Detecting Multivariate Financial Data Outliers using Calibrated Robust Mahalanobis Distances

Chris Green*

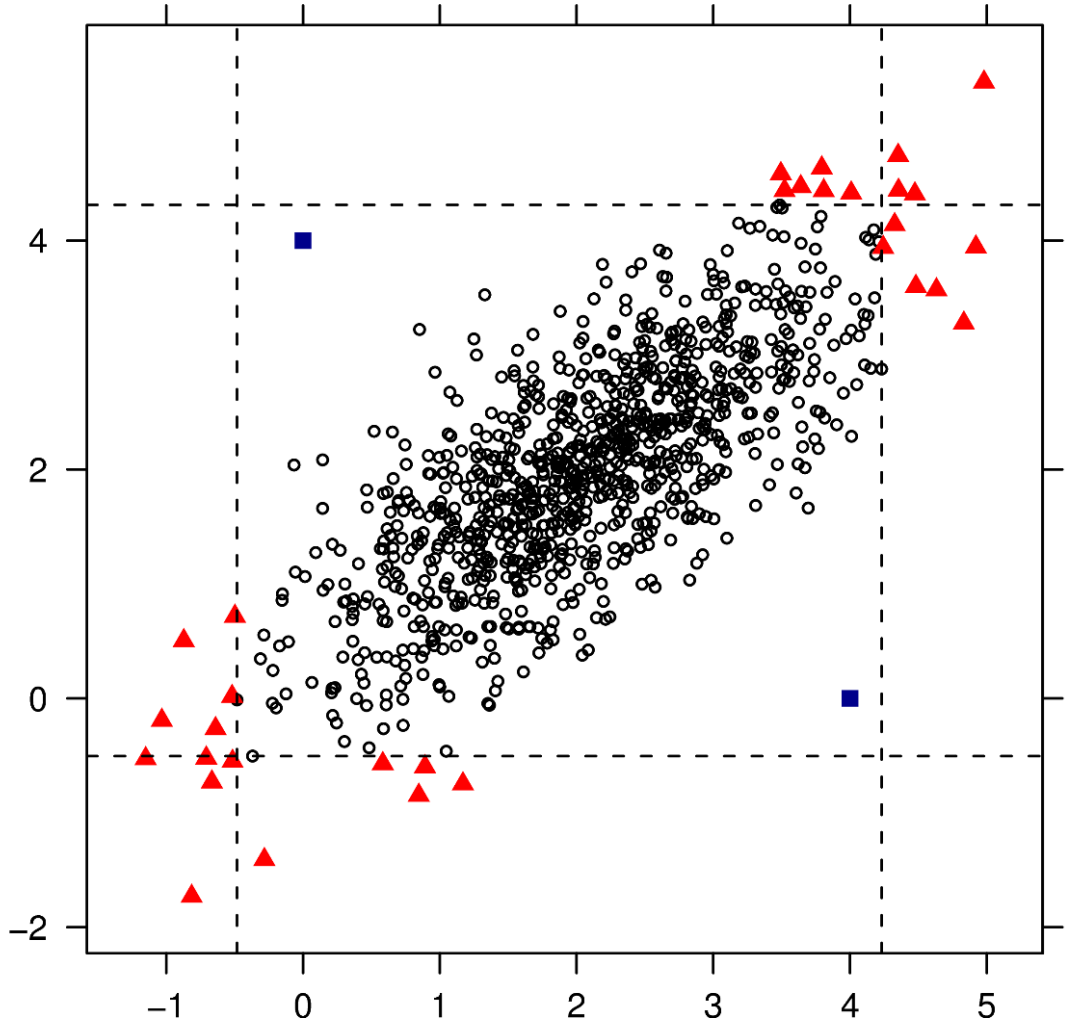
Doctoral Candidate
Department of Statistics
University of Washington
cggreen@uw.edu
http://students.washington.edu/cggreen/uwstat/

R/Finance 2015, May 29-30, 2015, University of Illinois, Chicago

* Joint work with R. Douglas Martin (University of Washington)

Multivariate Outliers May Not Be Mitigated By Univariate Techniques!

- Outliers are deviations from the model followed by the “bulk” of the data
- Will probably influence the construction of financial models (e. g., fundamental factor models)
- Commonly use 1-D trimming or Winsorization to deal with outliers in each variable (red triangles)
- This approach can miss multivariate outliers (blue squares)



Robust Mahalanobis Distances Are An Easy Way To Detect Multivariate Outliers

- Squared Mahalanobis distance (MD) of observation $\boldsymbol{x}$ from $\boldsymbol{\mu}$

$$\boldsymbol{d^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)}$$

- Plug in classical mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ estimates for unknown mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
- Problem: classical mean and covariance are not robust to outliers
- Solution: replace classical estimates with robust ones
  - Minimum Covariance Determinant (MCD) is a commonly-used robust estimate of dispersion which can be used to construct robust MDs

For MCD-Based Robust Mahalanobis Distances, Cerioli (2010) Provides An Accurate Test Procedure

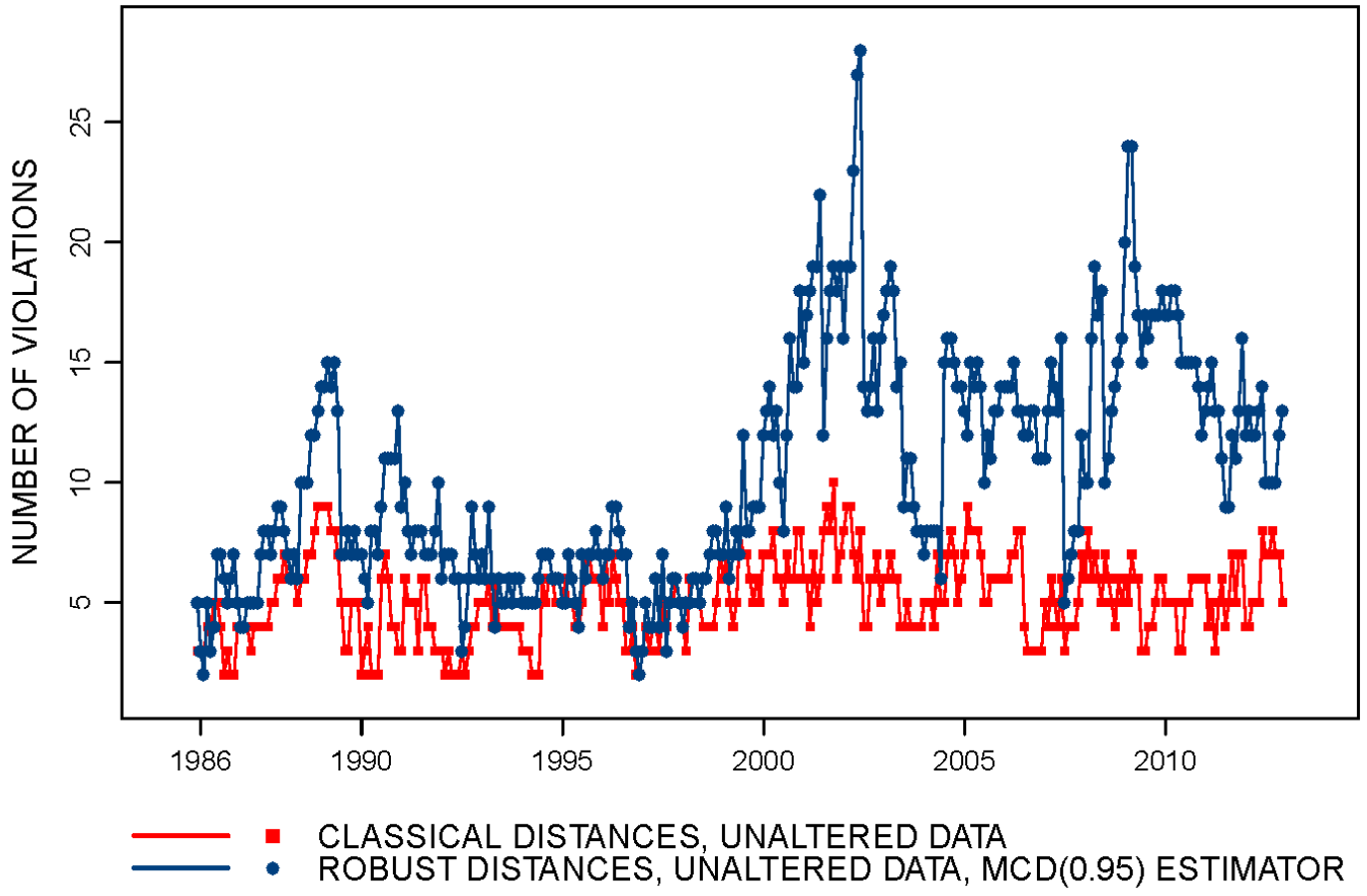
- Detect outliers by looking for MD that are "too large"
- Classical approach: test MD against a chi-squared χ^2_ν critical value (ν is the dimension)
- Hardin and Rocke (2005), Cerioli et al. (2009) showed that the χ^2_ν leads to tests with more false-positive detections than expected for robust MD
- Cerioli (2010) developed a calibration methodology, Iterated Reweighted MCD (IRMCD), which provides outlier detection tests with the correct Type I error behavior for robust MD
- We implemented this method in the `CerioliOutlierDetection` R package

Example: Outlier Detection In Financial Data Using Calibrated Robust MDs

- Apply Cerioli's IRMCD method to data one might use to build a fundamental factor model
  - Large-cap stocks (from CRSP database via WRDS)
  - Book-to-market, size (log. of mkt. cap.), earnings-to-price, momentum (12 mo. MA of returns) (from Compustat via WRDS)
  - 325 months of data, Dec. 1985-Dec. 2012
- Test for outliers using Cerioli approach and Bonferroni-corrected significance level of $0.025/325 \approx 0.00008$.
- Use a conservative version of MCD that uses approx. 95% of the data to estimate the robust dispersion matrix

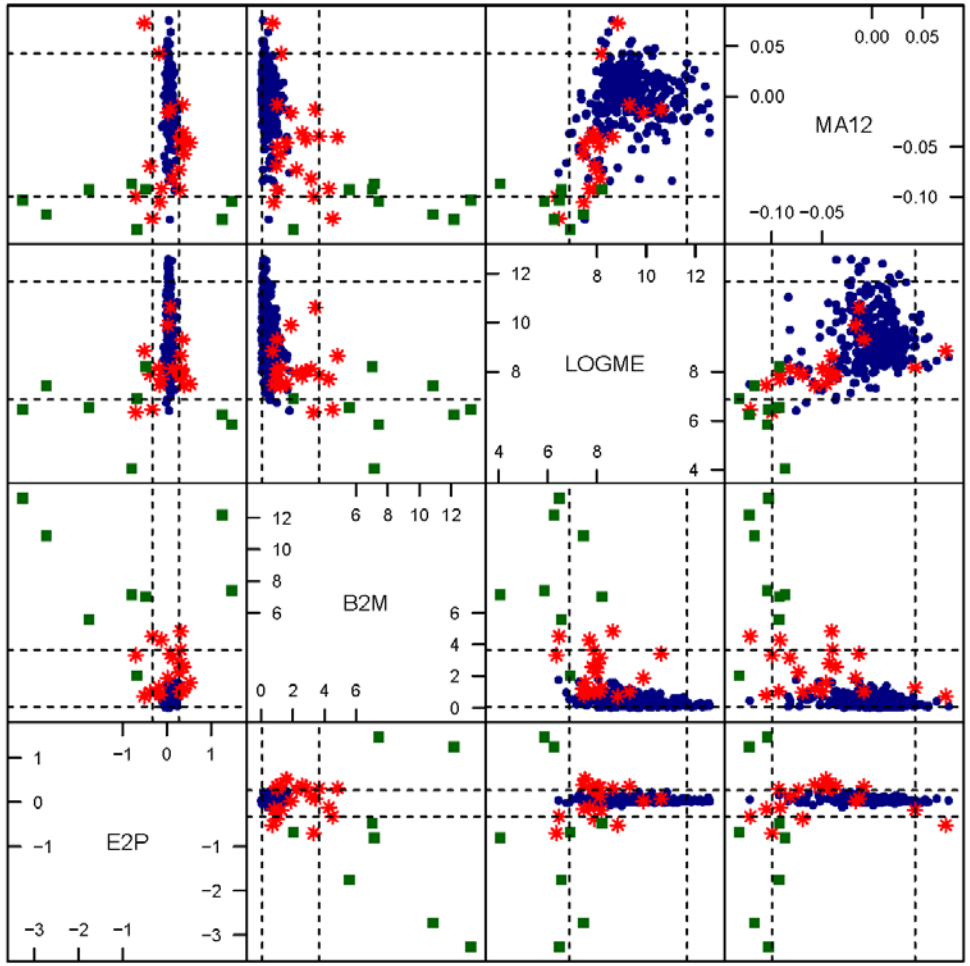
Robust MDs Detect Far More Outliers In The Data Than Classical MDs

- Classical MDs (red) suggest calm, unchanging markets 1985–2012
- Robust MDs (blue) suggest significant departures from multivariate normality, especially after the dot-com crash, 9/11, 2008 financial crisis



Univariate Trimming/Winsorization Would Not Have Mitigated All The Outliers

- June 2002 snapshot showing outliers found using classical and robust MDs (green squares) and those found only by robust MDs (red stars)
- Dashed lines are 2.5% trimming/Winsorization boundaries
- Many multivariate outliers are inside the box defined by the trimming boundaries and would not have been touched by univariate trimming!



Robust, Multivariate Methods Are The Right Way To Find Outliers In Factor Model Data

- We've demonstrated that multivariate outliers (1) exist in fundamental factor data and (2) may not be mitigated by 1-D trimming or Winsorization

- Current research work revolves around how multivariate outliers affect fundamental factor models
  - Many different ways outliers can arise in the data: errors, firm-specific events, market-wide events
  - Might want different approaches for different types of outliers

Thank You!

- R package `CerioliOutlierDetection`
  - available on CRAN:

http://cran.r-project.org/web/packages/CerioliOutlierDetection/index.html

- Working Paper
  - Green and Martin, “Diagnosing the Presence of Multivariate Outliers in Financial Data using Calibrated Robust Mahalanobis Distances” (2015). Available from http://students.washington.edu/cggreen/uwstat/papers/mvoutliersfinance.pdf

References

- Cerioli, A. (2010). “Multivariate outlier detection with high-breakdown estimators.” *Journal of the American Statistical Association*, 105(489):147–156.
- Cerioli, A., Riani, M., and Atkinson, A. C. (2009). “Controlling the size of multivariate outlier tests with the MCD estimator of scatter.” *Statistical Computing*, 19:341–353.
- Hardin, J. and Rocke, D. M. (2005). “The distribution of robust distances.” *Journal of Computational and Graphical Statistics*, 14:928–946.