Lecture 4

Multivariate visualization

- identify structure in data (e.g. existence of clusters of observations), outliers, etc.
- look at informative 2D scatterplots

Different approaches

① Find "interesting" 2D projections of the data.

$$X = \underset{1\times p}{\begin{pmatrix} \underset{\sim}{x}_1^T \\ \vdots \\ \underset{\sim}{x}_n^T \end{pmatrix}} \qquad A = \underset{p\times 2}{(\underset{\sim}{a}_1, \underset{\sim}{a}_2)} \qquad \text{Typically } \underset{\sim}{a}_1, \underset{\sim}{a}_2 \text{ orthogonal: } \underset{\sim}{a}_1^T \underset{\sim}{a}_2 = 0$$

$$Y = XA = \underset{n\times 2}{\begin{pmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n1} & y_{n2} \end{pmatrix}}$$

Plot $\{y_{i1}\}$ vs. $\{y_{i2}\}$

- $\underset{\sim}{a}_1, \underset{\sim}{a}_2$ should be chosen so that $\{\underset{\sim}{x}_i^T \underset{\sim}{a}_1\}, \{\underset{\sim}{x}_i^T \underset{\sim}{a}_2\}$ are "interesting" ← identify groups, outliers etc.

② Grand Tour Method

Asimov (1980s)

- do an exhaustive tour of all 2D projections t = time parameters

$$A(t) = (\underset{\sim}{a}_1(t) \;\; \underset{\sim}{a}_2(t))$$

$Y(t) = XA(t)$

↑

$n \times 2$ ⇒ scatterplot movie

- how to define $\underset{\sim}{a}_1(t)$ & $\underset{\sim}{a}_2(t)$ as t varies

Simple visualizations

Scatterplot matrix: Look at all $\binom{p}{2} = \frac{p(p-1)}{2}$ pairwise scatterplots of variables.

In R: pairs-displays scatterplots in a matrix

- more informative version of a correlation matrix
- non-linear pairwise dependencies revealed

Downside: Miss out on higher dimensional structure.

Mathematically, we're looking at projections $\underset{\sim}{a} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \underset{i\text{th-element}}{1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

How do we find "interesting" projections?

- how do we define "interesting"?

Random vector $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim f(\underset{\sim}{x})$ → joint distn

Sps that for any $\underset{\sim}{a}$ the dist'n of $\underset{\sim}{a}^T\underset{\sim}{X}$ is the same up to location & scale.
i.e. for some $\mu(\underset{\sim}{a})$, $\sigma(\underset{\sim}{a})$, the dist'n of $\frac{\underset{\sim}{a}^T\underset{\sim}{X}-\mu(\underset{\sim}{a})}{\sigma(\underset{\sim}{a})}$ is the same for all $\underset{\sim}{a}$

- suggests that all projections are uninteresting

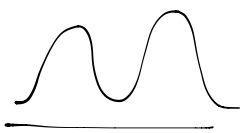
Multivariate normal has the property: all one dimensional projections are normal!

"Def'n": Non-interesting projection = projections that are normal.
Look for non-normal projections
Projection pursuit: (interesting)
- try to find non-normal projections of the data.
How to do this? → tractable
- need indices of non-normality
- $I(\underset{\sim}{a}^T\underset{\sim}{x}_1,\cdots,\underset{\sim}{a}^T\underset{\sim}{x}_n)$ is large if $\{\underset{\sim}{a}^T\underset{\sim}{x}_i\}$ is non-normal
  ↓ Index
- try to find $\underset{\sim}{a}$ to maximize I.

Example: Kurtosis
- for a random variable X with mean μ & variance σ^2 we define

$$K(X)=\frac{E[(X-\mu)^4]}{\sigma^4}$$
$$K(aX+b)=K(X)$$

If $X\sim N(\mu,\sigma^2)$ then $K(X)=3$
- short(light)-tailed dist'n $K<3$
- long(heavy)-tailed dist'n $K>3$
- multi-model dist'n $K>3$
  modes? or model?



Estimating Kurtosis
data $\underset{\sim}{x}_1,\cdots,\underset{\sim}{x}_n$ with mean $\bar{\underset{\sim}{x}}=\frac{1}{n}\sum_{i=1}^n\underset{\sim}{x}_i$ and covariance matrix $\frac{1}{n-1}\sum_{i=1}^n(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})^T$

$$\therefore \hat{K}(\underset{\sim}{a})=\frac{\frac{1}{n}\sum_{i=1}^n(\underset{\sim}{a}^T\underset{\sim}{x}_i-\underset{\sim}{a}^T\bar{\underset{\sim}{x}})^4}{(\underset{\sim}{a}^T\underset{\sim}{S}\underset{\sim}{a})^2}$$

- Find $\underset{\sim}{a}$ to maximize (or minimize) $\hat{K}(\underset{\sim}{a})$
- Alternative (and equivalently) we can maximize (or minimize) $\frac{1}{n}\sum_{i=1}^n(\underset{\sim}{a}^T\underset{\sim}{x}_i-\underset{\sim}{a}^T\bar{\underset{\sim}{x}})^4$ subject to $\underset{\sim}{a}^T\underset{\sim}{S}\underset{\sim}{a}=1$

$$\frac{1}{n}\sum_{i=1}^n(\underset{\sim}{a}^T\underset{\sim}{x}_i-\underset{\sim}{a}^T\bar{\underset{\sim}{x}})^4+\lambda[\underset{\sim}{a}^T\underset{\sim}{S}\underset{\sim}{a}-1]$$

$$\frac{\partial}{\partial\underset{\sim}{a}}:\ \frac{4}{n}\sum_{i=1}^n(\underset{\sim}{a}^T\underset{\sim}{x}_i-\underset{\sim}{a}^T\bar{\underset{\sim}{x}})^3(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})+2\lambda\underset{\sim}{S}\underset{\sim}{a}$$

$$\underset{\sim}{a}=\frac{-4}{2\lambda n}\sum_{i=1}^n(\underset{\sim}{a}^T\underset{\sim}{x}-\underset{\sim}{a}^T\bar{\underset{\sim}{x}})^3\underset{\sim}{S}^{-1}(\underset{\sim}{x}_i-\bar{\underset{\sim}{x}})$$

Example: Exam marks data 5 exam marks, 88 students

$$\underset{\sim}{a}=\begin{pmatrix}0.18\\0.76\\-0.61\\-0.09\\-0.02\end{pmatrix}$$

- maximized Kurtosis = 5.26

Isolate one observation as an outlier

$$\underset{\sim}{x}_{81} = \begin{pmatrix} 3 \\ 9 \\ 51 \\ 47 \\ 40 \end{pmatrix} \qquad \bar{\underset{\sim}{x}} = \begin{pmatrix} 39.0 \\ 30.6 \\ 30.6 \\ 46.7 \\ 42.3 \end{pmatrix}$$

- also look at $(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})^T S^{-1} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})$ ← compare to $\chi^2(5)$ quantiles
- minimized Kurtosis = 2.06

$$\underset{\sim}{a} = \begin{pmatrix} -0.45 \\ 0.84 \\ 0.30 \\ 0.02 \\ -0.05 \end{pmatrix}$$

From 1 to 2 dimensions

- maximize sample kurtosis $\hat{K}(\underset{\sim}{a})$ at $\underset{\sim}{a} = \underset{\sim}{a}_1$
- now maximize $\hat{K}(\underset{\sim}{a})$ over $\underset{\sim}{a}$ with $\underset{\sim}{a}^T \underset{\sim}{a}_1 = 0$

Maximize $\frac{1}{n} \sum_{i=1}^{n} (\underset{\sim}{a}^T \underset{\sim}{x} - \underset{\sim}{a}^T \bar{\underset{\sim}{x}})^4$ subject to $\underset{\sim}{a}^T S \underset{\sim}{a} = 1$ and $\underset{\sim}{a}^T \underset{\sim}{a}_1 = 0$
$\Rightarrow \underset{\sim}{a}_2$

Plot $\underset{\sim}{a}_1^T \underset{\sim}{x}_i$ vs $\underset{\sim}{a}_2^T \underset{\sim}{x}_i$ for $i = 1, \cdots, n$
Example: Exam mark data
- plots and R code on Blackboard

$$\underset{\sim}{a}_1 = \begin{pmatrix} -0.45 \\ -0.84 \\ 0.30 \\ 0.02 \\ -0.05 \end{pmatrix} \qquad \underset{\sim}{a}_2 = \begin{pmatrix} 0.40 \\ -0.17 \\ 0.02 \\ -0.82 \\ 0.36 \end{pmatrix}$$

Plot of $\underset{\sim}{a}_1^T \underset{\sim}{x}_i$ vs $\underset{\sim}{a}_2^T \underset{\sim}{x}_i$ again reveals: $\underset{\sim}{x}_{81}$ as unusual

Density Estimation

Given univariate or bivariate data

$$x_1, \cdots, x_n \qquad \underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$$

- want to estimate underlying density f.

Kernel density estimation (1 dimension)

- symmetric kernel function $K(x)$ ← density
  - $K(-x) = K(x)$
  - $K(x) \geq 0$
  - $\int_{-\infty}^{\infty} K(x)\,dx = 1$
- for some bandwidth parameter h, define $\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$
- R function: density
- Bandwidth h controls smoothness of $\hat{f}_h(x)$
- increase $h \Rightarrow$ smoother density estimate
- decrease $h \Rightarrow$ less smooth estimate i.e. more modes

Motivation

- spread out the probability mass of $\frac{1}{n}$ at each point
- at each x, add up contributions from each x_i
- around x_i, prob max ($= \frac{1}{n}$) is redist'd according to the density function

$$\frac{1}{nh} K\left(\frac{x-x_i}{h}\right)$$

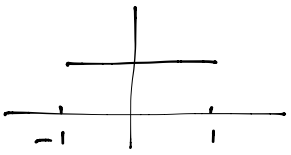
- note that $\int_{-\infty}^{\infty} \frac{1}{nh} K\left(\frac{x-x_i}{h}\right) dx$, $y = \frac{x-x_i}{h}$, $dx = h\,dy$

$$= \frac{1}{n} \int_{-\infty}^{\infty} K(y)\,dy = \frac{1}{n}$$

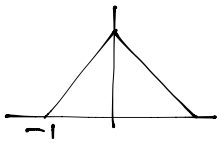
- add contributions from each x_i $\quad \hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$

Types of kernels

① Rectangular kernel $\quad K(x) = \frac{1}{2}$ for $|x| \le 1$

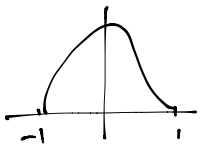


② Triangular kernel $\quad K(x) = 1 - |x|$ for $|x| \le 1$



③ Gaussian Kernel $\quad K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$

④ Epanechnikov kernel $\quad K(x) = \frac{3}{4}(1 - x^2)$ for $|x| \le 1$



Notes

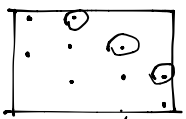
① Choice of kernel is much less important than bandwidth choice
- rectangular kernel should be avoided
- bias/variance tradeoff in choice of h
  - as $h \uparrow$, bias $\uparrow$, variance $\downarrow$
    $h \downarrow$, bias $\downarrow$, variance $\uparrow$
- given the true f, can derive optimal values of h
- automatic bandwidth selection in R is good.

② R-function density, scales the kernels to have variance 1

i.e. $\int_{-\infty}^{\infty} x^2 K(x)\,dx = 1$

- makes choice of bandwidth for different kernels easy → same h gives very similar estimates for different kernels.

From 1 to 2 dimensions (briefly)



- kernel density estimation more complicated in 2 dimensions