Finding Multivariate Outlier

Applied Multivariate Statistics – Spring 2012
Goals

- Concept: Detecting outliers with (robustly) estimated Mahalanobis distance and QQ-plot
- R: chisq.plot, pcout from package “mvoutlier”
Outlier in one dimension - easy

- Look at scatterplots
- Find dimensions of outliers
- Find extreme samples just in these dimensions
- Remove outlier
2d: More tricky

No outlier in $x$ or $y$
Recap: Mahalanobis distance

- True Mahalanobis distance:

  \[ MD(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} \]

- Estimated Mahalanobis distance:

  \[ \hat{MD}(x) = \sqrt{(x - \hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu})} \]

Sq. Mahalanobis Distance \( MD^2(x) \)

\[ = \text{Sq. distance from mean in standard deviations} \]

IN DIRECTION OF X
Mahalanobis distance: Example

\[ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \]

\[ \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix} \]
Mahalanobis distance: Example

\[ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \]
\[ \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix} \]

\[ \mathbf{MD} = 4 \]

(20,0)
Mahalanobis distance: Example

\[
\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix},
\]

\[
\Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix}
\]

MD = 10
Mahalanobis distance: Example

\[ \mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \]
\[ \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 1 \end{pmatrix} \]

\( (10, 7) \)

\[ \text{MD} = 7.3 \]
Theory of Mahalanobis Distance

Assume data is multivariate normally distributed (d dimensions)

Mahalanobis distance of samples follows a Chi-Square distribution with d degrees of freedom

(“By definition”: Sum of d standard normal random variables has Chi-Square distribution with d degrees of freedom.)
Check for multivariate outlier

- Are there samples with estimated Mahalanobis distance that don’t fit at all to a Chi-Square distribution?
- Check with a QQ-Plot
- Technical details:
  - Chi-Square distribution is still reasonably good for estimated Mahalanobis distance
  - use robust estimates for $\mu, \Sigma$
Robust Estimates: Income of 7 people

Robust Scatter

Std. Dev.
Robust

Std. Dev.
Robust Estimates for outlier detection

- If scatter is estimated robustly, outlier “stick out” much more
- Robust Mahalanobis distance: Mean and Covariance matrix estimated robustly
Example - continued

Outlier easily detected!
Outliers in >2d can be well hidden!

No outlier, right?
Outliers in >2d can be well hidden!
Outliers in >2d can be well hidden!

This outlier can’t be seen in the scatterplot-matrix (but in a 3d plot)
Method 1: Quantile of Chi-Square distribution

- Compute for each sample (in d dimensions) the robustly estimated Mahalanobis distance $MD(x_i)$
- Compute the 97.5%-Quantile $Q$ of the Chi-Square distribution with d degrees of freedom
- All samples with $MD(x_i) > Q$ are declared outlier
Method 2: Adjusted Quantile

- Adjusted Quantile for outlier: Depends on distance between cdf of Chi-Square and ecdf of samples in tails
- Simulate “normal” deviations in the tails
- Outlier have “abnormally large” deviations in the tails (e.g. more than seen in 100 simulations without outliers)
Method 2: Adjusted Quantile

ECDF leaves “plausible” range

Defines adaptive cutoff
Method 2: Adjusted Quantile Function “aq.plot”
Method 3: State of the art - pcout

- Complex method based on robust principal components
- Pretty involved methodology
- Very fast – good for high dimensions

- R: Function “pcout” in package “mvoutlier”
- $w_{final01}: 0$ is outlier
- $w_{final}:$ Small values are more severe outlier

Automatic outlier detection

- It is **always better** to look at a QQ-plot to find outlier! Just find points “sticking out”; no distributional assumption
- If you can’t: Automatic outlier detection
  - finds usually too many or too few outlier depending on parameter settings
  - depends on distribution assumptions (e.g. multivariate normality)
- + good for screening of large amounts of data
Concepts to know

- Find multivariate outlier with robustly estimated Mahalanobis distance
- Cutoff
  - by eye (best method)
  - quantile of Chi-Square distribution
R commands to know

- chisq.plot, pcout in package “mvoutlier”
Next week

- Missing values