1

Looking at Outliers

Nick Fieller

Department of Probability & Statistics
University of Sheffield, UK

Manta Workshop,Tampere, May 2009,

2

Introduction

- Topics to be discussed:–
  - Introduction
    - What are outliers?
    - What are the objectives of analyses involving outliers?
  - Motivating data sets
  - Strategies for identification
    - pairwise original components or PCs
  - **Outlier Displaying components**
    - definition
    - exploitation
    - illustration

3

Introduction

- Outliers
  - '*Extreme*' observations which cause '***surprise or suspicion***' to the data analyst.
    - Possibly generated by some mechanism different from that governing the majority of the data
  - A ***discordant outlier*** is one which is sufficiently extreme for a formal test of the hypothesis that all observations are of the same random variable to be rejected (e.g 5%)
    - Typically tests of discordancy presume a location &/or scale shift alternative

4

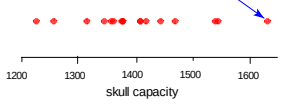
Introduction

- Examples:

Karl Pearson's Moriori skull capacities

1230	1318	1380	1420	**1630**	1378
1348	1380	1470	1445	1360	1410
1540	1260	1364	1410	1545	

An upper outlier, but not discordant



5

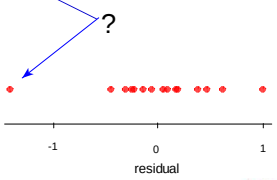
Introduction

- Exs (Ct[d])

Lt. Hearndon's measurements of Venus

–.30	+0.48	+0.63	–0.22	+0.18
–0.44	–0.24	–0.13	-0.15	+0.39
+1.01	+0.06	**–1.40**	+0.20	+0.10

A lower outlier, and discordant

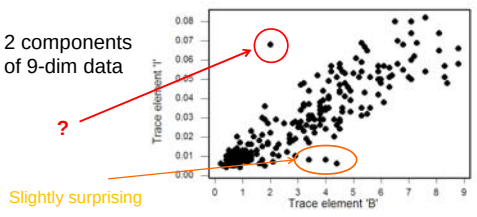


6

Introduction

- Univariate outliers 'stick out' at one end or the other
- Multivariate outliers "just stick out somewhere" (Anon)

2 components of 9-dim data



Introduction

7

- Objectives of analyses involving outliers
  - Identification
    - Obtain a deeper understanding of data
      - Perhaps correct errors in measurement or classification
      - Informal evaluation of likely effect on later analyses – will outlier bias later results?
      - Many outliers 'don't matter'
  - Test for discordancy
    - Fine for those addicted to hypothesis tests and p-values

Introduction

8

- Objectives (Ct[d])
  - Accommodation
    - Routine use of robust methods
      - 'blinkered approach to data analysis'
      - (blinkers – used on horses to shield their eyes so they only look straight ahead and so are *not frightened by the unexpected*)
  - Outliers are interesting in their own right
    - In moderate sized data sets there is the luxury of ***identification & investigation***
    - Moderate = a few hundred observations fewer than fifty dimensions

Data

9

- Motivating Data (an archaeological example)
  - Amounts of 9 *trace elements* in Greek ancient clay pots from known[??] sites
    - Profile of trace elements is specific to area made
    - Objective:
      to discover source of manufacture of some clay plots believed to have been made in some other city
      – evidence of trading between cities

Data

10

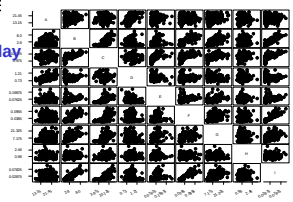
- A standard discrimination/classification problem [almost]
  - Need to have 'clean' training examples
  - Could be outliers since never 100% certain that the archaeologist is correct in belief of 'known sources'
    - e.g. a pot found in Thebes and believed to be made in Thebes might actually have been made in Sparta and so appear as an outlier on plot of Thebes data
- Full data set is 269 samples with 9 trace elements (A, B, …, I) from ~10 known sites
  - Subset from Thebes of 53 pots

Strategies

11

- Strategies for identification
  - Pairwise plots of individual components
    - Not useful for e

9 components is already too many for such a display
(some outliers visible, brushing reveals they are the same points)



Strategies

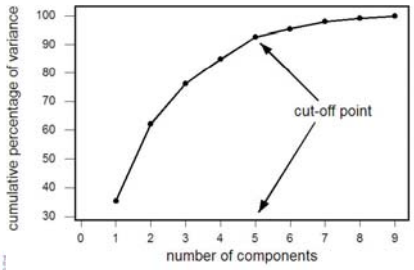
12

- Strategies for identification
  - Pairwise plots of individual components
    - Not useful for even moderate dimensions
  - Pairwise plots of principal components
    - Same problem with more than ~5
  - Pairwise plots of selected PCs
    - Authors disagree whether outliers appear on first few or last few components
  - I find plots on 'cut-off PCs' often useful

Strategies

13

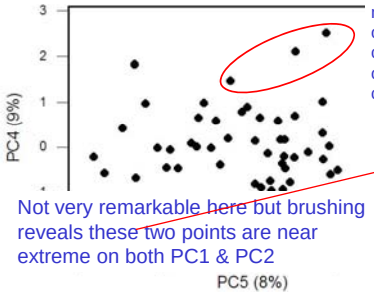
- Inherent dimensionality is around 5
  - Cut-off PC is 5th
  - Suggests plotting 4th, 5th (…) for revealing outliers



Strategies

14

- Plot on cut-off PCs (Thebes)



Brushing reveals these three have near identical values on other PCs so could be a group of outliers from a common source

Not very remarkable here but brushing reveals these two points are near extreme on both PC1 & PC2

ODCs

15

- None of these strategies for identification help in ***evaluation*** of impact of outliers nor understanding of their source
  - This is provided by

    Outlier Displaying Components

    - Essentially a 1–outlier ODC is the discriminant coordinate (crimcoord/canonical variate) between the outlier and the remaining (n–1) observations

ODCs

16

- **Key property of ODCs**
  - The 1–outlier ODC contains **all numerical information** on the discordancy of that outlier
    - i.e. value of discordancy test statistic for an observation calculated in p-dimensions is exactly same as the value calculated from the 1-dimensional projection of the data onto that observation's ODC
      - (assuming MV Normality and LRT against location shift alternatives)
  - Exploitation:
    - Use this dimension for displaying original & supplementary data

ODCs

17

- To demonstrate this consider UIT approach
  - **U**nion-**I**ntersection **T**est obtained by projecting data into one dimension & maximising value of equivalent 1-dim test statistics with respect to projection

ODCs

18

Let X' be $n \times p$ data matrix (centred to 0)

sample variance is $S = XX'/(n-1)$

The LRT statistic for testing x_j as discordant is

$$u_j = x_j' S^{-1} x_j$$

– *the squared Mahalanobis distance of x_j from the mean 0*

Projecting data into one dimension by $Y = \beta' X$

– (β a $p \times 1$ vector)

and letting $y_j = \beta' x_j$ gives discordancy test statistic for y_j as $U_j(\beta) = (n-1) y_j'(YY')^{-1} y_j = (n-1) x_j' \beta (\beta' XX' \beta)^{-1} \beta' x_j$

– *the squared Studentised distance of y_j from the mean 0*

19

ODCs

Need to maximise $U_j(\beta) = (n-1)x_j'\beta(\beta'XX'\beta)^{-1}\beta'x_j$ with respect to β

Noting invariance with respect to scale of β & so imposing a non-restrictive scale constraint that $\beta'XX'\beta=1$ and differentiating w.r.t. vector β converts problem to an eigenvalue problem,

showing that we need β to be the [right] eigenvector of $(n-1)(XX')^{-1}x_j\,x_j'=S^{-1}x_j\,x_j$ which is of rank 1 (and so has only one non-zero eigenvalue which is $x_j'S^{-1}x_j$ and eigenvector proportional to $S^{-1}x_j$, i.e. $\beta_{opt} = S^{-1}x_j$

Direct substitution shows $U_j(\beta_{opt}) = u_j$ i.e. numerical value of test statistic is preserved under optimal projection

20

ODCs

- Thus, the one-dimensional data projected onto the ODC capture all the numerical information on the discordancy of that observation
  - (though the reference distribution for a formal test of discordancy does depend on the dimensionality p).
- Picking that with the highest value of the statistic will reveal the most extreme outlier.
  - Generalization to two outliers & beyond depends on whether they arise from a common or two distinct slippages.
    - for three outliers there are five possibilities … … ….

21

ODCs

- Exploitation of ODCs
  - For pure detection use is limited to modest data sets with low contamination rates
    - Sensible to use your favourite **robust estimates** of location & variance
      - i.e. not sample mean and variance which are very sensitive to outliers
  - procedure can be effective for displaying outliers already identified & for plotting subsidiary data in same coordinate system
    - Examination of loadings of original components may give information on the nature of the outliers
      - in the spirit of union-intersection test procedures

22

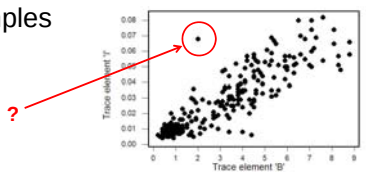
ODCs

- Since only 1 dimension is needed for display of outlier a second dimension can be chosen for displaying something else
  - Possibilities are PCs, component maximising variance subject to orthogonality with ODC, etc
- Note that a 1-outlier ODC is not in general orthogonal to any PC and the two ODCs for 2 distinct outliers are not orthogonal
  - just as canonical variates (discriminant coordinates) are not orthogonal but are plotted as if they are

23

ODCs

- Examples

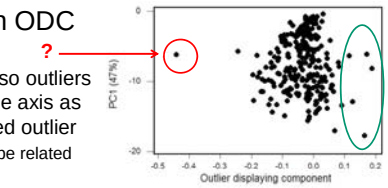


  - Outlier could reflect an abnormal **low** value on trace element B
    - or a recording error on element B
  - or an abnormal **high** value on trace element I
    - or a recording error on element I

24

ODCs

- Display on ODC

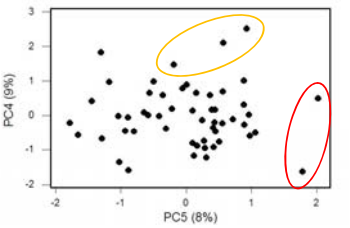


  - Note also outliers on same axis as identified outlier
    - maybe related
  - outlier display component is
    - (–0.0, 0.1, –0.0, 0.1, **0.3, –0.7**, 0.0, –0.1, **–6.0**)
    - (A, B, C, D, **E,** **F**, G, H, **I**)
    - suggests that if any recording error then it is on element I not on B

25

Strategies

- Plot on cut-off PCs for Thebes subgroup
  - rotate onto ODCs for 2 groups, i.e. one group of three with a common slippage and another of two with a distinct common slippage

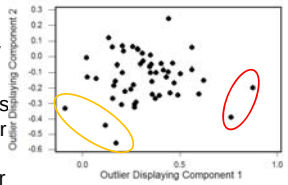


  - Superimposition of data from other cities may suggest whether these groups are actually foreign imports

26

ODCs

- Outliers not remarkable but use is to investigate source of highlighted pots by plotting further data as supplementary points.
- examination of loadings shows ODC for the pair is dominated by trace elements D & E and for the triple it is dominated by E, F and I (compare loadings of earlier example)
- Note two points at top of diagram may be worthy of examination



27

Final Comments

- Techniques are labour intensive
  - Need facilities for matrix manipulation and interactive brushing
    - S-plus, R, [Minitab]
- OK for modest data sets but not for routine analysis of high dimensional big data sets
  - Much better methods available though not for informal evaluation & assessment as shown above

28

http://nickfieller.staff.shef.ac.uk

nick.fieller@sheffield.ac.uk

29

30