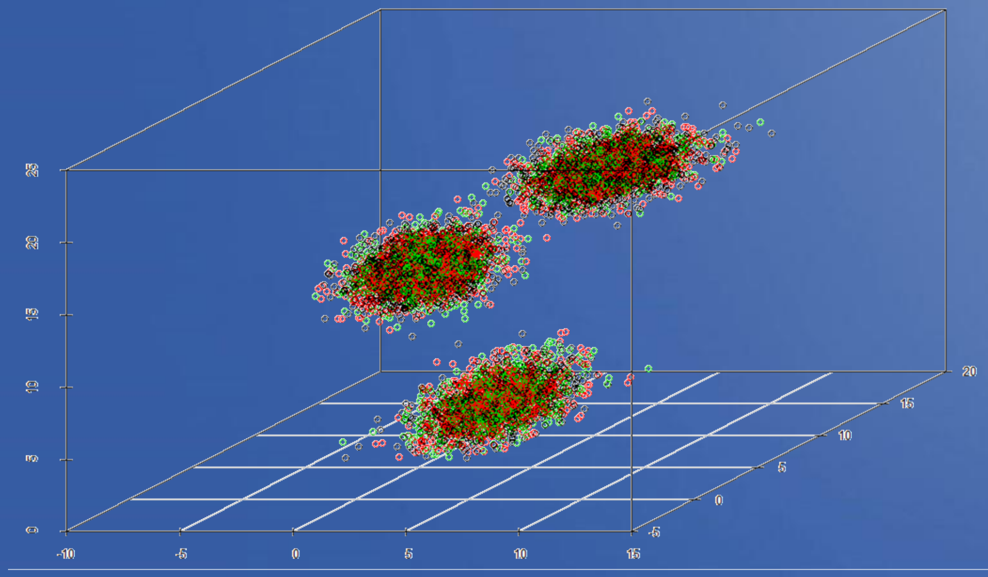
The Effect of Survey Measurement Error on Clustering Algorithms



ITSEW 2019
International Total Survey Error Workshop

Utrecht University

VU VRIJE UNIVERSITEIT AMSTERDAM

Paulina Pankowska and Dr Daniel Oberski

Presentation outline

- Background
  - ✓ Clustering algorithms: GMM and DBSCAN
  - ✓ (Survey) measurement error
  - ✓ Measurement error & clustering

- Our research
  - ✓ Testing the sensitivity of clustering results to measurement error- a simulation study

Universiteit Utrecht

VU

Clustering / cluster analysis

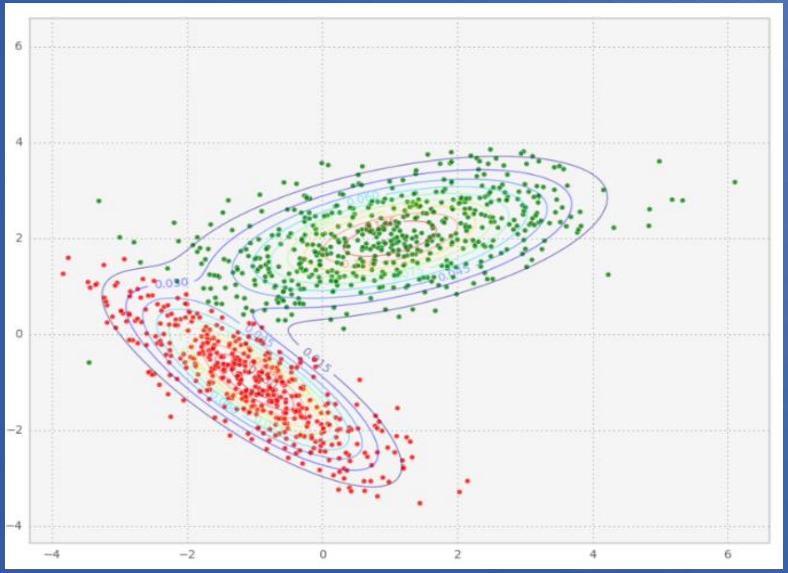
- A process classifying observations into groups (i.e. clusters) such that similar ones belong to the same group and dissimilar to different groups

- An unsupervised learning problem (unlabeled data/ no predefined classes)

- Has a wide range of applications (e.g. marketing, insurance and banking- fraud detection, adaptive survey designs)

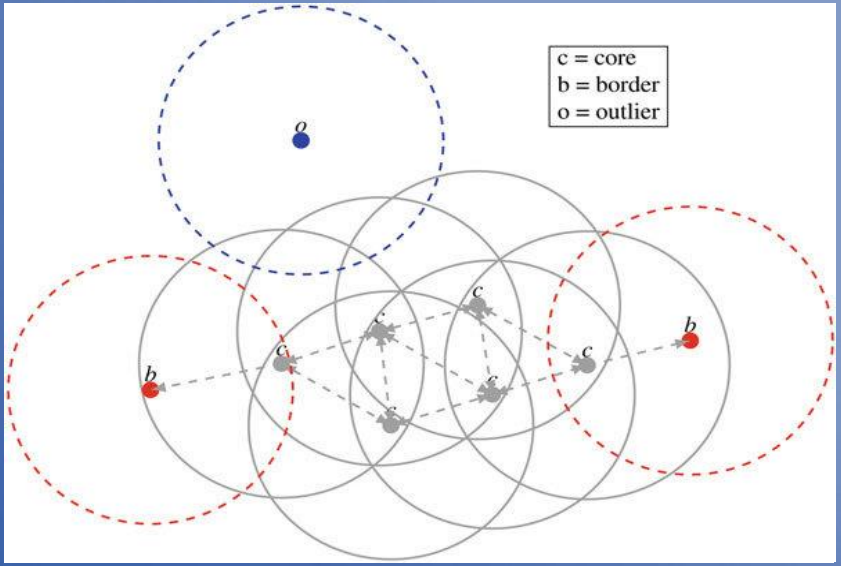
GMM and DBSCAN

Gaussian Mixture Models (GMM)



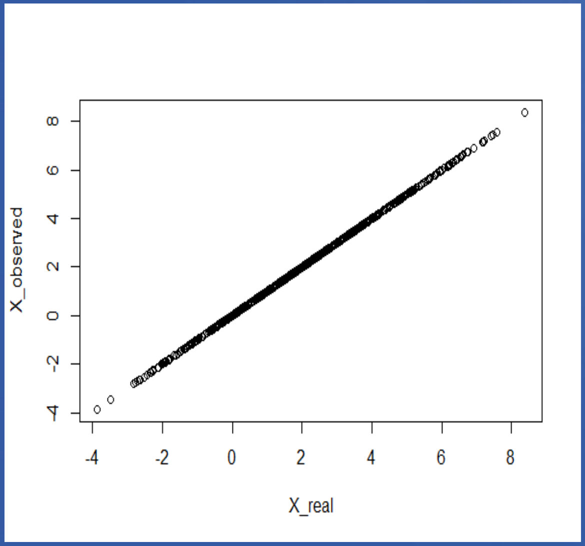
Source: http://www.nehalemlabs.net/prototype/blog/2014/04/03/quick-introduction-to-gaussian-mixture-models-with-python/

Density-based spatial clustering of applications with noise (DBSCAN)

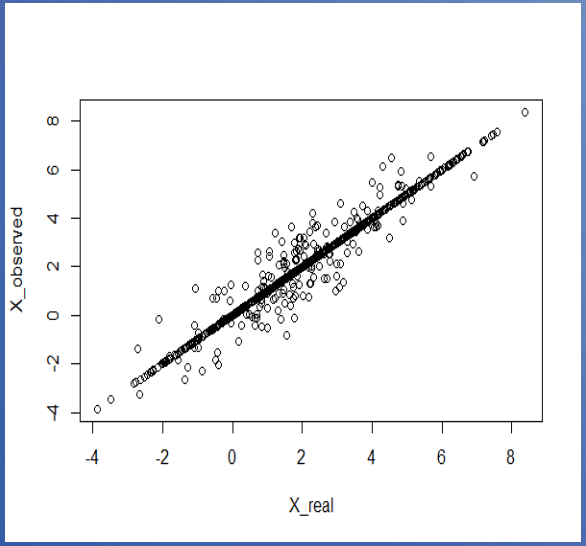


Source: Izzo et al. (2016)

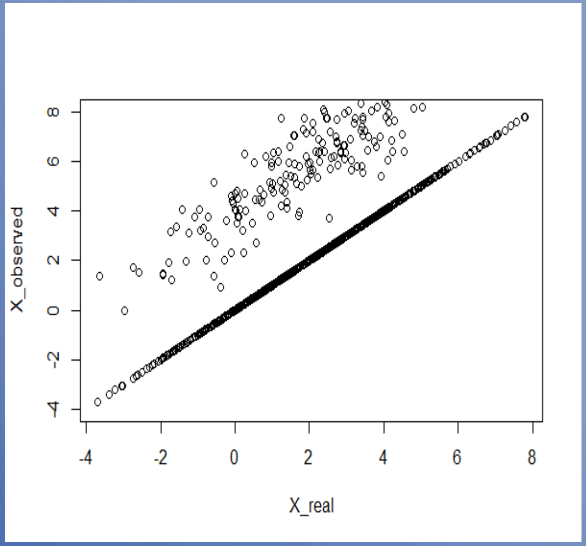
(Survey) measurement error



No measurement error



Random measurement error



Systematic measurement error

Our analysis- simulation setup

- Simulate a 3-dimensional dataset from a mixture of three multivariate Gaussian distributions (N = 1,000)

- Introduce measurement error based on 24 conditions:

Error type		Var's with error		Error rate		Magnitude
1. Random 2. Systematic	X	1. One 2. All (three)	X	1. 0.2 2. 0.4	X	1. Low 2. Medium 3. High

- Compare clustering results for 'original' data and datasets with error:

  ✓Number of clusters

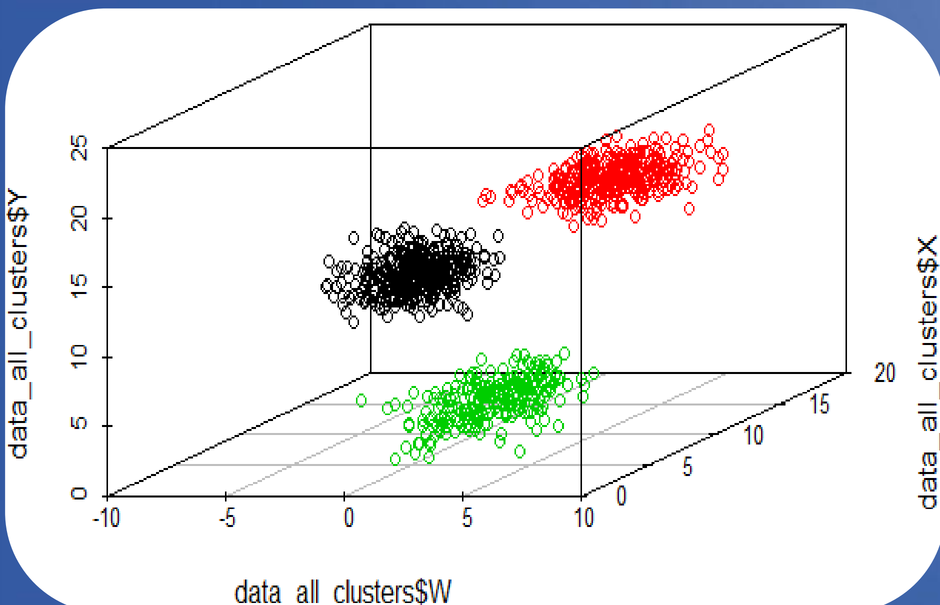
  ✓Cluster similarity of (i) 'raw' clustering results and (ii) merged/ stable clusters (adjusted Rand index)

Universiteit Utrecht

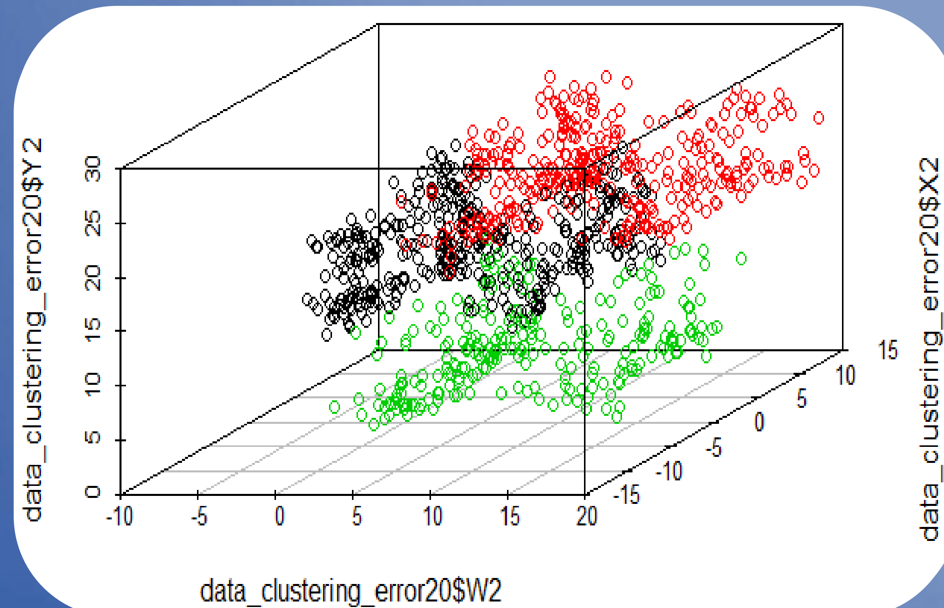
VU

Our analysis- visualization of the data

Simulated dataset ('original data')



Dataset with systematic measurement error (3 var)

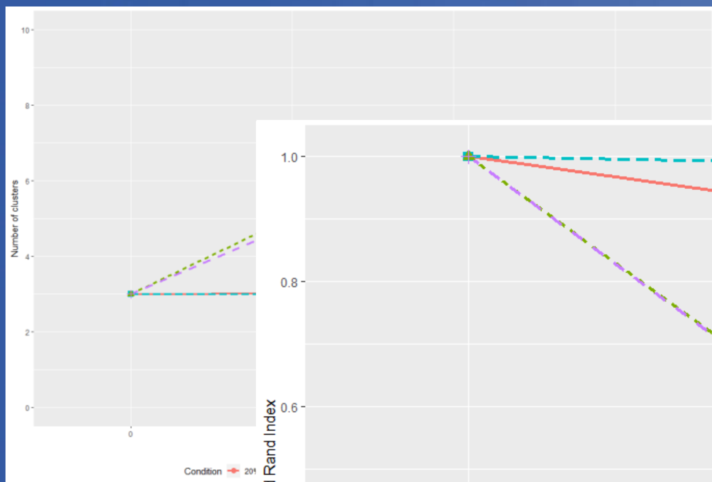


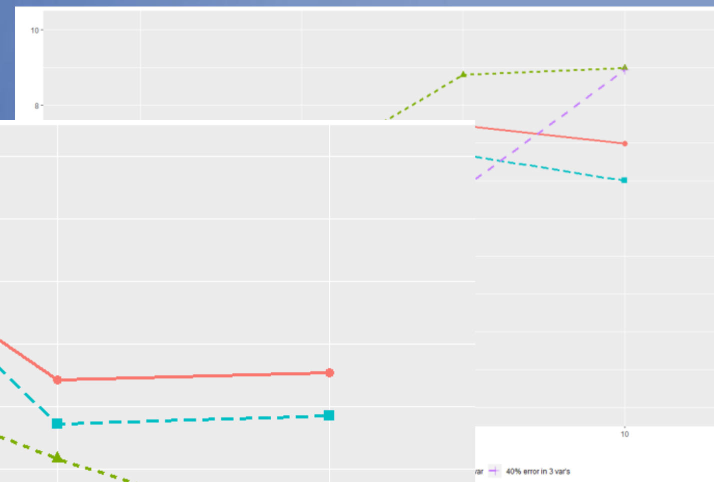
Universiteit Utrecht

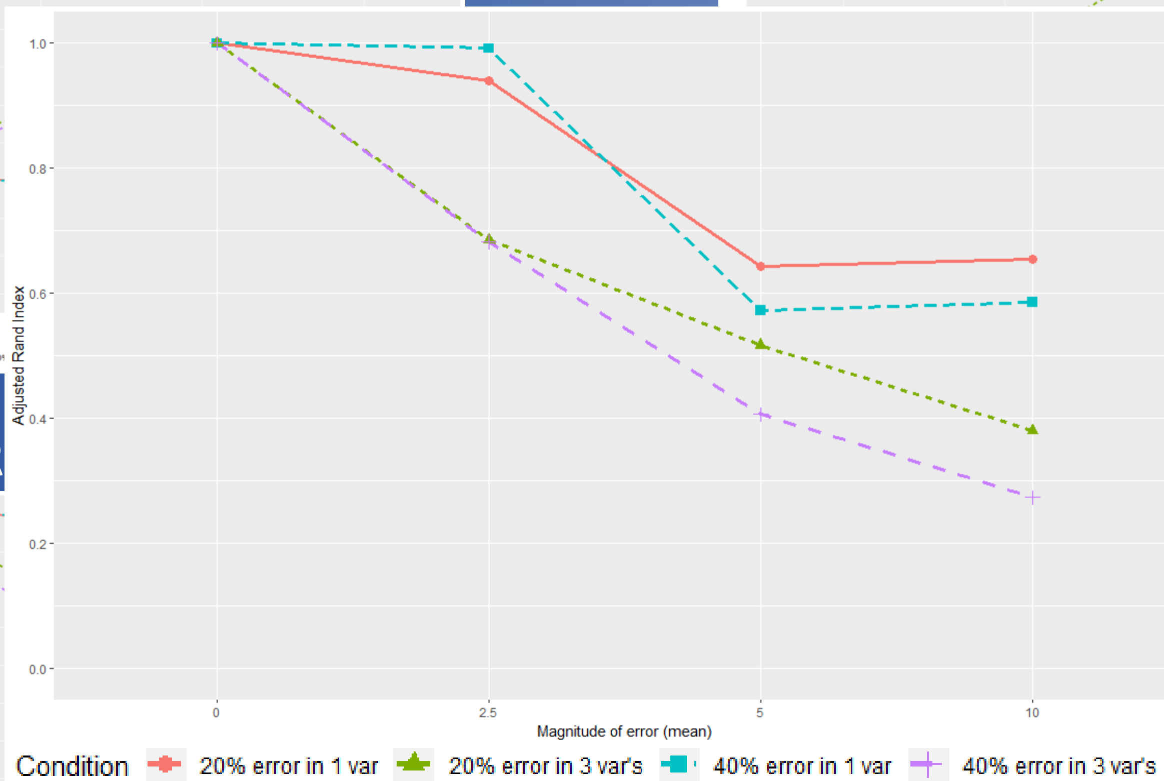
VU

GMM Results: components

- Number of clusters

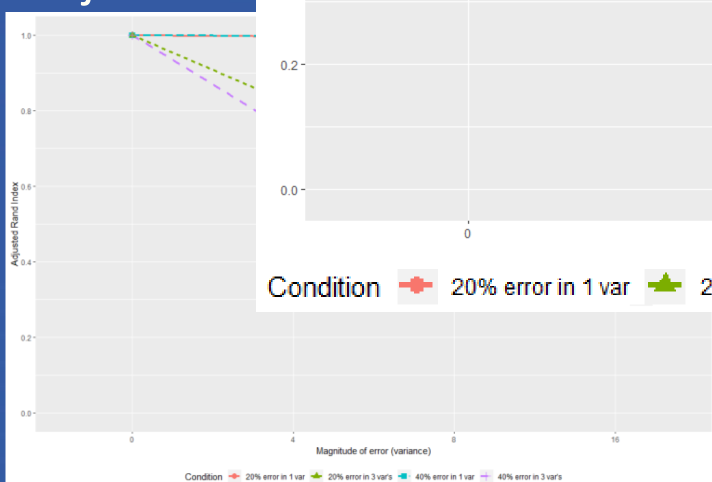


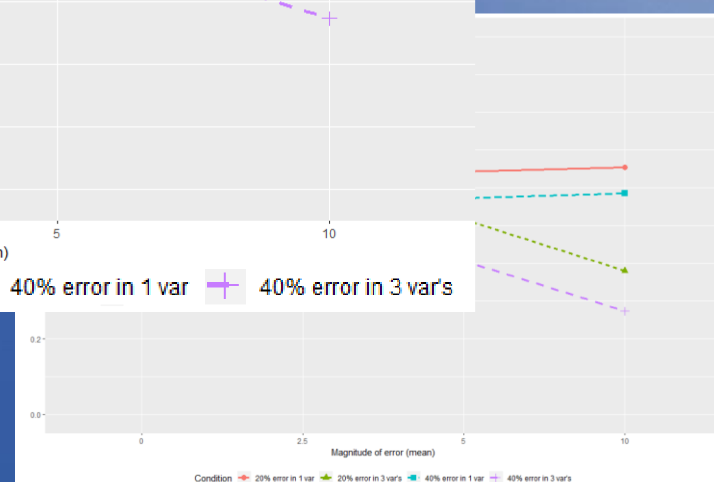




ror

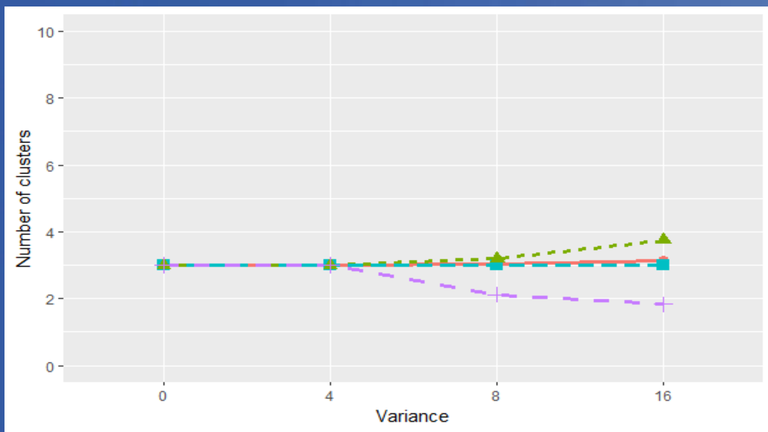
- Adjusted R



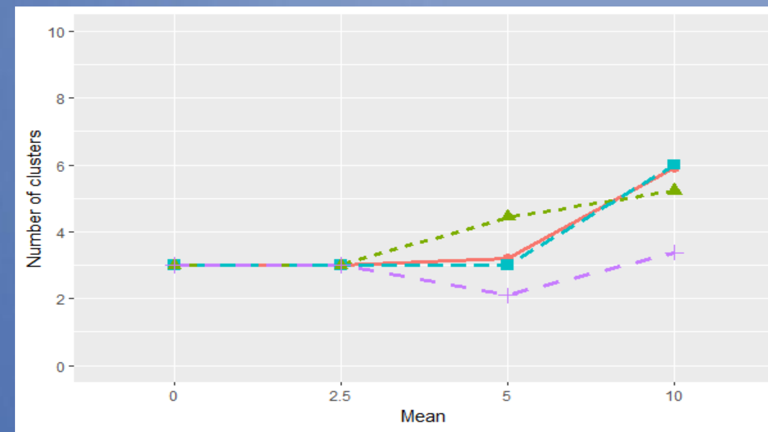


GMM Results: (merged) clusters

- Number of clusters

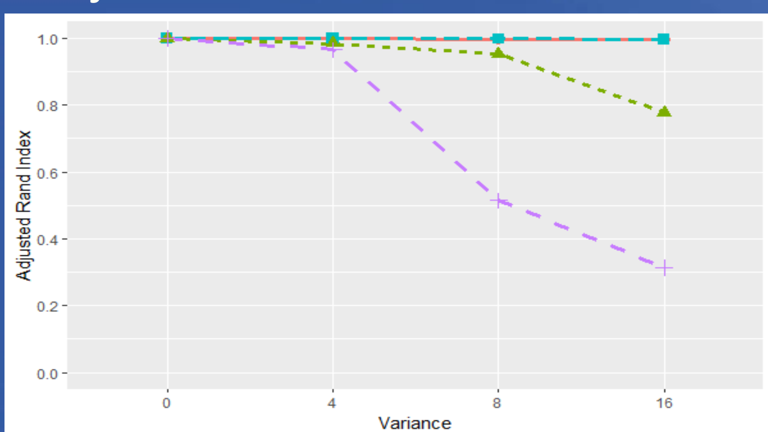


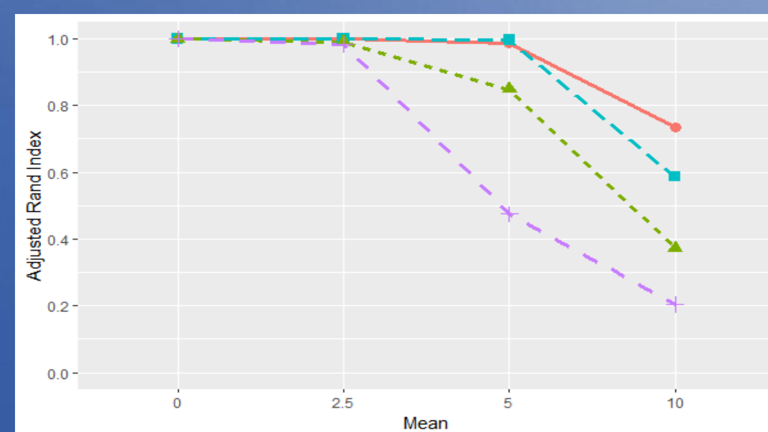
Random error



Systematic error

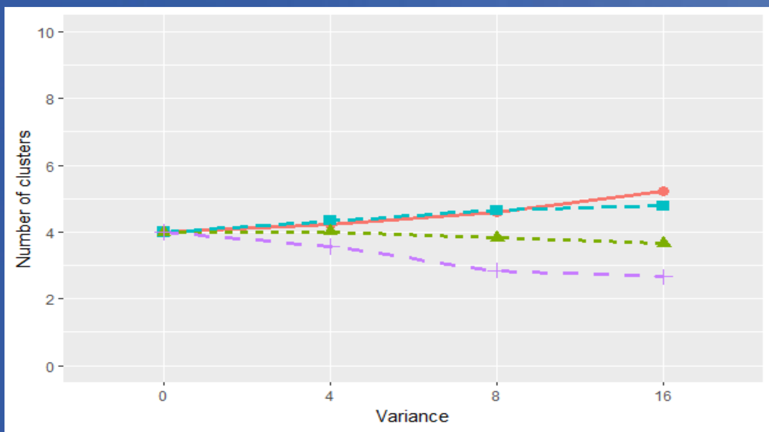
- Adjusted Rand index



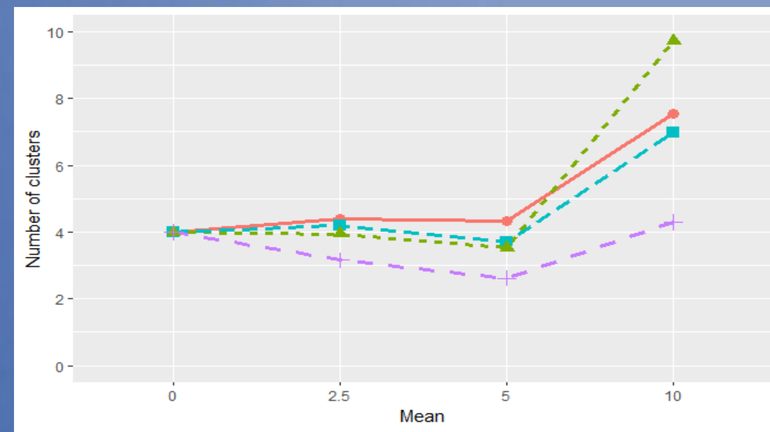


DBSCAN Results: all clusters

- Number of clusters

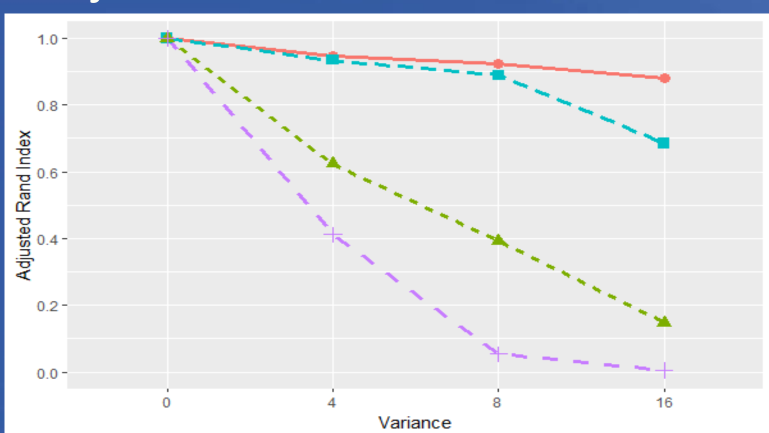


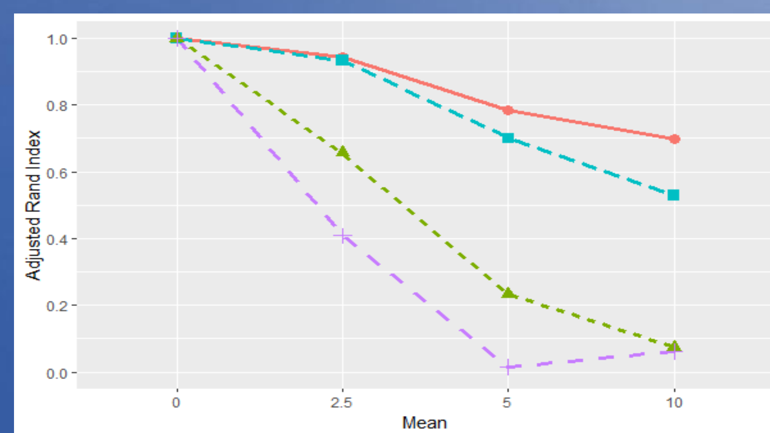
Random error



Systematic error

- Adjusted Rand index

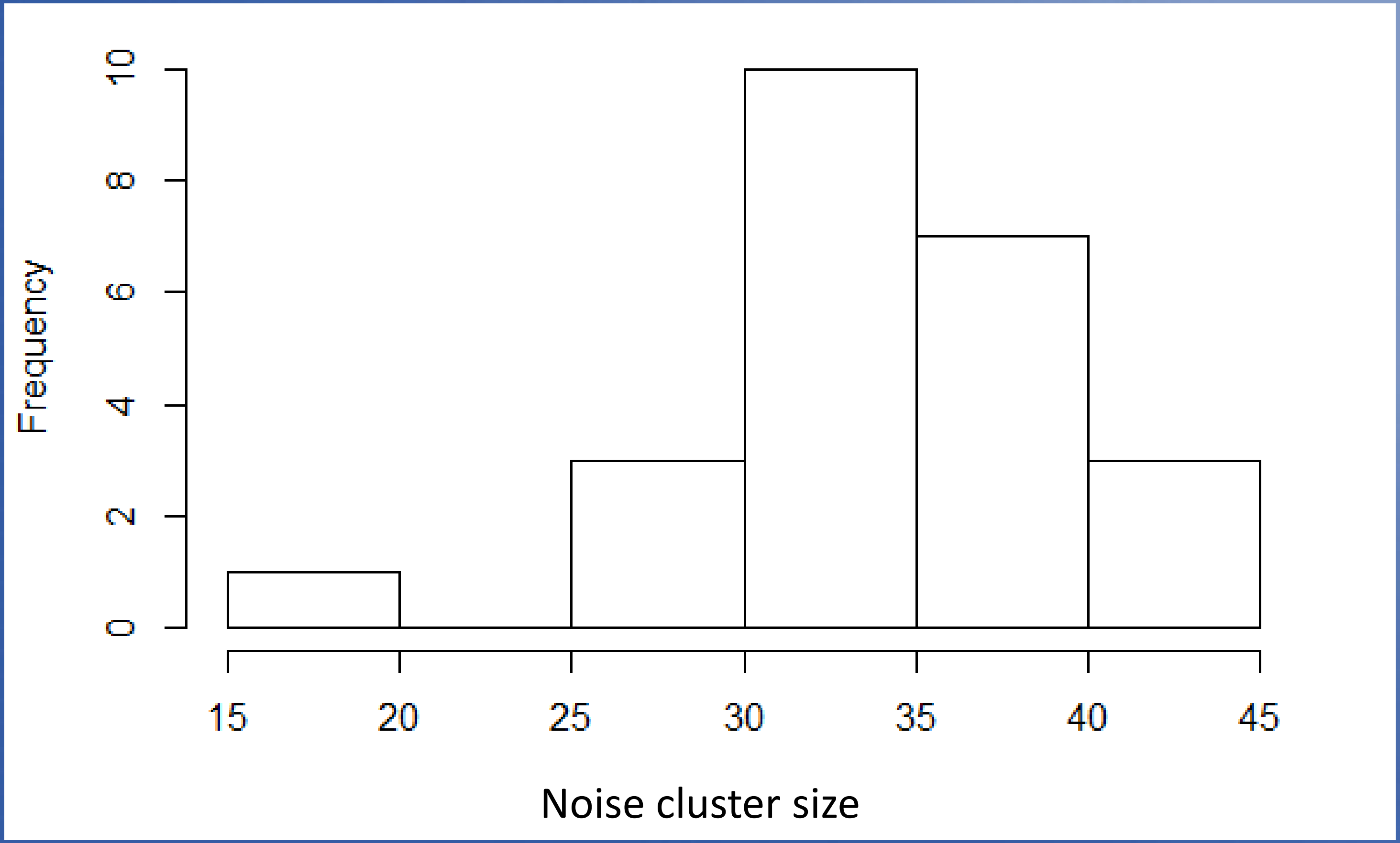




Condition: 20% error in 1 var; 20% error in 3 var's; 40% error in 1 var; 40% error in 3 var's

DBSCAN Results: size of noise cluster

- **Original noise cluster size: 32**



Summary and conclusions

- **GMM is less sensitive to measurement error than DBSCAN**
  - ✓ In particular when GMM components are merged into clusters
  - ✓ Only looking at stable DBSCAN clusters does not help
  - ✓ The noise cluster in DBSCAN does not capture measurement error
- **Measurement error has very strong biasing effects when**
  - ✓ It is systematic as opposed to random
  - ✓ It affects all (three) variables rather than only one
  - ✓ The magnitude is high
  - ✓ Error rate does not appear to matter much

Universiteit Utrecht

VU

Questions about next steps

1. Should we try other clustering algorithms and/or other techniques to extract and compare clusters?
   - If so, which ones?

2. Should we also correct for measurement error in this paper (i.e. using latent variable modelling)?

3. Would a real-life data application be interesting?

Universiteit Utrecht

Thank you!

- Paulina Pankowska, p.k.p.pankowska@vu.nl
- Daniel Oberski, d.l.oberski@uu.nl

Universiteit Utrecht

VU

References

Aggarwal, C. C. (2009). Managing and Mining Uncertain Data. Advances in Database Systems (Vol. 35). Springer

Aggarwal, C. C., & Reddy, C. K. (Eds.). (2013). Data clustering: algorithms and applications. CRC press.

Bishop, C. M. (2012). Pattern recognition and machine learning, 2006. 대한토목학회지, 60(1), 78-78.

Chaudhuri, B. B., & Bhowmik, P. R. (1998). An approach of clustering data with noisy or imprecise feature measurement. Pattern Recognition Letters, 19(14), 1307-1317.

Dave, R. N. (1991). Characterization and detection of noise in clustering. Pattern Recognition Letters, 12(11), 657-664.

Frigui, H., & Krishnapuram, R. (1996). A robust algorithm for automatic extraction of an unknown number of clusters from noisy data. Pattern Recognition Letters, 17(12), 1223-1232.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis, 52(1), 258-271.

Hennig, C. (2010). Methods for merging Gaussian mixture components. Advances in data analysis and classification, 4(1), 3-34.

Hennig, C. (2013). fpc: Flexible procedures for clustering. R package version 2.1-5.

Izzo, D., Hennes, D., Simões, L. F., & Märtens, M. (2016). Designing complex interplanetary trajectories for the global trajectory optimization competitions. In Space Engineering (pp. 151-176). Springer, Cham.

Jolion, J. M., & Rosenfeld, A. (1989). Cluster detection in background noise. Pattern Recognition, 22(5), 603-607.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons.

Kumar, M., & Patel, N. R. (2007). Clustering data with measurement errors. Computational Statistics & Data Analysis, 51(12), 6084-6101.

Kumar, M., Patel, N. R., & Woo, J. (2002, July). Clustering seasonality patterns in the presence of errors. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 557-563). ACM.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45(3), 325-342.

Nevo, D., Zucker, D. M., Tamimi, R. M., & Wang, M. (2016). Accounting for measurement error in biomarker data and misclassification of subtypes in the analysis of tumor data. Statistics in medicine, 35(30), 5686-5700.

Universiteit Utrecht

VU