A New Contamination Model

Ruben Zamar
Department of Statistics, UBC

October 21, 2015

A NEW ROBUSTNESS MODEL

More General (and Realistic) Robustness Model

- "Classical" robust methods work as follows

More General (and Realistic) Robustness Model

- "Classical" robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)

More General (and Realistic) Robustness Model

- "Classical" robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases

More General (and Realistic) Robustness Model

- "Classical" robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases
- Important assumption underlying classical robust procedures

More General (and Realistic) Robustness Model

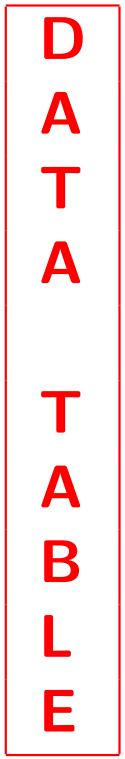
- "Classical" robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases
- Important assumption underlying classical robust procedures
  - Percentage of unusual data points is relatively small

More General (and Realistic) Robustness Model

- "Classical" robust methods work as follows
  - Identify unusual data points in the dataset (rows in the data table)
  - Downweight the unusual data cases
- Important assumption underlying classical robust procedures
  - Percentage of unusual data points is relatively small
  - Hopefully way below 50%

The Shape of Data Tables

- Typical data table (before)



LARGE and Shallow Data Tables

- In some applications we deal with datasets like this

Some Examples of This Type of Data

- Microarray data

p	number of genes	several thousands
n	number of patients	at best a few hundreds

Some Examples of This Type of Data

- Microarray data

p	number of genes	several thousands
n	number of patients	at best a few hundreds

- Asthenosphere Data

p	number of locations	about 5000
n	number of days	3650 days

Implications for Robustness Methods

- Downweighting entire rows may be too "wasteful"

Implications for Robustness Methods

- Downweighting entire rows may be too "wasteful"
- Rows may be only partially spoiled

Implications for Robustness Methods

- Downweighting entire rows may be too "wasteful"
- Rows may be only partially spoiled
- Consider "cell contamination" as opposed to "row contamination"

Implications for Robustness Methods

- Downweighting entire rows may be too "wasteful"
- Rows may be only partially spoiled
- Consider "cell contamination" as opposed to "row contamination"
- Need for more flexible robustness methods and models

An Example: Cell-wise Contamination Model

- From Alqallaf's PhD thesis and Alqallaf et al (2009)

$$\mathbf{X} = (I - B)\,\mathbf{Y} + B\mathbf{Z}$$

$$B = diag\,(B_1, B_2, ..., B_p)$$

$$P\,(B_i = 1) = 1 - P\,(B_i = 0) = \epsilon_i$$

An Example: Cell-wise Contamination Model

- From Alqallaf's PhD thesis and Alqallaf et al (2009)

$$\mathbf{X} = (I - B)\,\mathbf{Y} + B\mathbf{Z}$$

$$B = diag\,(B_1, B_2, ..., B_p)$$

$$P\,(B_i = 1) = 1 - P\,(B_i = 0) = \epsilon_i$$

- Lot's of room for research at the MSc and PhD levels on this area

Open Problems

- How to measure robustness in this more general context?

Open Problems

- How to measure robustness in this more general context?
- How to achieve robustness in this more genreral context?

Open Problems

- How to measure robustness in this more general context?
- How to achieve robustness in this more genreral context?
- Some initial work was done in Alqallaf (2009), for the multivariate location/scatter model

Open Problems

- How to measure robustness in this more general context?
- How to achieve robustness in this more genreral context?
- Some initial work was done in Alqallaf (2009), for the multivariate location/scatter model
- Lot's of room for research at the MSc and PhD levels on this area