Actuaries Institute

# Recent outlier detection methods with illustrations in loss reserving

**Benjamin Avanzi**, **Mark Lavender**, **Greg Taylor, Bernard Wong**
School of Risk and Actuarial Studies, UNSW Sydney

Insights, 18 September 2017

Actuaries Institute

# The Reserving Problem

| $i/j$ | **1** | **2** | $\cdots$ | $j$ | $\cdots$ | $J$ |
|---|---|---|---|---|---|---|
| **1** | $X_{1,1}$ | $X_{1,2}$ | $\cdots$ | $X_{1,j}$ | $\cdots$ | $X_{1,J}$ |
| **2** | $X_{2,1}$ | $X_{2,2}$ | $\cdots$ | $X_{2,j}$ | $\cdots$ | |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | |
| $i$ | $X_{i,1}$ | $X_{i,2}$ | $\cdots$ | $X_{i,j}$ | | |
| $\vdots$ | $\vdots$ | $\vdots$ | | | | |
| **I** | $X_{I,1}$ | | | | | |

Figure: Aggregate claims run-off triangle

- Complete the square (or rectangle)
- Also - multivariate extensions.

# Common Reserving Techniques

- Deterministic Chain-Ladder
- Stochastic Chain-Ladder (Hachmeister and Stanard, 1975; England and Verrall, 2002)
- Mack's Model (Mack, 1993)
- GLMs (Wright, 1990)
- Regression - e.g. Taylor and Ashe (1983); Taylor (1988); Kremer (1997)
- Bornhuetter-Ferguson technique (Bornhuetter and Ferguson, 1972)

# One of the (many) challenges for practitioners

Imagine one of the observations in the triangle is fairly large (or small)—larger (smaller) than you would generally expect (*really*?). What do you do?

- Even one data point may have a major impact on the outcome—either positive or negative
(refer to our previous GIS presentation, and associated paper Avanzi, Lavender, Taylor, and Wong, 2016)
- on the one hand, to adjust it too much will lead to an under- (over-)estimation of the reserve
- on the other hand, to leave it may lead to an over- (under-)estimation of the reserve

How can we determine in a rigorous way whether this is an outlier?
How can we best allow for this / adjust the observation?



Actuaries Institute

## Outliers in Reserving (Van Wouwe et al., 2009)

- One or more observations differ from majority
- Can be due to
  - Exceptional circumstances
  - Mistakes
- Outliers are present in real data (Verdonck, Van Wouwe, and Dhaene, 2009; Van Wouwe, Verdonck, and Van Rompay, 2009; Verdonck and Van Wouwe, 2011)
- Existance of such outliers can potentially have significant impact on the reserves and also the Standard Error of the reserve estimate.

# Robustness: Definition

**Ability of model or estimation procedure to not be overtly influenced by outliers and/or deviations from assumptions.**

Why Make Something Robust?

- More closely reflect reality → Enhanced accuracy & reliability
- Guard against oversimplification for mathematical convenience
- Identify and further inspect aberrations in the data

What are the trade-offs?

- Bias *(data is manipulated)*
- Efficiency *(precisely wrong vs approximatively right)*
  - A tale of two extremes
- Complexity *(should we bother?)*

# Illustration - Efficiency Trade-Off

Standard Error of Reserves are

- Marginally increased when no outliers are present
- Always decreased when one outlier present. Often significantly (e.g. $X_{1,9}$ and $X_{2,4}$)

| outl | clasCL | robCL |
|---|---|---|
| — | 1,599,681 | 2,190,625 |
| $X_{11}$ | 2,707,399 | 2,138,764 |
| $X_{12}$ | 2,508,137 | 2,203,704 |
| $X_{13}$ | 3,121,649 | 2,350,954 |
| $X_{14}$ | 2,841,367 | 2,261,444 |
| $X_{15}$ | 3,904,724 | 2,320,335 |
| $X_{16}$ | 4,688,239 | 2,048,458 |
| $X_{17}$ | 2,237,214 | 2,208,538 |
| $X_{18}$ | 2,465,573 | 2,338,262 |
| $X_{19}$ | 3,171,971 | 2,037,788 |
| $X_{1,10}$ | 2,075,936 | 2,159,432 |
| $X_{21}$ | 2,378,233 | 2,366,410 |
| $X_{22}$ | 2,529,390 | 2,258,192 |
| $X_{23}$ | 3,520,701 | 2,215,397 |
| $X_{24}$ | 4,510,697 | 2,365,810 |
| $X_{25}$ | 3,342,324 | 2,216,542 |

(Verdonck, Van Wouwe, and Dhaene, 2009)

# Robustness in Reserving Models

- Deterministic chain-ladder:
  - Development factors are essentially means, and means are non-robust.
- Stochastic chain-ladder :
  - Maximum or quasi-likelihood estimation of parameters is non-robust
    (Pregibon, 1982; Künsch, Stefanski, and Carroll, 1989; Verdonck and Debruyne, 2011)

(applies also to other parametric reserving techniques)

# Robustness in Reserving - Additional complications

Motivation for our work:

- Claims data is **often skewed** $\rightarrow$ not fully accounted for in existing robust reserving
(see, e.g. Verdonck and Van Wouwe, 2011).
- General insurers often operate across multiple lines of business $\rightarrow$ robust **multivariate technique** required

# Summary of problem

Loss reserving techniques are based on simplifying assumptions.
Not all observations will conform to these assumptions.

## Challenges

- Identifying outliers in a statistically rigorous manner, including observations that violate the dependence structure of multiple data sets
  - Misclassification risk
  - Particularly when skewness is present
- Adjusting detected outliers appropriately

# Introduction - Robustness criteria

**Three fundamental features** (Huber and Ronchetti, 2009):

1. **Efficiency**: Optimal or nearly optimal efficiency at the assumed model
2. **Stability**: Small divergences should only have a minor effect on performance
3. **Breakdown**: Moderately greater divergences should not lead to a disaster

# Robust Statistical Techniques

- Heuristic Tools - Facilitate understanding of the robustness of the procedure
- Robust M-estimation - ensure parameters are estimated in a robust fashion
- Outlier Detection - Identify aberrant observations

# Heuristic Tools

- Influence Functions (Huber and Ronchetti, 2009; Hampel, 1968)

$$IF(x, F, T) = \lim_{h \to 0} \frac{T((1-h)F + h\delta_x) - T(F)}{h}$$

  - Verdonck and Debruyne (2011): Influence functions for estimates of $\alpha_i$, $\beta_j$, future claims & reserves in Poisson GLM CL unbounded.

# Heuristic Tools

- Impact Functions (Venter and Tampubolon, 2008; Avanzi, Lavender, Taylor, and Wong, 2016)

$$IF_{k,j}(T) = \frac{\partial T}{\partial X_{k,j}}$$

  - Highlight sensitivity of statistics to individual observations
  - Venter and Tampubolon (2008): Calculate impact functions and GDFs (Ye, 1998) for a range of models.
- Gross Error Sensitivity
- Breakdown Point

## M-estimators

$$\sum m(x_i; T_n) = \text{min!}$$

Where $T_n$ is the estimate

OR implicitly

$$\sum \psi(x_i; T_n) = 0$$

where $\psi$ represents the derivative of $m$

Not necessarily robust in their own right (Pregibon, 1982; Künsch, Stefanski, and Carroll, 1989; Stefanski, Carroll, and Ruppert, 1986) $\rightarrow$ depends how observations are weighted

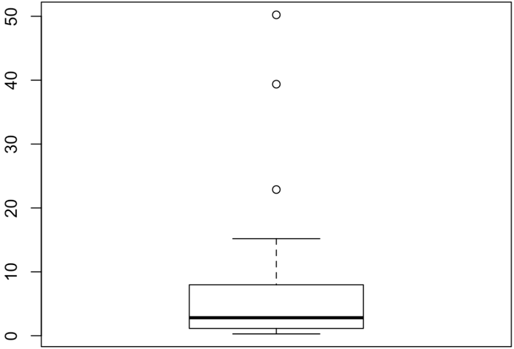
Verdonck and Debruyne (2011) provide a robust GLM reserving methodology by utilising a Huber Function.

# Common Outlier Detection Techniques

- Boxplots (Tukey, 1977; Hubert and Vandervieren, 2008)
- Bagplots (Rousseeuw, Ruts, and Tukey, 1999)
- *bagdistance* (Hubert, Rousseeuw, and Segaert, 2016)
- Adjusted Outlyingness (Hubert and Van der Veeken, 2008)
- MCD (Rousseeuw, 1984) Mahalanobis Distance

# BoxPlot

- Univariate data: $x_1, x_2, ..., x_n$
- $Q_1$: First quartile
- $Q_3$: Third quartile
- IQR: $Q_3 - Q_1$
- Non-outlying observations: $[Q1 - 1.5^{*}\text{IQR}, Q3 + 1.5^{*}\text{IQR}]$


0
10
20
30
40
50

# Adjusted Boxplot

- $\text{med}_n$: median of the data
- $h(x_i, x_j)$: $\frac{(x_j - \text{med}_n) - (\text{med}_n - x_i)}{x_j - x_i}$, $x_i < \text{med}_n < x_j$
- $MC(X_n) = \text{med}_{x_i < \text{med}_n < x_j} h(x_i, x_j)$
- Adjusted boxplot:

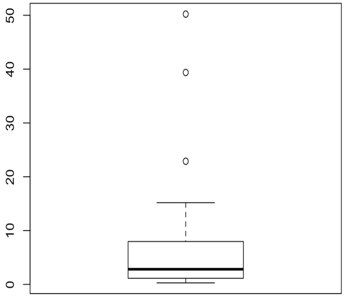
$$[Q1 - 1.5e^{-4MC}\text{IQR}, Q3 + 1.5e^{3MC}\text{IQR}],\ \ MC > 0$$

$$[Q1 - 1.5e^{-3MC}\text{IQR}, Q3 + 1.5e^{4MC}\text{IQR}],\ \ MC < 0$$

# Boxplot and Adjusted Boxplot

**Classical**

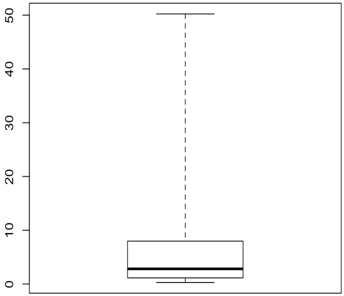
(Tukey, 1977)




$$[Q1 - 1.5IQR, Q3 + 1.5IQR]$$

**Skew-Adjusted**

(Hubert and Van der Veeken, 2008)




$$[Q1 - h_l(MC)IQR, Q3 + h_u(MC)IQR]$$

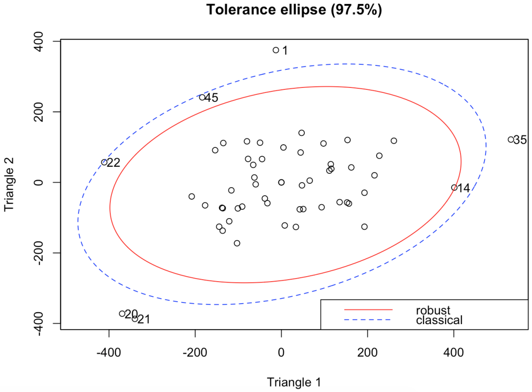
# Boxplots in Robust Reserving Literature

Deterministic Robust Chain-Ladder of
Verdonck, Van Wouwe, and Dhaene (2009):

- Applies classical boxplot on residuals after testing for normality of data at two stages in methodology
- Recommends utilising Skew-Adjusted boxplot if normality rejected

# Minimum Covariance Determinant (MCD)

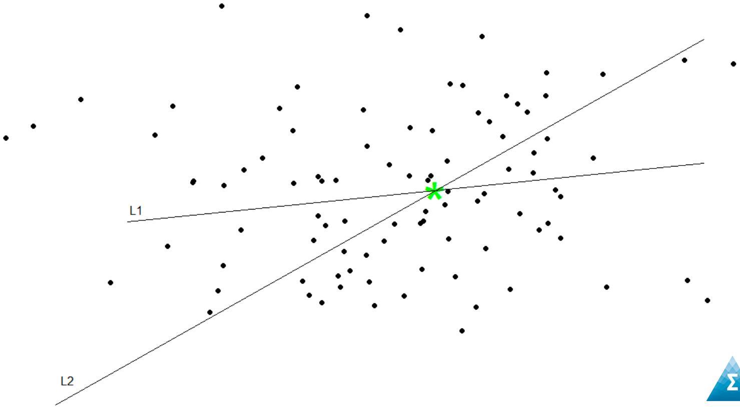
$$MD(\mathbf{x}_i) = \sqrt{(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_0)' \Sigma_0^{-1} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_0)} \qquad \text{(Rousseeuw, 1984)}$$



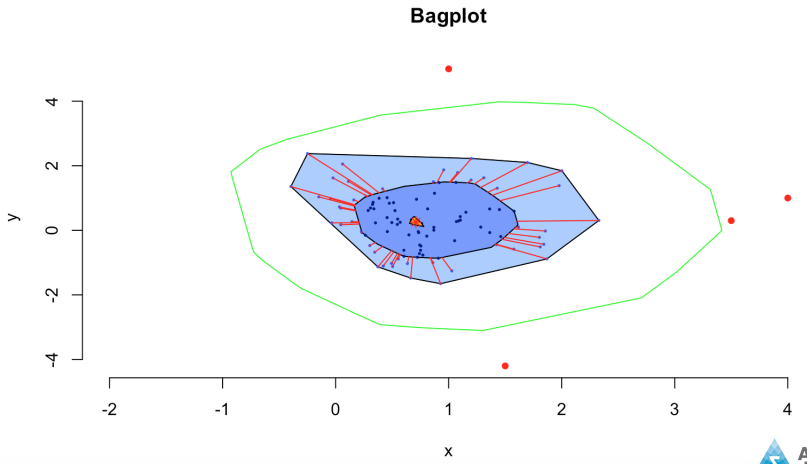

Adjust outliers by bringing them back to the tolerance ellipse given by the 95$^{th}$ percentile of $\chi^2$ distribution

# Halfspace Depth (Tukey, 1975)

- Bounded Influence Function (Romanazzi, 2001)
- Tukey Median has high Breakdown Point (up to $\frac{1}{3}$) (Donoho and Gasko, 1992)



# Bagplot



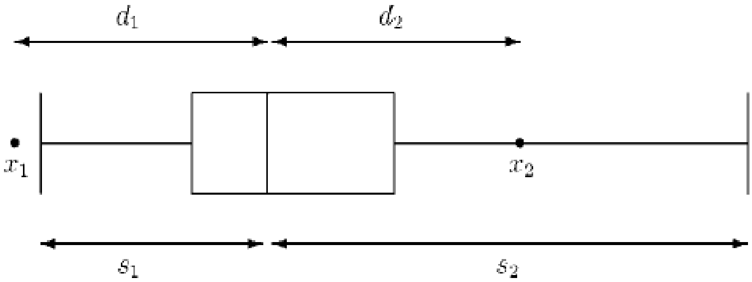
# Adjusted Outlyingness

- Adjustment of Stahel-Donoho Outlyingness (Donoho, 1982) to account for skewness
- Adjusted Outlyingness:

$$\mathrm{AO}_i = \mathrm{AO}^{(1)}(x_i, X_n) = \begin{cases} \frac{x_i - \mathrm{med}(X_n)}{w_2 - \mathrm{med}(X_n)}, & \text{if } x_i > \mathrm{med}(X_n) \\ \frac{\mathrm{med}(X_n) - x_i}{\mathrm{med}(X_n) - w_1}, & \text{if } x_i < \mathrm{med}(X_n) \end{cases}$$

Actuaries Institute

- $x_1$ and $x_2$ are the same distance to the median
- But the scales are different
- $x_1$ is an outlier, $x_2$ is not




(Hubert and Van der Veeken, 2008)

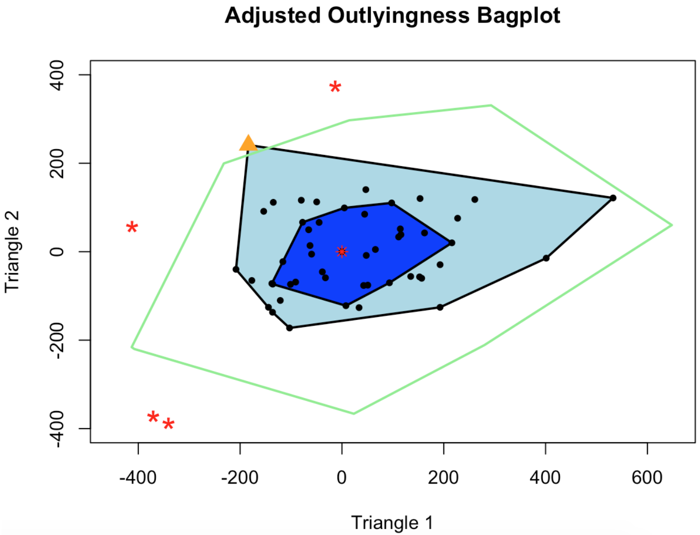
# Multivariate Data - Adjusted Outlyingness

- Project the data in the direction $\mathbf{a}$ to maximize the univariate outlyingness measure
- If large AO in one direction then observation may be an outlier
- Adjusted Outlyingness:

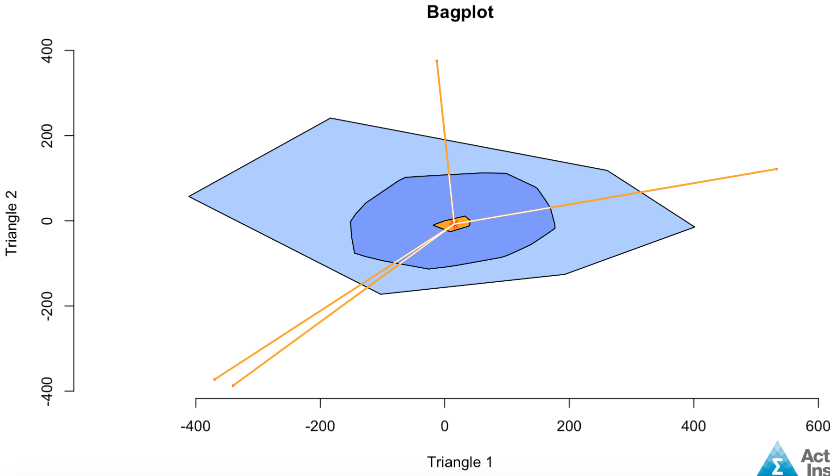
$$\mathrm{AO}_i = \mathrm{AO}(\mathbf{x}_i, \mathbf{X}_n) = \sup_{\mathbf{a} \in \mathbb{R}^p} \mathrm{AO}^{(1)}(\mathbf{a}^T \mathbf{x}_i, \mathbf{X}_n \mathbf{a})$$

- Plot Adjusted Boxplot for all $\mathrm{AO}_i$
- Cutoff value: $Q_3 + 1.5e^{3MC}\mathrm{IQR}$

# AO-Based Bagplot



# *bagdistance*



# Recent Developments

**mrfDepth Package**

- Previously, bagplot fence factor was chosen to be 3.
- Hubert, Rousseeuw, and Segaert (2015) derived distribution of *bagdistance* when underlying data is elliptical.
- As a result Segaert, Hubert, Rousseeuw, Raymaekers, and Vakili (2017) have developed a new fence factor given by $\sqrt{\chi^2_{\{99,N\}}}$
- This new cut-off has more rigorous foundations.
- A new cutoff value for Adjusted Outlyingness given by $\sqrt{\chi^2_{\{99,N\}}} \cdot \text{median}(AO)$

Actuaries Institute

# Robust Bivariate Chain-Ladder

## General Approach

1. Perform robust GLM chain-ladder on each triangle separately
2. Store residuals as bivariate data set
3. Detect bivariate outliers
4. Treat outliers
5. Perform multivariate time series chain-ladder

# Existing Robust Bivariate Chain-Ladder

## Shortcomings of Current Techniques

- MCD Mahalanobis Distance
  - Does not incorporate skewness at all

    $\rightarrow$ FALSE or MISSED detections
- Bagplot
  - Does not provide unique measure of outlyingness
    $\rightarrow$ limits communicability and potential treatment options
  - Only can handle mild skewness
  - Computationally expensive
    $\rightarrow O(n^2(\log n)^2)$.

# New Robust Bivariate Chain-Ladder

## New Approaches

- Adjusted Outlyingness
  - Explicitly incorporates robust measure of skewness
    $\rightarrow$ Suitable for more extreme levels of skewness
  - Provides unique measure of outlyingness
  - Less computationally expensive than bagplot ($O(mnp \log n)$)
    $\rightarrow$ Important for extensions to higher dimensions
- *bagdistance*
  - Derived from bagplot
  - Provides unique measure of outlyingness
    $\rightarrow$ Can determine how outlying observations are
    $\rightarrow$ More treatment options become available

# Robust Bivariate Chain-Ladder Example

- Real data over 10 years from Shi, Basu, and Meyers (2012)
- Personal and Commercial Auto Insurance Lines
- 55 observations
- Bivariate Skewness = 7.21

# MCD Mahalanobis Distance

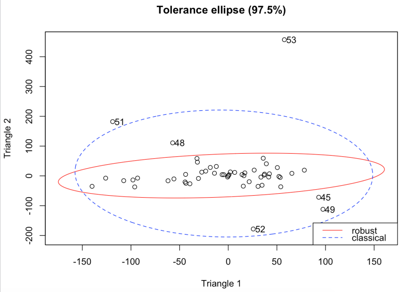



Figure: Initial Tolerance Ellipses

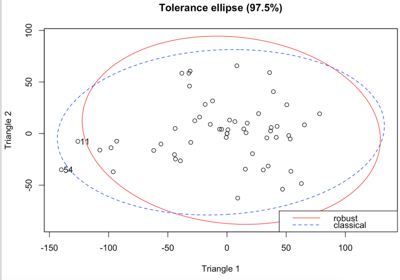



Figure: Tolerance Ellipses After Adjustment

# Bagplot

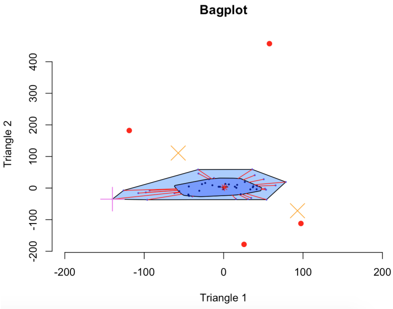



Figure: Bagplot without fence

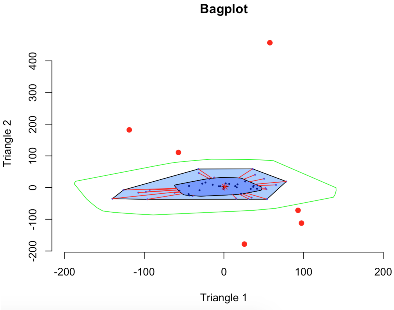



Figure: Bagplot with fence drawn in green

# Adjusted Outlyingness

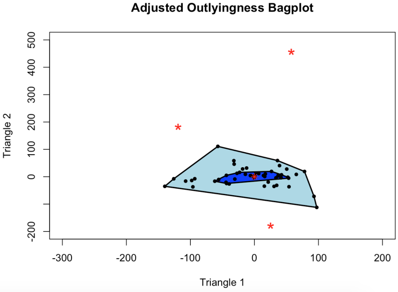



Figure: Adjusted Outlyingness bagplot without fence

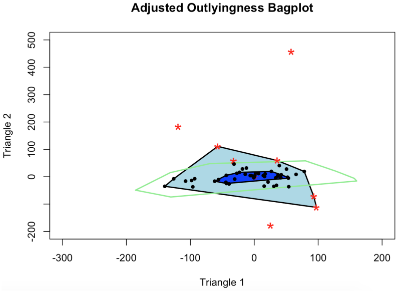



Figure: Adjusted Outlyingness bagplot with fence drawn in green

# Adjusted Outlyingness using alternative mrfDepth cut-off value

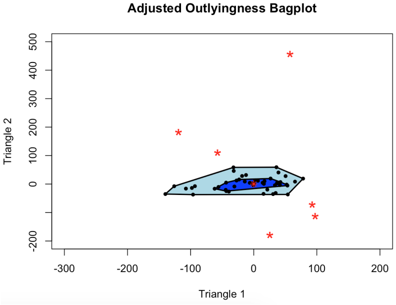



Figure: Adjusted-Outlyingness bagplot using mrfDepth cut-off

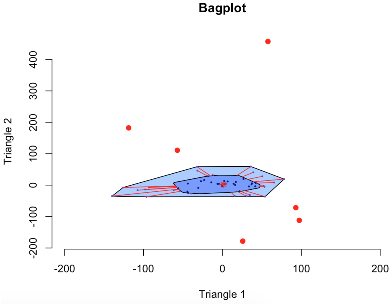



Figure: Bagplot

# Outlier Detection Results

| Outliers | Bagplot | MCD | AO* | *bagdistance*** |
|---|---|---|---|---|
| $X_{6,5}$ | ✓(2) | ✓(13.1420) | X (3.8184) | ✓(3.6055) |
| $X_{7,3}$ | ✓(3) | ✓(20.0908) | X (3.7012) | ✓(3.9901) |
| $X_{7,4}$ | ✓(1) | ✓(25.8186) | X (5.5598) | ✓(4.9688) |
| $X_{8,2}$ | ✓(1) | ✓(58.1493) | ✓(6.2001) | ✓(6.7914) |
| $X_{8,3}$ | ✓(1) | ✓(48.6677) | ✓(7.9516) | ✓(7.0146) |
| $X_{9,1}$ | ✓(1) | ✓(289.9173) | ✓(15.6375) | ✓(15.7328) |
| $X_{9,2}$ | X (1) | ✓(7.6540) | X (1.6196) | X (2.4845) |

Table: Outlier Detection Results

*All were detected only when using traditional AO cut-off value.

**Using cut-off distance of $\sqrt{\chi^2_{0.99,2}}$.

Actuaries Institute

# Weighting adjustment of outliers based on bagdistance

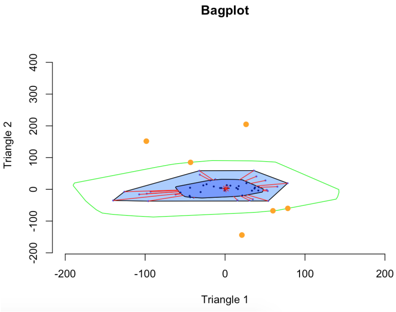



Figure: bagdistance adjustment mechanism with no limit

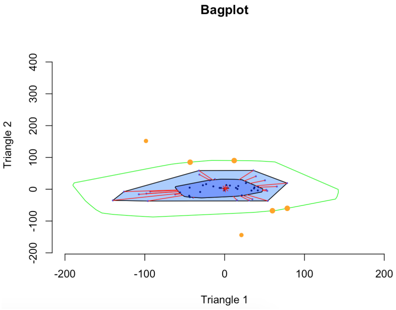



Figure: bagdistance adjustment mechanism with limit

# Final Reserves and rmse

| | Triangle 1 | | Triangle 2 | | Total | |
|---|---|---|---|---|---|---|
| | Reserve | rmse | Reserve | rmse | Reserve | rmse |
| Original | 6 435 951 | 322 573 | 489 028 | 90 542 | **6 924 978** | **337 001** |
| MCD | 6 438 541 | 283 054 | 438 497 | 45 552 | **6 877 037** | **293 870** |
| Bagplot-Fence | 6 427 881 | 291 820 | 441 746 | 51 132 | **6 869 627** | **302 233** |
| Bagplot-Loop | 6 416 205 | 283 260 | 445 063 | 41 656 | **6 861 268** | **296 087** |
| AO-Fence | 6 883 147 | 300 431 | 431 610 | 50 330 | **6 883 147** | **300 431** |
| AO-Loop | 6 438 760 | 299 719 | 434 768 | 56 658 | **6 873 528** | **304 282** |
| AO-mrfDepth | 6 413 096 | 283 024 | 444 902 | 41 590 | **6 857 998** | **295 890** |
| *bd* (no limit) | 6 416 900 | 304 215 | 456 717 | 68 371 | **6 873 617** | **315 127** |
| *bd* (limit) | 6 412 356 | 301 108 | 438 695 | 61 462 | **6 851 051** | **305 999** |

Table: Bivariate Example Reserves and rmse

- Modest adjustment in reserves
- Significant reduction in rmse of reserves

# Robust N-Dimensional Chain-Ladder

Extension of Robust Bivariate Chain-Ladder including techniques based on

- Adjusted Outlyingness
- Halfspace Depth

# Adjusted Outlyingness Approach

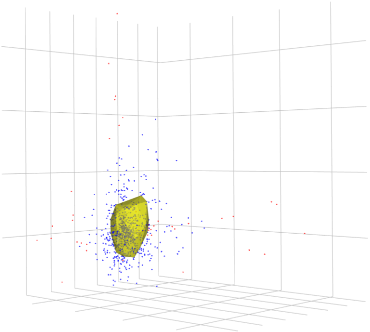


Figure: Trivariate Residuals with 3D AO-based Bag

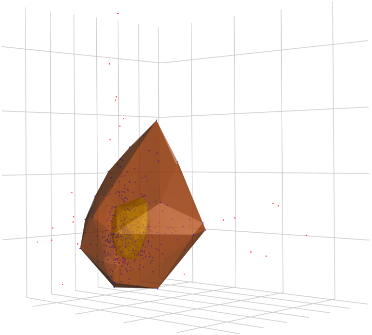
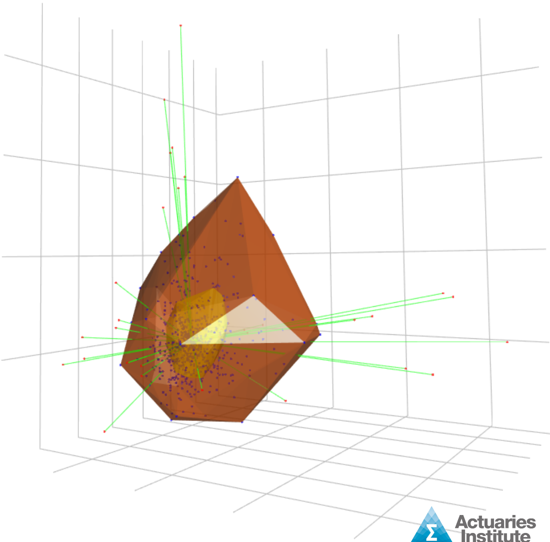


Figure: Trivariate Residuals with 3D AO-based Bag and Loop

Adjust residuals back to relevant convex polyhedron using parametric line-clipping algorithm (Cyrus and Beck, 1978; Liang and Barsky, 1984)



# Some remarks

- We put forward extensions to robust chain ladder techniques
- We do not recommend that these methodologies should necessarily *replace* existing methodologies.
- Rather, we recommend this analysis is done *in addition* to existing analysis as a **diagnostic tool** so as to gain further **insights about possible outliers**
- This should then inform how to proceed with existing techniques, and may prompt the actuary to make some additional investigations about the data

# Summary

In this presentation we

- Reviewed outlier detection techniques, and the (scant) literature in robust reserving
- Implemented and explored recent developments in cut-off values for the bagplot and Adjusted-Outlyingness techniques
- Provided novel alternative adjustment mechanisms based on bagdistance and Adjusted Outliness
- Explained how to extend the robust bivariate chain-ladder to an N-dimensional framework

# Conclusions

- Claims data are often skewed and contain outliers
- Using our techniques, such features can be more reliably identified in multivariate settings
- This should inform the reserving process, and ultimately lead to more reliable reserves

# References I

Avanzi, B., Lavender, M., Taylor, G., Wong, B., 2016. On the impact, detection and treatment of outliers in robust loss reserving. In: General Insurance Seminar, Melbourne, Australia.

Bornhuetter, R., Ferguson, R., 1972. The actuary and IBNR. Proceedings of The Casualty Actuarial Society 59, 181–195.

Cyrus, M., Beck, J., 1978. Generalized two- and three-dimensional clipping. Computers and Graphics 3, 23–28.

Donoho, D., 1982. Breakdown properties of multivariate location estimators. Qualifying Paper, Harvard University, Boston.

Donoho, D. L., Gasko, M., Dec 1992. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. The Annals of Statistics 20 (4), 1803–1827.

England, P., Verrall, R., 2002. Stochastic claims reserving in general insurance. British Actuarial Journal 8(3), 443–544.

# References II

Hachmeister, C. A., Stanard, J. N., 1975. Ibnr claims count estimation with static lag functions. Spring meeting of the Casualty Actuarial Society.

Hampel, F., 1968. Contributions to the theory of robust estimation, ph.d. thesis,. University of California, Berkeley.

Huber, P., Ronchetti, E., 2009. Robust Statistics, 2nd Edition. John Wiley & Sons Inc. Publications.

Hubert, M., Rousseeuw, P., Segaert, P., 2015. Multivariate functional outlier detection. Statistical Methods & Applications 24, 177–202.

Hubert, M., Rousseeuw, P., Segaert, P., 2016. Multivariate and functional classification using depth and distance. Advances in Data Analysis and Classification, 1–22.

Hubert, M., Van der Veeken, S., 2008. Outlier detection for skewed data. Journal of Chemometrics 22, 235–246.

Hubert, M., Vandervieren, E., 2008. An adjusted boxplot for skewed distributions. Computational Statistics and Data Analysis 52, 5186–5201.

# References III

Kremer, E., 1997. Robust lagfactors. ASTIN Colloquium 1997 at Cairns, 137–145.

Künsch, H., Stefanski, L., Carroll, R., 1989. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. Journal of the American Statistical Association 84(406), 460–466.

Liang, Y., Barsky, B., 1984. A new concept and method for line clipping. ACM Transaction on Graphics 3, 1–22.

Mack, T., 1993. Distribution-free calculation of the standard error of chain-ladder reserve estimates. ASTIN Bulletin 23(2), 213–225.

Pregibon, D., 1982. Resistant fits for some commonly used logistic models with medical applications. Biometrics 38(2), 485–498.

Romanazzi, M., 2001. Influence function of halfspace depth. Journal of Multivariate Analysis 77, 138–161.

# References IV

Rousseeuw, P., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.

Rousseeuw, P., Ruts, I., Tukey, J., 1999. The bagplot: A bivariate boxplot. The American Statistician 54(4), 282–287.

Segaert, P., Hubert, M., Rousseeuw, P., Raymaekers, J., Vakili, K., 2017. Package 'mrfdepth'.
URL
https://cran.r-project.org/web/packages/mrfDepth/mrfDepth.pdf

Shi, P., Basu, S., Meyers, G., 2012. A bayesian log-normal model for multivariate loss reserving. North American Actuarial Journal 16(1), 29–51.

Stefanski, A., Carroll, R., Ruppert, D., 1986. Optimally bounded score functions for generalized linear models with applications to logistic regression. Biometrika 73(2), 413–424.

Taylor, G., 1988. Regression models in claims analysis i: theory. Proceeding of the Casualty Actuarial Society 74, 354–383.

Actuaries Institute

# References V

Taylor, G., Ashe, F., 1983. Second moments of estimates of outstanding claims. Journal of Econometrics 23(1), 37–61.

Tukey, 1977. Exploratory Data Analysis, 1st Edition. Reading Mass, Addison-Wesley Publishing Co.

Tukey, J., 1975. Mathematics and the picturing of data*. Proceedings of the International Congress of Mathematicians, 523–531.

Van Wouwe, M., Verdonck, T., Van Rompay, K., 2009. Application of classical and robust chain-ladder methods: results for the belgian non-life business. Global Business and Economics Review 11, 99–115.

Venter, G., Tampubolon, R., 2008. Robustifying reserving. Variance 4(2), 136–154.

Verdonck, T., Debruyne, M., 2011. The influence of individual claims on the chain-ladder estimates: Analysis and diagnostic tool. Insurance: Mathematics and Economics 48, 85–98.

# References VI

Verdonck, T., Van Wouwe, M., 2011. Detection and correction of outliers in the bivariate chain-ladder method. Insurance: Mathematics and Economics 49, 188–193.

Verdonck, T., Van Wouwe, M., Dhaene, J., 2009. A robustification of the chain-ladder method. North American Actuarial Journal 13, 280–298.

Wright, T. S., 1990. A stochastic method for claims reserving in general insurance. Journal of the Institute of Actuaries 117(3), 677–731.

Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association 93, 120–131.