of the k-variate random vector $A^T\mathbf{X}$ are
    $$A^T\mathbf{X} \sim (A^T\boldsymbol{\mu}, A^T\Sigma A).$$
    2. The random vectors $A^T\mathbf{X}$ and $B^T\mathbf{X}$ are uncorrelated if and only if $A^T\Sigma B = \mathbf{0}_{k\times l}$, where $\mathbf{0}_{k\times l}$ is the $k \times l$ matrix all of whose entries are 0.
4. Both these results can be strengthened when $\mathbf{X}$ is Gaussian.

1.3.2 The Random Sample Case

1. Let $\mathbf{X}_1, \cdots, \mathbf{X}_n$ be d-dimensional random vectors. Assume that the $\mathbf{X}_i$ are independent and from the same distribution $F : \mathcal{R}^d \to [0, 1]$ with finite mean $\boldsymbol{\mu}$ and covariance matrix Σ.
2. In statistics one often identifies a random vector with its observed values and writes $\mathbf{X}_i = \mathbf{x}_i$. We explore properties of random samples but only encounter observed values of random vectors in the examples.
3. Write $\mathcal{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n]$ for the sample of independent random vectors $\mathbf{X}_i$ and call this collection a random sample or data.

1.3.2 The Random Sample Case

1. Write

$$\mathcal{X} = \begin{bmatrix} X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \vdots & \vdots & \cdots & \vdots \\ X_{1d} & X_{2d} & \cdots & X_{nd} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{\cdot 1} \\ \mathbf{X}_{\cdot 2} \\ \vdots \\ \mathbf{X}_{\cdot d} \end{bmatrix}$$

2. The ith column of $\mathcal{X}$ is the ith random vector $\mathbf{X}_i$, and the jth row $\mathbf{X}_{\cdot j}$ is the jth variable across all n random vectors. The first subscript i in X_{ij} refers to the ith vector $\mathbf{X}_i$, and the second subscript j refers to the jth variable.

3. The sample mean $\bar{\mathbf{X}}$ and the sample covariance matrix $\mathbf{S}$ and sometimes write $\mathcal{X} \sim \mathrm{Sam}(\bar{\mathbf{X}}, \mathbf{S})$ in order to emphasise that we refer to the sample quantities, where

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \quad \text{and } \mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T.$$

1.3.2 The Random Sample Case

1. The sample mean and sample covariance matrix depend on the sample size n. If the dependence on n is important, for example, in the asymptotic developments, write $\mathbf{S}_n$ instead of $\mathbf{S}$.
2. Data are often centred. We write $\mathcal{X}_{cent}$ for the centred data

$$\mathcal{X}_{cent} \equiv \mathcal{X} - \bar{\mathbf{X}} = [\mathbf{X}_1 - \bar{\mathbf{X}}, \cdots, \mathbf{X}_n - \bar{\mathbf{X}}]$$

3. The centred data are of size $d \times n$. Using this notation, the $d \times d$ sample covariance matrix $\mathbf{S}$ becomes

$$S = \frac{1}{n-1}(\mathcal{X} - \bar{\mathbf{X}})(\mathcal{X} - \bar{\mathbf{X}})^T, \quad \text{with entries}$$

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - m_j)(X_{ik} - m_k)$$

with $\bar{\mathbf{X}} = [m_1, \cdots, m_d]^T$, and m_j is the sample mean of the jth variable.