Some outlierness measures for unsupervised anomaly detection

matthieu.lesnoff@cirad.fr

ChemHouse, Montpellier, 29 April 2025

X
(spectra)

x

Is x isolated/extreme?

Example of 4 outlierness measures

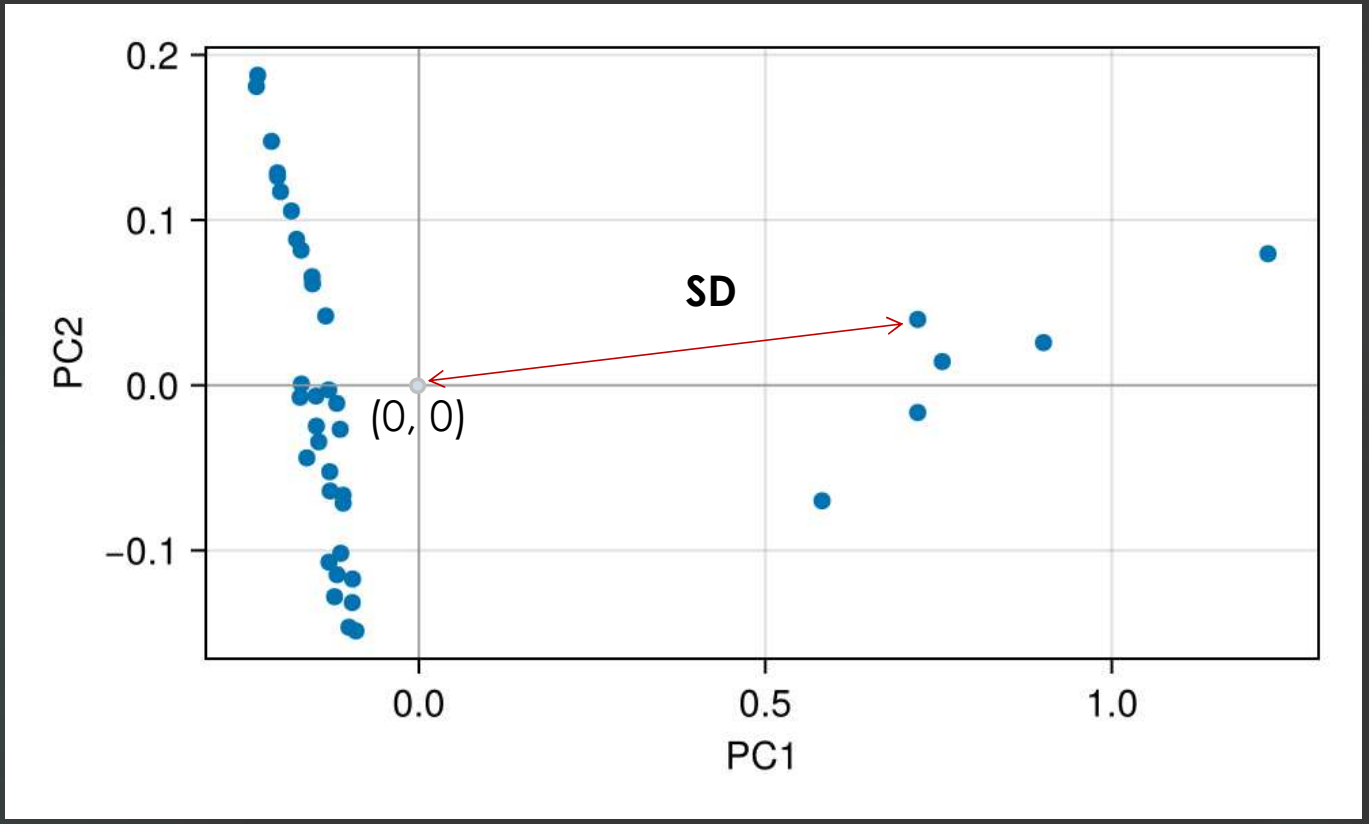
PCA

1. Score distance (SD)
2. Orthogonal distance (OD)

KNN Distance-based

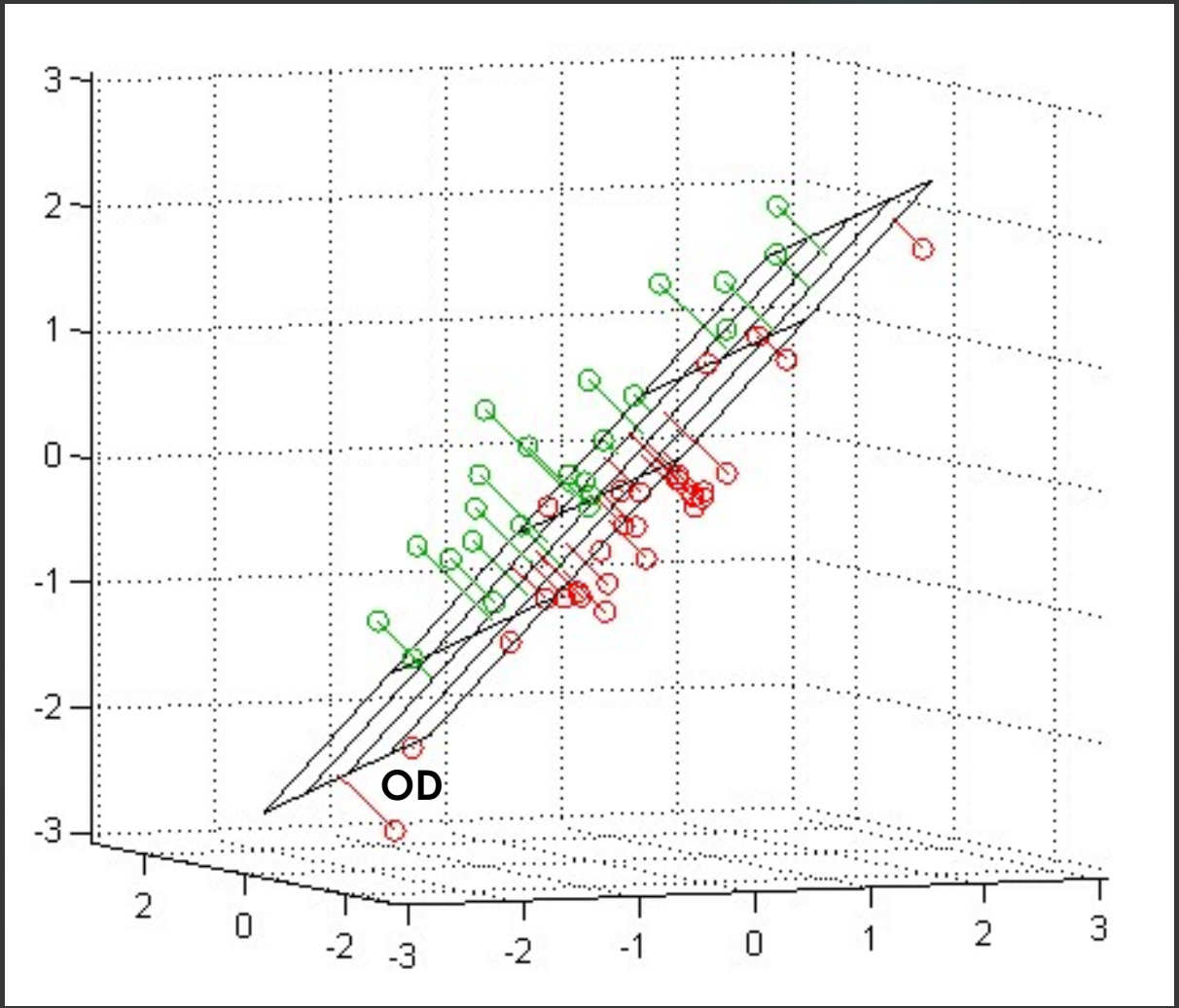
3. Global
4. Local

1. SD: Mahalanobis distance between the projection and the center of the score space



2. OD: Euclidean distance between the observation and its projection on the score space

= X-residuals



SD/OD can be summarized to a compromise

- $\sqrt{.5 \times \left(\frac{\mathrm{SD}}{\mathrm{cutoff}}\right)^2 + .5 \times \left(\frac{\mathrm{OD}}{\mathrm{cutoff}}\right)^2}$
- $\sqrt{\frac{\mathrm{SD}}{\mathrm{cutoff}} \times \frac{\mathrm{OD}}{\mathrm{cutoff}}}$

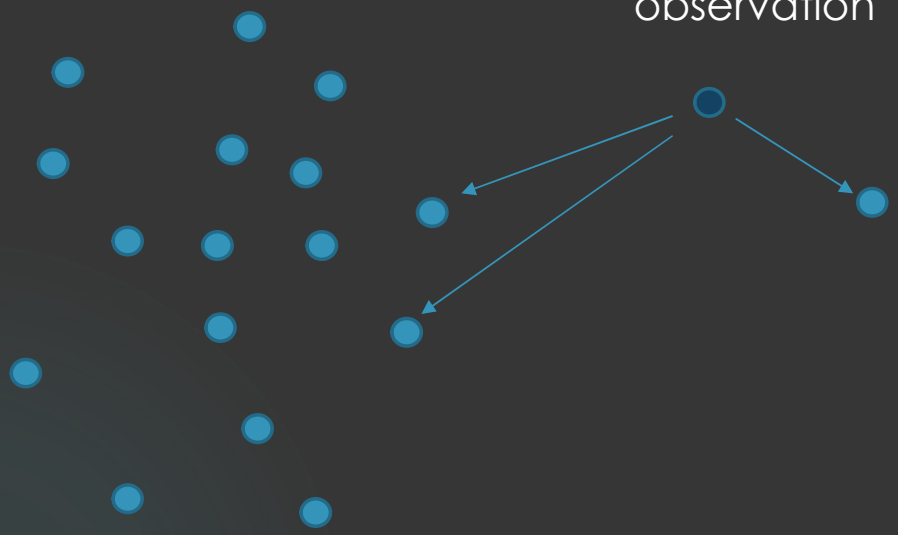
3. Global KNN Distance-based

For each observation

- Find its k nearest neighbors
- Summarize the k distances, e.g. sum or maximum

  (estimate of 1 / density)

Angiulli, F., Pizzuti, C., 2005. https://doi.org/10.1109/TKDE.2005.31
Angiulli, F. et al. 2006. https://doi.org/10.1109/TKDE.2006.2
Campos et al. 2016 https://doi.org/10.1007/s10618-015-0444-8
Ramaswamy et al. 2000. https://doi.org/10.1145/342009.335437



k = 3

High value ⇒ neighbors are far ⇒ the observation is expected to be isolated

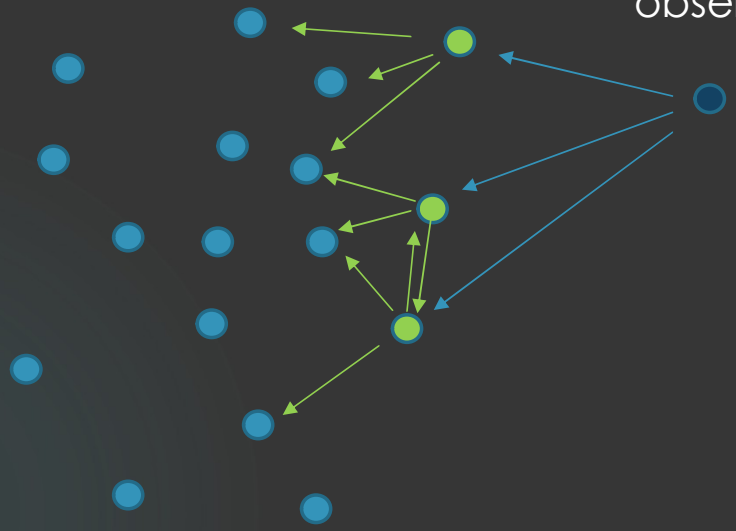
4. Local KNN Distance-based

For each observation

- Find its k nearest neighbors
- Summarize the k distances (e.g. sum or maximum) ➔ **out1**
- For each of the k neighbors
  - find the k nearest neighbors and summarize the k distances
- Average the k summary values ➔ **out2**
- Outlierness = **out1 / out2**

~ density around the neighbors / density at the observation

Simplified-LOF
Campos et al. 2016 https://doi.org/10.1007/s10618-015-0444-8
Schubert et al. 2014. https://doi.org/10.1007/s10618-012-0300-z



For each observation

k = 3

Can be computed

- In the X-space,

  or after dimension reduction (PCA, tSNE, UMAP, etc.)

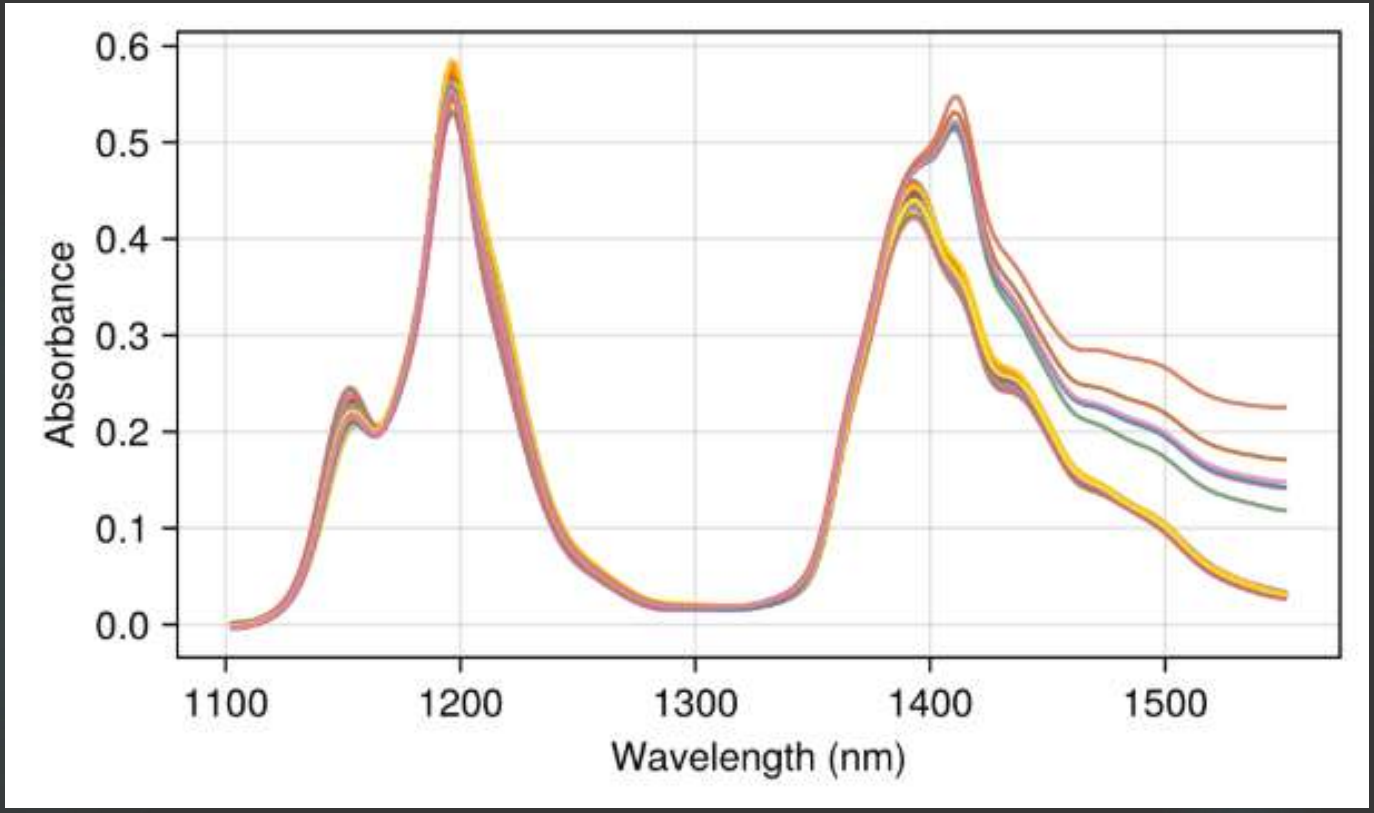
- With different metrics

Illustrations

- Octane
- Challenge2018

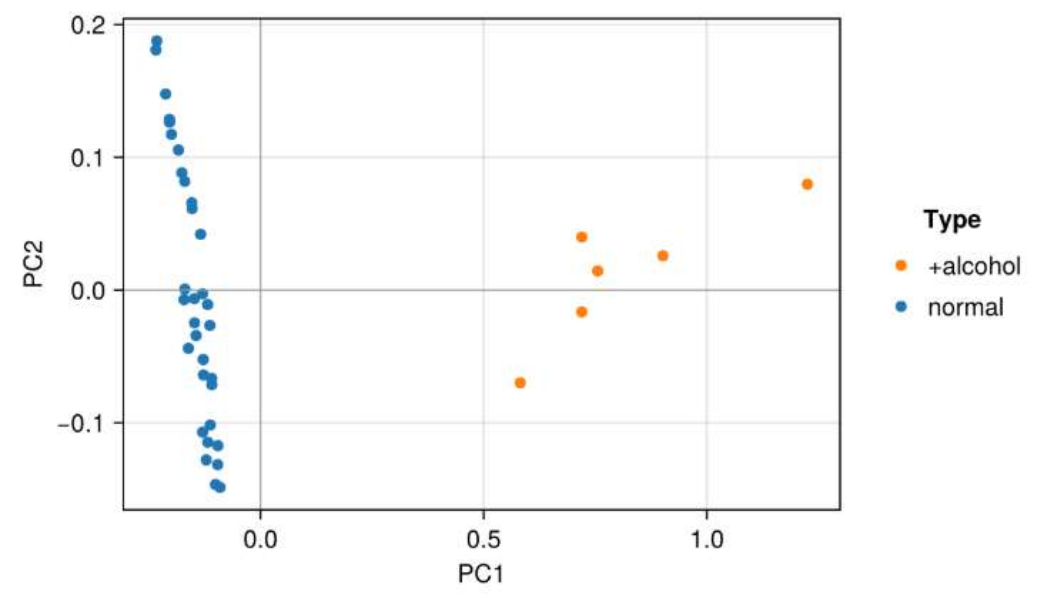
Octane dataset

n = 39 NIR spectra of gasoline samples (1102 -1552 nm step 2)
Six of the samples contain added alcohol
Hubert et al. 2005, Technometrics, 47, 64–79

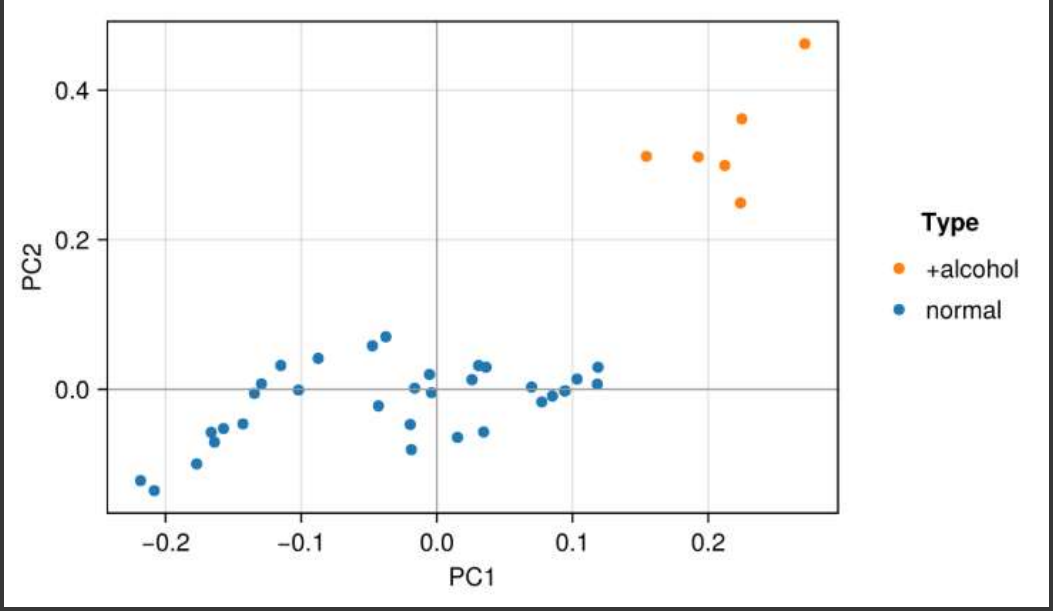


PCA nlv = 3

Usual



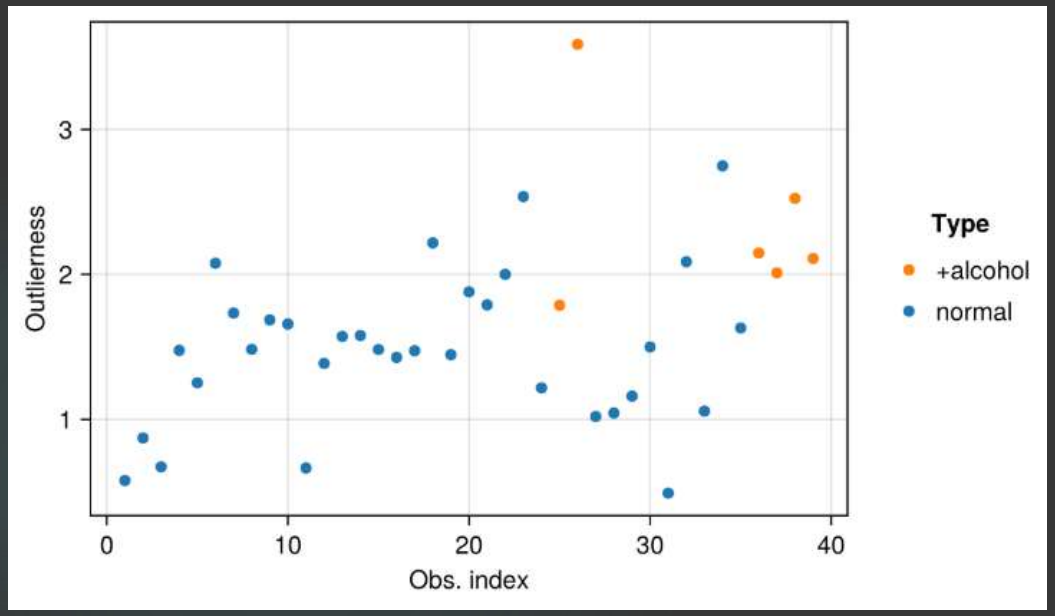
Robust



… Detection with SD is expected to be difficult

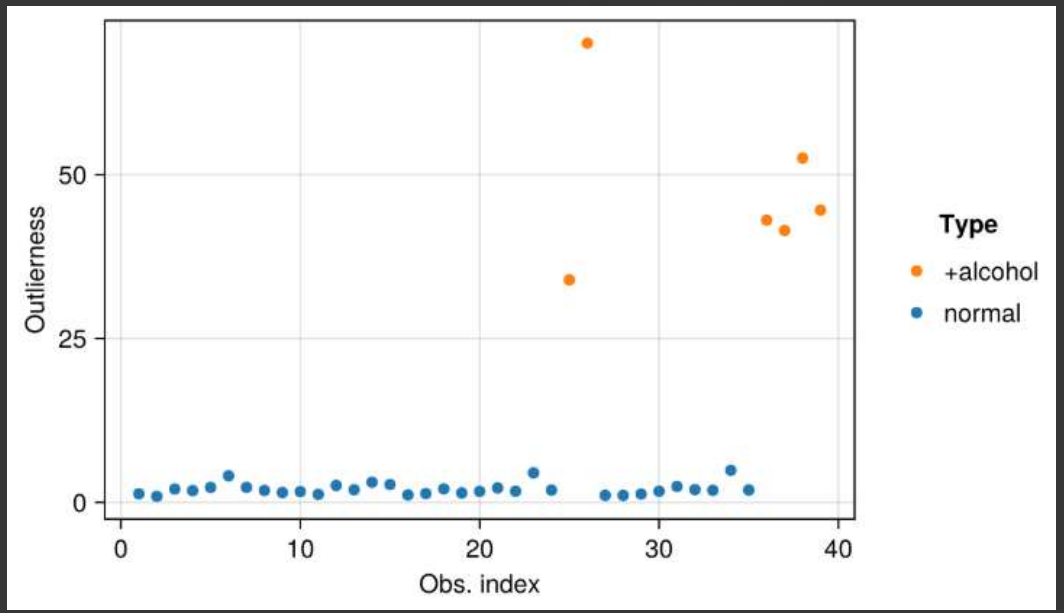
SD

Usual



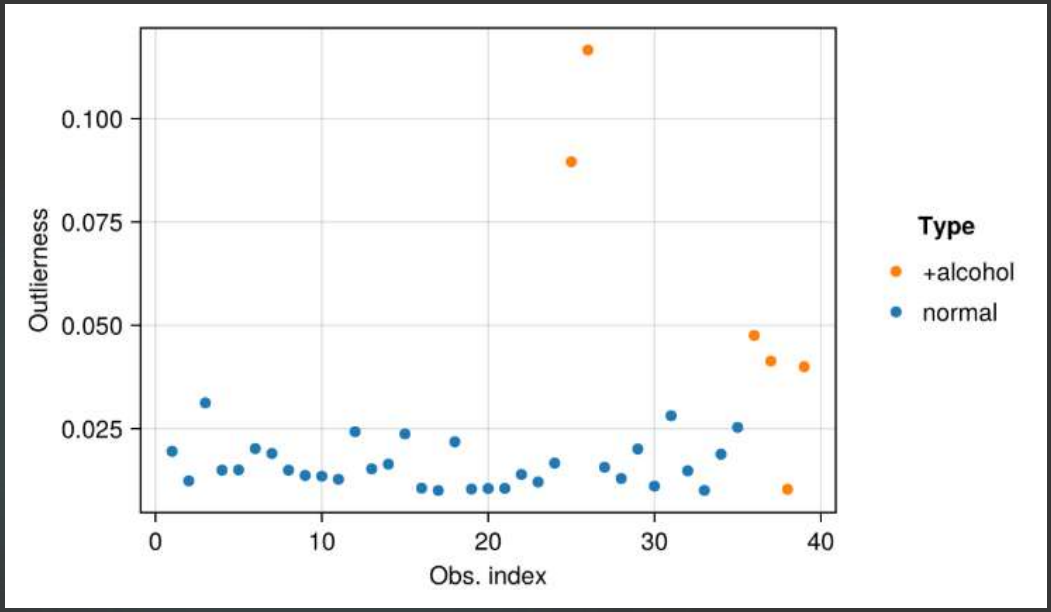
No detection

Robust

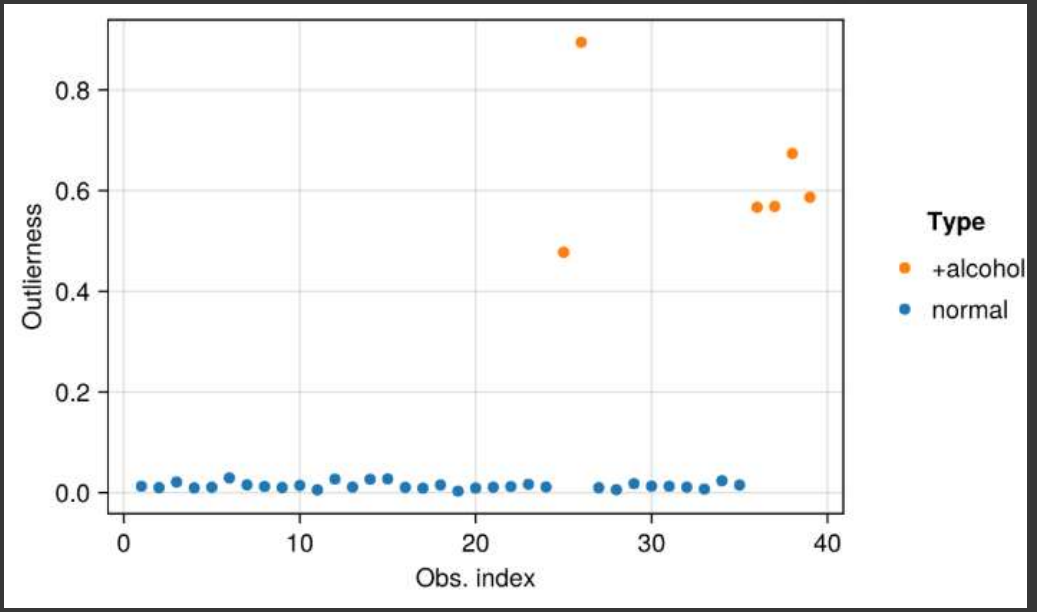


OD

Usual

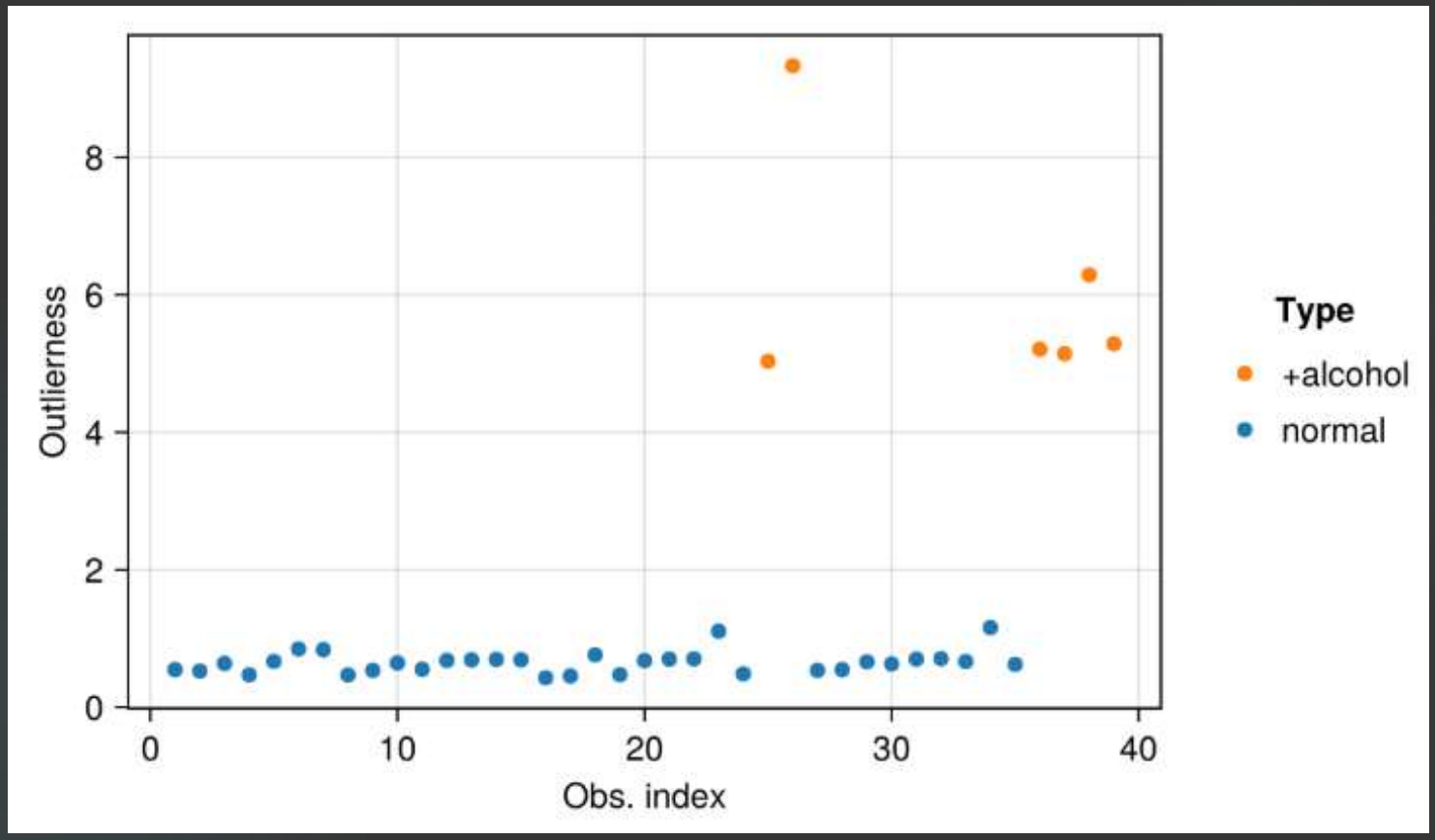


Robust



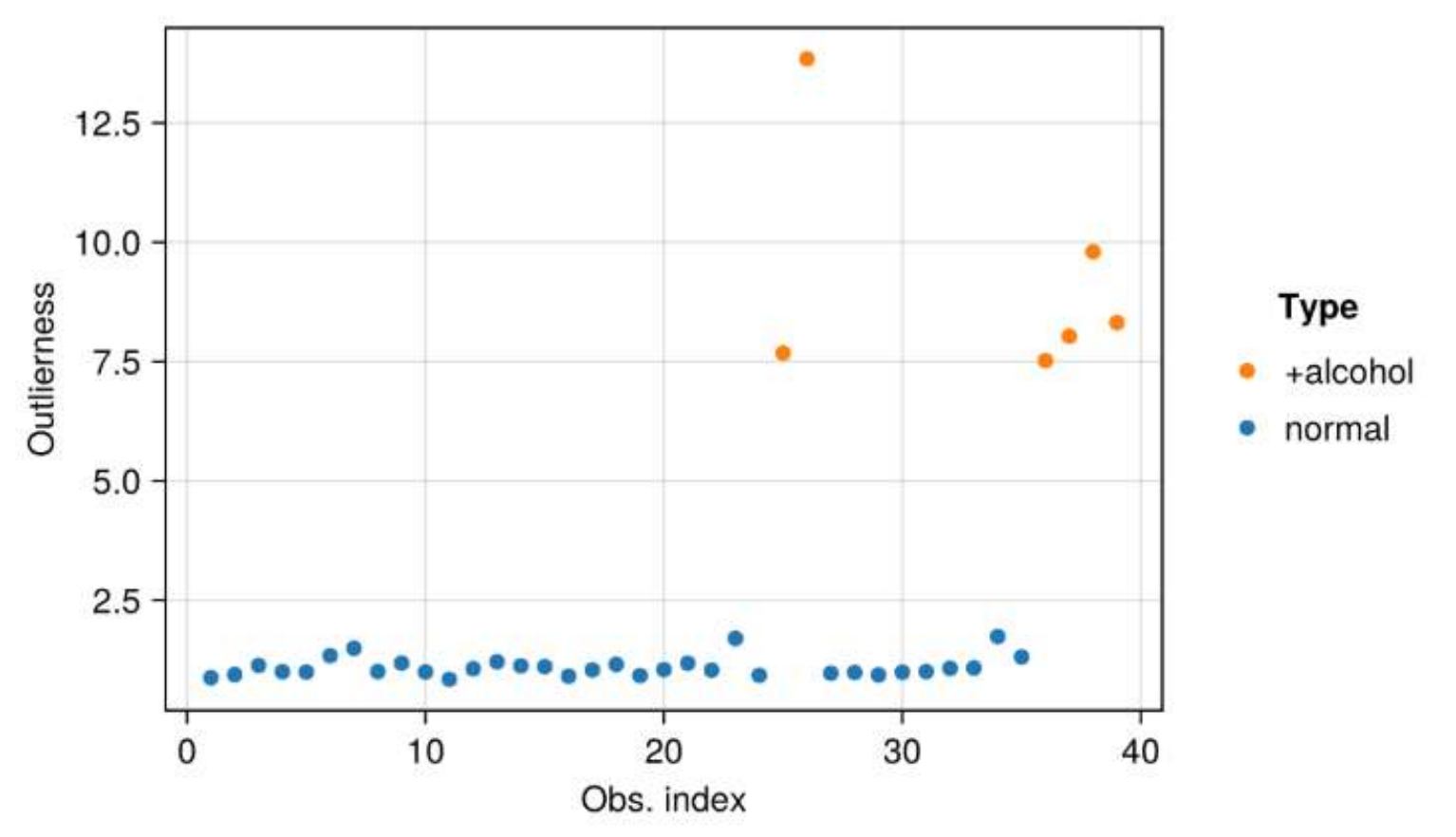
Partial detection

Global KNN X-space, Euclidean, k = 15



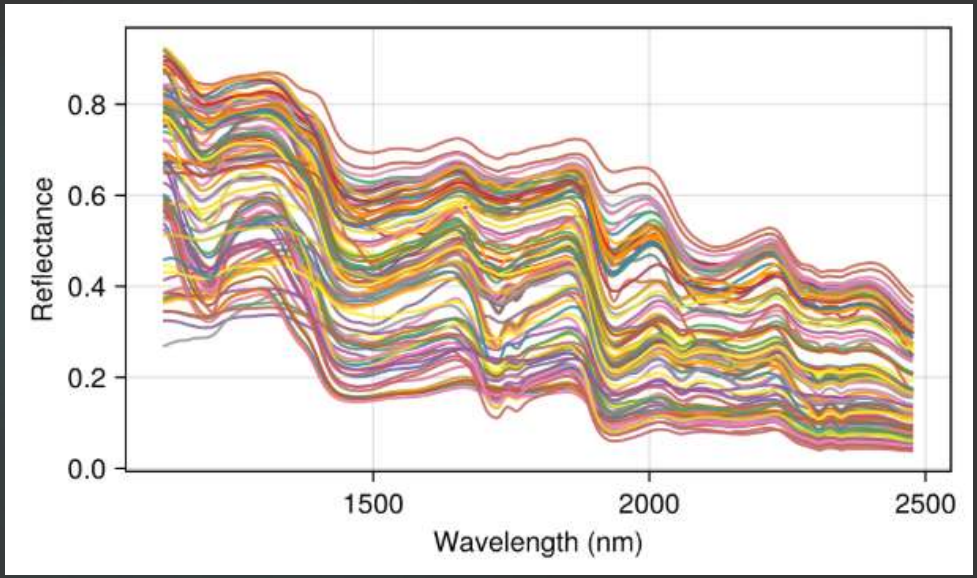
Rk: Same if KNN computed from PCA scores since PCA preserves the global distances

Local KNN same parameters

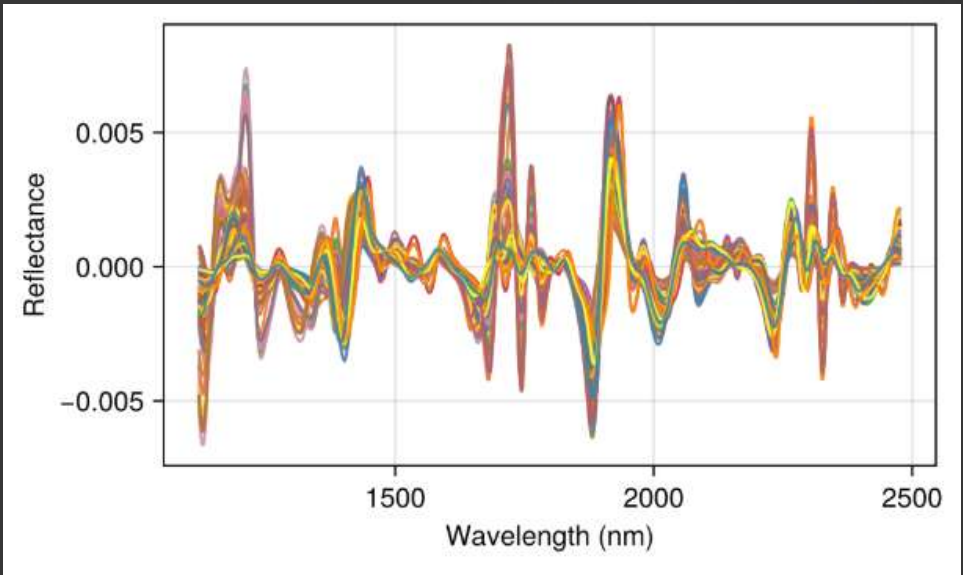


Challenge2018 dataset

NIRS data on forages, feed and food
used in the challenge of the congress Chemometrics2018
(Paris, January 2018)



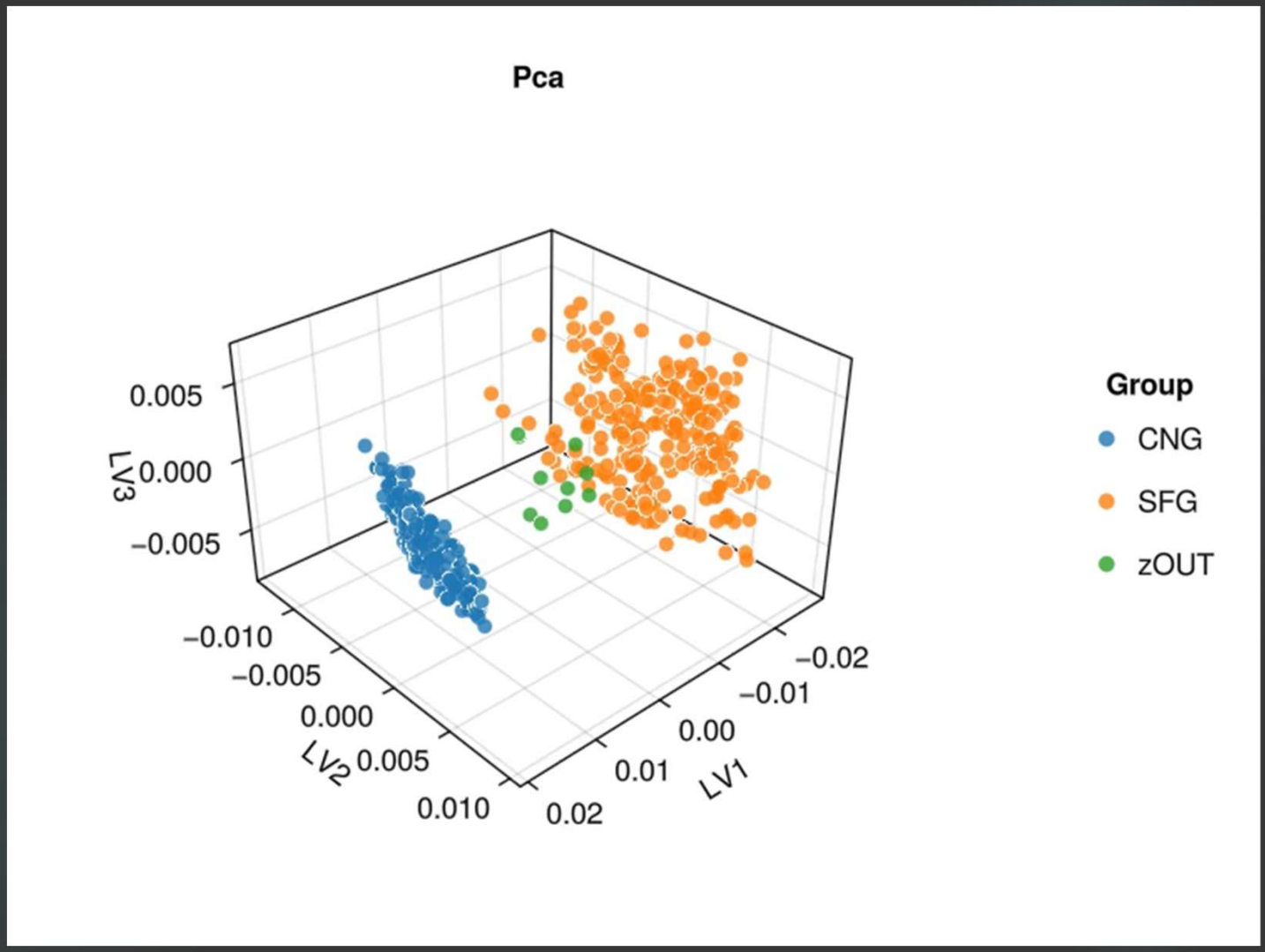
SNV + SavGol2



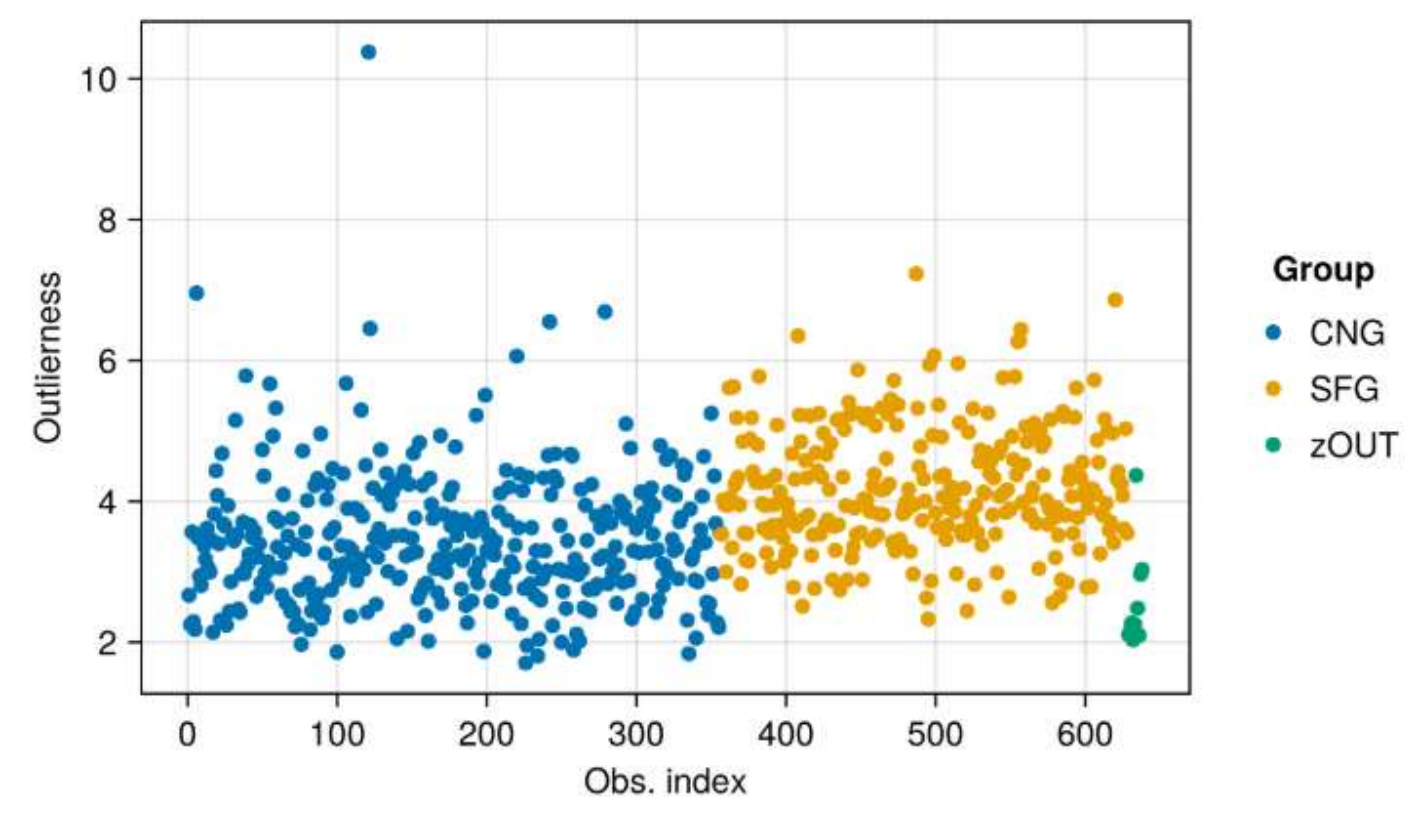
For this example: Extract of two of the 10 present categories

- CNG corn gluten n = 356
- SFG sun flower seed n = 272

+ n = 10 fictive outliers

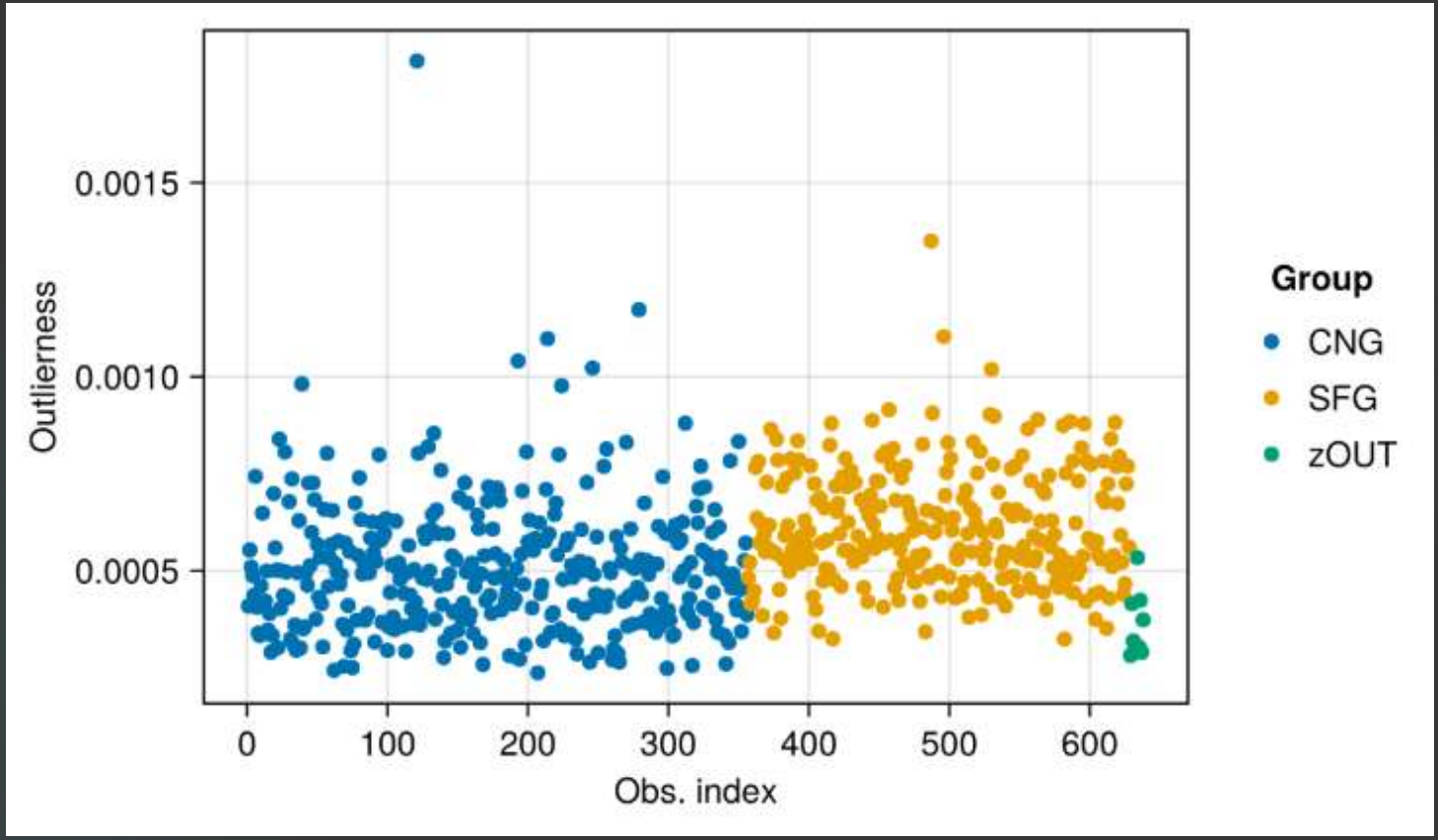

Pca
0.005
0.000
−0.005
LV3
−0.010
−0.005
0.000
0.005
0.010
LV2
−0.02
−0.01
0.00
0.01
0.02
LV1
Group
CNG
SFG
zOUT

PCA nlv = 15 SD



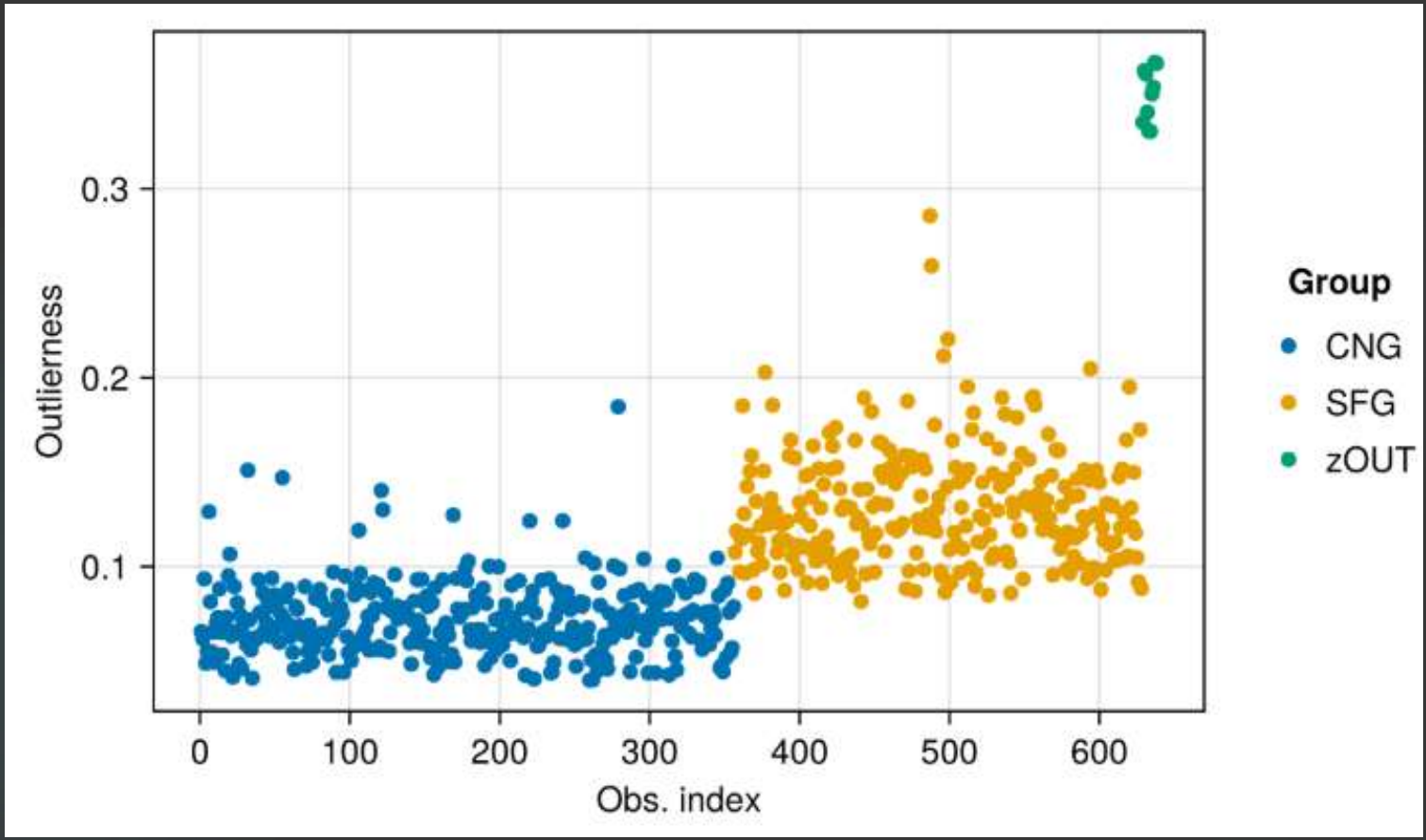
No detection

PCA nlv = 15 OD

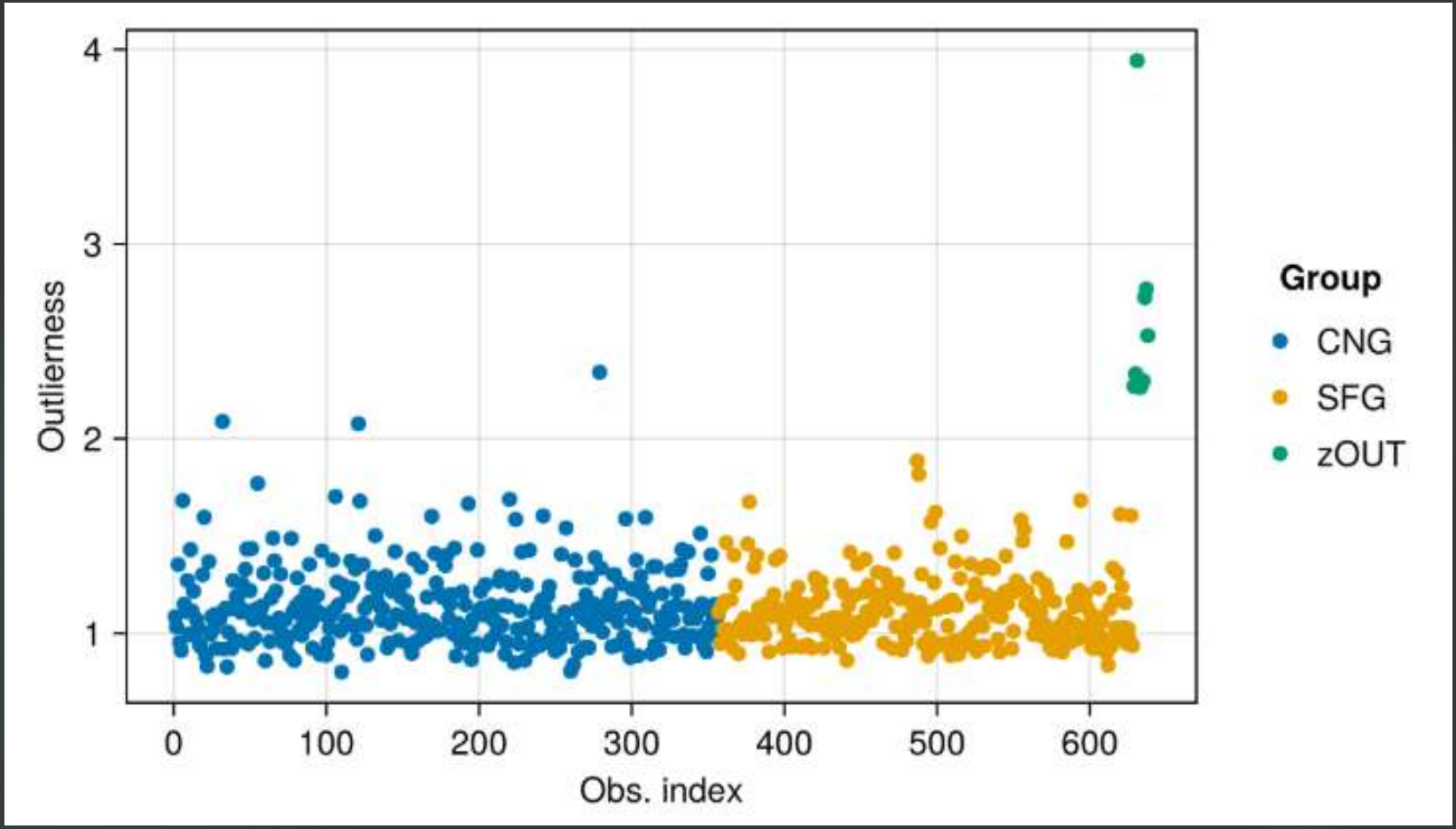


No detection

Global KNN Euclidean in X-space, k = 30



Local KNN Same parameters



⇒ PCA SD/OD are **not always the gold-standard**

It depends on the configuration of the data

Which cutoff values? (to automatize detections, etc.)

Parametric

- But require hypotheses on the outlierness distribution

Non parametric

- e.g.: Median(Out) + 3 × MAD(Out)

How to compute?

https://github.com/mlesnoff/Jchemo.jl

Functions

- occsd
- occod
- occsdod
- outknn
- outlknn