Outlier Detection in Rating-Scale Data via Autoencoders

Max Welz[1,2] Andreas Alfons[1]

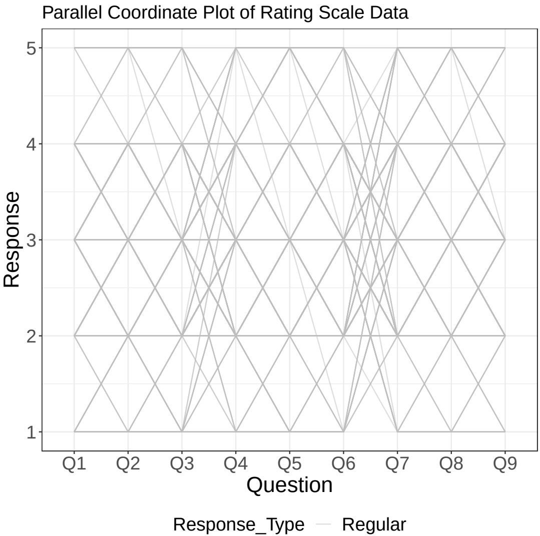
[1]Erasmus University Rotterdam, Dept. of Econometrics

[2]Erasmus Medical Center, Dept. of Public Health

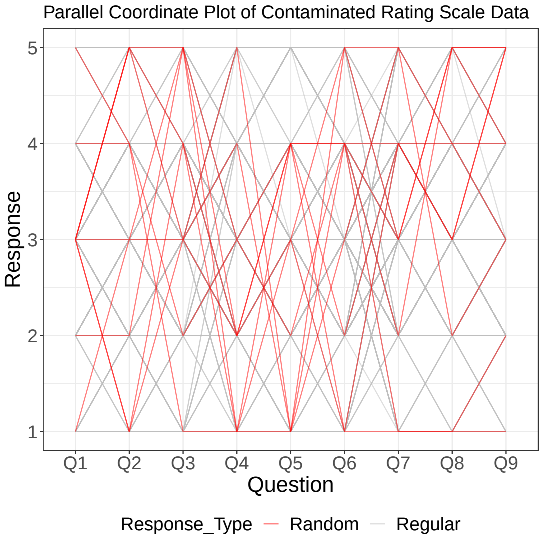
September 23, 2021

ICORS 2021, Vienna, Austria

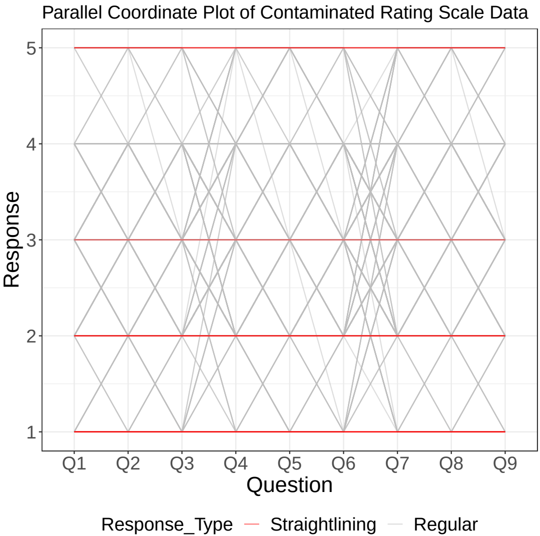
Parallel Coordinate Plot of Rating-Scale Dataset



Parallel Coordinate Plot of Rating-Scale Dataset



Parallel Coordinate Plot of Rating-Scale Dataset



Types of Outliers in Rating-Scale Data

The psychological literature defines various types of rating-scale outliers. We focus on *content nonresponsivity* (Nichols et al., 1989):

1. perfect straightlining;
2. imperfect straightlining;
3. perfect extreme responding;
4. imperfect extreme responding;
5. random responding.

Type 1.1: Perfect Straightlining

Perfect straightlining is the tendency to consistently choose the same answer category, regardless of question content.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
ICORS 2021 is awesome.	☐	☐	☐	☐	■
I like chocolate.	☐	☐	☐	☐	■
Oxygen is important.	☐	☐	☐	☐	■
I like my job.	☐	☐	☐	☐	■
I dislike my job.	☐	☐	☐	☐	■
Getting bitten by a shark would be fun.	☐	☐	☐	☐	■

Figure: Perfect straightlining with "Strongly Agree" as focal response.

Type 1.2: Imperfect Straightlining

Imperfect straightlining is the tendency to consistently choose responses around the same focal answer category, regardless of question content.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
ICORS 2021 is awesome.	□	□	□	■	□
I like chocolate.	□	□	□	□	■
Oxygen is important.	□	□	■	□	□
I like my job.	□	□	□	■	□
I dislike my job.	□	□	□	■	□
Getting bitten by a shark would be fun.	□	□	□	■	□

Figure: Imperfect straightlining with "Agree" as focal response.

Type 2.1: Perfect Extreme Responding

Perfect extreme responding is the tendency to choose solely the most extreme answer categories, regardless of question content.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
ICORS 2021 is awesome.	□	□	□	□	■
I like chocolate.	■	□	□	□	□
Oxygen is important.	■	□	□	□	□
I like my job.	□	□	□	□	■
I dislike my job.	□	□	□	□	■
Getting bitten by a shark would be fun.	■	□	□	□	□

Figure: Perfect extreme responding.

Type 2.2: Imperfect Extreme Responding

Imperfect extreme responding is the tendency to choose extreme answer categories (albeit not necessarily the most extreme category), regardless of question content.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
ICORS 2021 is awesome.	□	□	□	□	■
I like chocolate.	□	■	□	□	□
Oxygen is important.	■	□	□	□	□
I like my job.	□	□	□	□	■
I dislike my job.	□	□	□	■	□
Getting bitten by a shark would be fun.	■	□	□	□	□

Figure: Imperfect extreme responding.

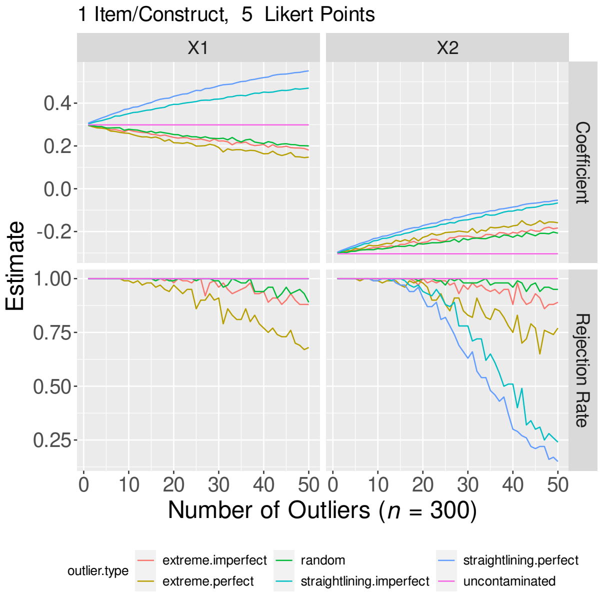
Type 3: Random Responding

Random responding is the tendency to randomly choose answer categories, regardless of question content.

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
ICORS 2021 is awesome.	□	□	□	■	□
I like chocolate.	□	□	□	□	■
Oxygen is important.	□	□	■	□	□
I like my job.	□	□	■	□	□
I dislike my job.	□	■	□	□	□
Getting bitten by a shark would be fun.	■	□	□	□	□

Figure: Random responding.

Effects of Outliers in Rating-Scale Data



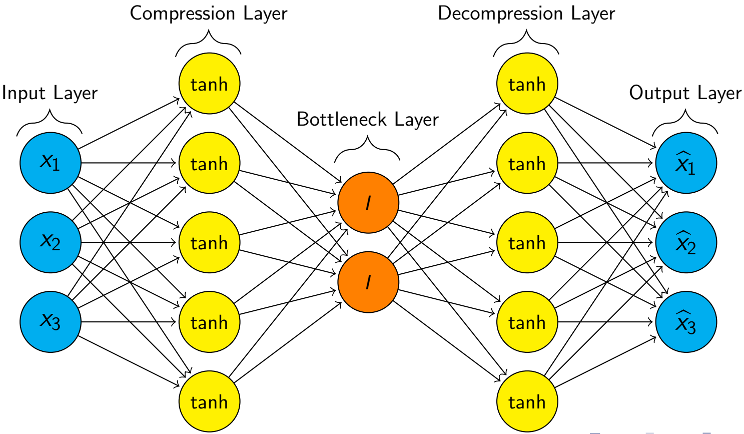
Summary of Proposed Method

For outlier detection in rating-scale data, we propose to:

- Use a form of auto-associative neural networks (*autoencoders*; Kramer, 1992);
- Transform the resulting outlier scores to central normality (Raymaekers and Rousseeuw, 2021).

Proposed Method (1/3)

- An autoencoder (Kramer, 1992) is a neural network with three hidden layers that attempts to reconstruct its input.
- Central hidden layer is crucial: compresses the input.
- Can be seen as nonlinear generalization of PCA (Kramer, 1991).



Proposed Method (2/3)

Use the per-individual mean squared reconstruction error as outlyingness score (OS) for each individual:

$$\mathrm{OS}(\mathbf{x}) = \frac{1}{p} \sum_{j=1}^{p} \left(\frac{x_j - \widehat{x}_j}{L_j} \right)^2,$$

where p is the number of questions, $\widehat{x}_j$ is the autoencoder's reconstruction of question response x_j, and L_j is the number of answer categories of question j.

Proposed Method (3/3)

We use the transformation by Raymaekers and Rousseeuw (2021).

- Let $\mathring{g}(\cdot)$ be the rectified Box-Cox transformation.[1]
- For scores $\mathrm{OS}_1, \dots, \mathrm{OS}_n$, robustly standardize the transformations by

$$z_i = \frac{\mathring{g}(\mathrm{OS}_i) - \mathrm{median}\{\mathring{g}(\mathrm{OS}_j) : j = 1, \dots, n\}}{\mathrm{MADN}\{\mathring{g}(\mathrm{OS}_j) : j = 1, \dots, n\}}.$$

- Flag observation i as *outlier* if

$$z_i > \sqrt{\chi^2_{0.975,1}}.$$

- Only allows for flagging in the right tail. Flagging in the left tail is work in progress (more on this later).

[1] We currently experiment with various choices of the rectification constant C_ℓ to find a recommended choice. Currently $C_\ell = 0.25$.

Simulation Design (1/2)

Data generating process:

- We generate $n = 300$ correlated rating-scale observations by using the sampling scheme in Kaiser et al. (2011);
- We consider three constructs, each consists of five questions (i.e. $p = 3 \times 5 = 15$ questions);
- Questions *within* the same construct have high correlation of 0.8;
- Questions from different constructs have medium correlation of ± 0.3;
- Each question has five answer categories;
- Each dataset is contaminated with up to 50 outliers;
- We average the considered performance measures over $R = 100$ such datasets.

Simulation Design (2/2)

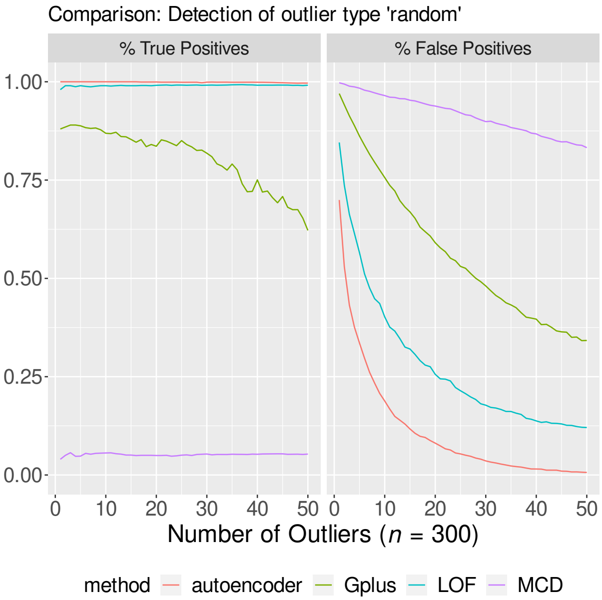
Benchmark methods:

- *Local Outlier Factor* (LOF; Breunig et al., 2000);
- G_+ score (from psychology; Zijlstra et al., 2007);
- Robust Mahalanobis distance via MCD (Rousseeuw, 1984);
- ... and seven more, as in Zijlstra et al. (2011).

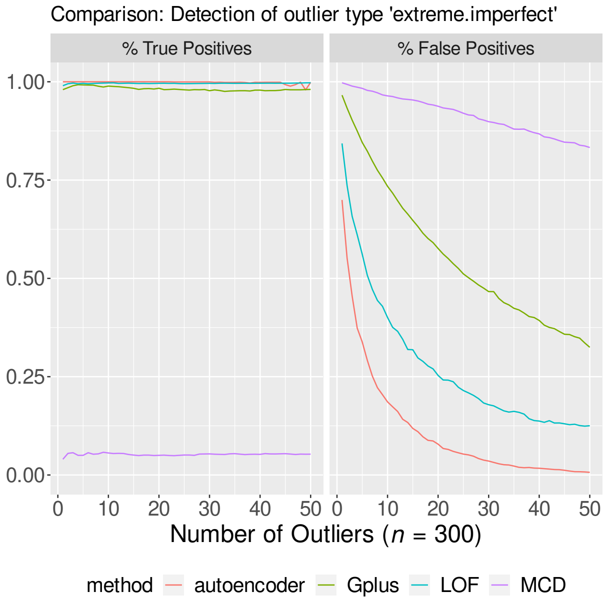
Performance measures:

- % True Positives = fraction of true positives. "How many of the true outliers are flagged?"
- % False Positives = fraction of false positives. "How many of the flagged points are no outliers?"

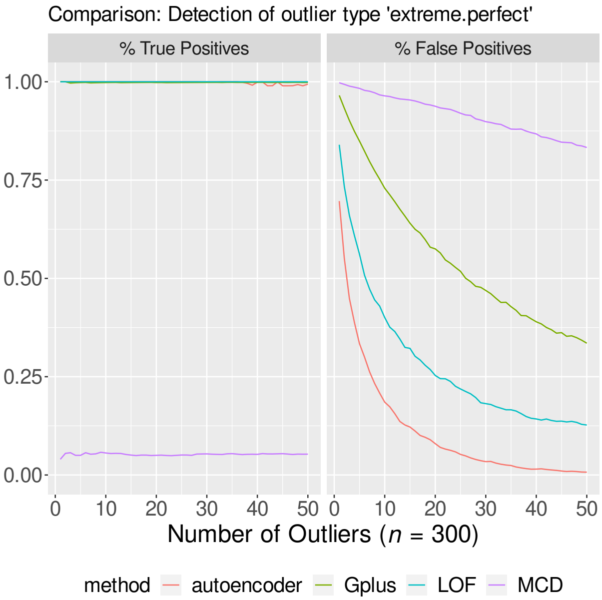
Simulation Results (1/5)



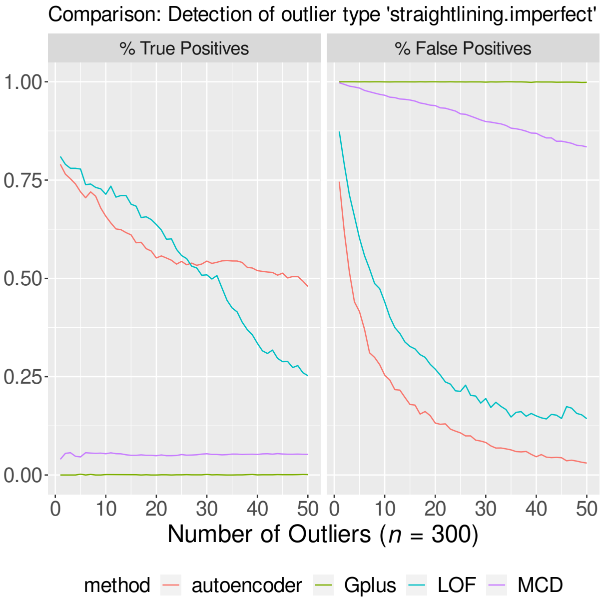
Simulation Results (2/5)



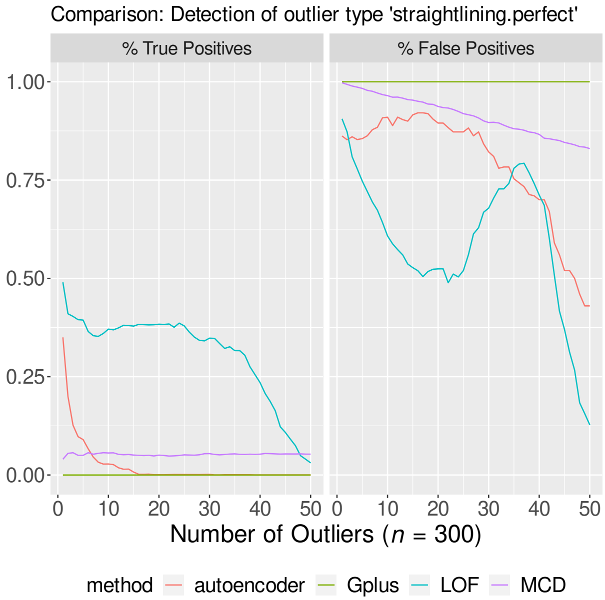
Simulation Results (3/5)



Simulation Results (4/5)



Simulation Results (5/5)



Simulation Conclusion

- Our method outperforms all benchmark methods and reliably detects all types of rating-scale outliers, except perfect straightliners;
- Perferct straighliners should be easy to detect with an additional rule:
  - ➔ Either adapt autoencoder or outlyingness score;
- Gap between our method and the benchmark methods widens in more complex scenarios.

General Conclusion

- Rating-scale outliers are different from conventional outliers, but they can be just as harmful;
- Autoencoders seem to be promising in detecting rating-scale outliers;
- Robust methods for rating-scale data/categorical data are underdeveloped. Potential for novel research ideas!

Thank you for the attention and a special thanks to the organizers of ICORS 2021! Let's have a discussion!

(Slides: `https://mwelz.github.io/assets/pdf/icors-2021.pdf`.)

References

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on management of data*, pages 93–104.

Kaiser, S., Träger, D., and Leisch, F. (2011). Generating correlated ordinal random values. Technical Report 94, Department of Statistics, University of Munich.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243.

Kramer, M. A. (1992). Autoassociative neural networks. *Computers & Chemical Engineering*, 16(4):313–328.

Nichols, D. S., Greene, R. L., and Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology*, 45(2):239–250.

Raymaekers, J. and Rousseeuw, P. J. (2021). Transforming variables to central normality. *Machine Learning*. In press.

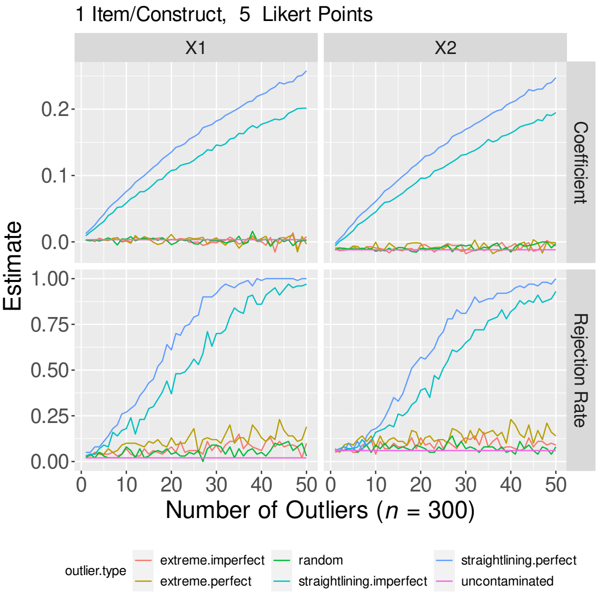
Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880.

Zijlstra, W. P., van der Ark, L. A., and Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research*, 42(3):531–555.

Zijlstra, W. P., van der Ark, L. A., and Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, 36(2):186–212.

Appendix

Effects of Outliers in Rating-Scale Data



Design Choices for the Qutoencoder

- Use robust pseudo-Huber loss for fitting:

$$x \mapsto \delta^2 \left(\sqrt{1 + (x/\delta)^2} - 1 \right),$$

  where $\delta > 0$ is a fixed constant.
- Central hidden layer: hyperbolic tangent activation (nonlinear):

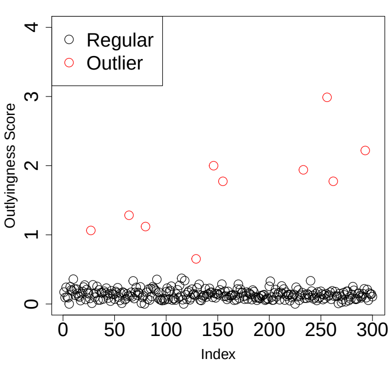
$$x \mapsto \tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}.$$

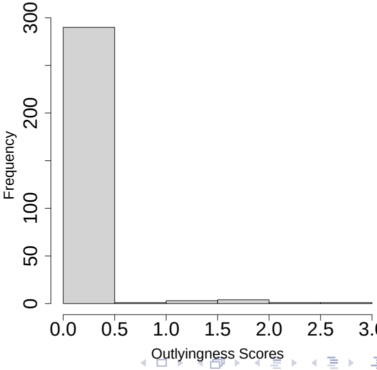
- Left and right hidden layers: linear *identity mapping*, $x \mapsto I(x) = x$; suggested by Kramer (1992).
- Use batch learning to avoid "too much" overfitting.

Outlyingness Scores of Random Respondents

We apply our autoencoder on a simulated rating-scale dataset with $n = 300$ individuals, of which 10 are random outliers. Their outlyingness scores are clearly separated.

- Same situation for all other types of outliers EXCEPT perfect straightliners.





We repeat the previous exercise, but this time, the 10 outliers are perfect straightliners. Their outlyingness scores tend to be among the very lowest.

- Not unsurprising; straightiners are easy to reconstruct;
- Outliers can be in both tails of the scores' distribution!