1 Outlier detection scheme

We now use the weighted projection quantiles for the purpose of detecting multivariate outliers. Relative to a multivariate data cloud, multiple outliers can either lie far away from majority of the data in a separate small cluster of points, or they may be scattered without any noticeable clumping. While there do exist k-nearest neighbor based [1] or depth-based [3,4] outlier detection methods, they only concentrate on local or global properties of a dataset, respectively. Here we combine our idea of depth outlined in section ? with a k-nearest neighbor distance measure to devise an outlier score for each observation in a multivariate dataset that can detect clustered or scattered outliers based on different values of a tuning parameter.

Definition Consider iid observations $\mathbf{X} = \{\mathbf{X_1}, ..., \mathbf{X_n}\} \in \mathbb{R}^{n \times p}$ from a multivariate distribution F. For any point $\mathbf{x} \in \mathbb{R}^p$ suppose $\bar{d}_k(\mathbf{x}, \mathbf{X})$ and $D(\mathbf{x}, \mathbf{X})$ are its k-nearest neighbor distance and WPQ-depth based on the data, respectively. The the depth-based outlier score for x is defined as:

$$O_{D,\alpha}(\mathbf{x}; \mathbf{X}) = \alpha . \log(\bar{d}_k(\mathbf{x}, \mathbf{X})) - (1 - \alpha) \log(D(\mathbf{x}, \mathbf{X}))$$

where $\alpha \in [0, 1]$ is the tuning parameter.

For $\alpha = 0$ this score becomes the negative log of the depth function, while $\alpha = 1$ makes this same as the log of mean kNN distance. This outlier score is defined based on the reasoning that a point far isolated from the rest of the data will always have a low depth, but whether it has a high kNN distance or not depends on if it is part of a small isolated clump of points or a single isolated point. For small values of α, $O_{D,\alpha}$ puts more emphasis on isolated points. On the other hand, for α close to 1 high values of the outlier score will tend to identify low-depth isolated points.

1.1 Simulations

We now consider two simulation scenarios to demonestrate the performance of our outlier score at different values of α. The k to obtain mean kNN distance is fixed at $\lfloor \sqrt{n} \rfloor$.

In the first setup we consider a 500-size sample, 95% of which are from $\mathcal{N}((0, 0)', I_2)$ and the other 5% drawn from $\mathcal{N}((10, 10)', I_2)$. The second setup also contains 500 data points, the last 25 having each element of their mean vector drawn from $\{\pm 6, ..., \pm 10\}$. Since the first group has clustered outliers, a smaller value of α should be able to identify points in the outlying cluster, while in the second setup a higher α should result in high score for points with a large mean nearest-neighbor distance and small depth, i.e. the scattered outlier points. Rows 2 to 4 in Fig. 1 give the index plots for outlier scores for these two scenarios, computed considering $\alpha = 0.05, 0.5, 0.95$. Points 1 to 475 are colored

green and the last 25, which are situated away from the main data cloud, are colored red. For the first group of samples, $\alpha = 0.05$ and 0.5 give a better distinction between the two populations, while in the second group this is best achieved for $\alpha = 0.95$.

1.2 Real data examples

Stackloss data This dataset due to Brownlee [2] has 21 observations from a plant regarding oxidation of Ammonia to Nitric Acid, and has 3 predictors: air flow, cooling temperature and concentration of acid; and percentage of ingoing Ammonia that escapes the oxidation process as response variable.The stackloss dataset has been widely used for detecting outliers in regression or unsupervised analysis. For example, the analysis of only the predictor variables due to Hadi [5] identifies observations {1,2,3,21} as potential outliers, while the bayesian model averaging-based approach taken by Hoeting *et al* [7] identifies observations 1, 3, 4 and 21 as outliers.

We set aside the response variable and calculate outlier scores based on the 3 predictors. Fig. 3 shows the scores for $\alpha = 0.05, 0.5$ and 0.95. For the first two plots, {1,2,3,7,8,21} are the top few point with high outlier scores. For $\alpha = 0.95$, there is considerable difference of scores between {1,2,3,21} and other points.

Hawkins, Bradu amd Kass data This artificial dataset given by Hawkins *et al* [6] consists of 75 observations and 4 variables (3 predictors and 1 response variable). The first 10 observations are high influential points while observations 11 to 14 are good leverage points. Analysis using classical methods like Mahalanobis' distance or Cook's distance only correctly identifies 11-14 as outliers and masks points 1 to 10. Hadi's unsupervised method of robust outlier detection [5] correctly identifies all 14 points as outliers. Our analysis of the 3 predictor variables using depth-based outlier scores (Figure 4) replicates Hadi's finding. The choice of the tuning parameter does not seem to matter here. Observation 14 is not shown in the index plots because its score was much higher than all others.

1.3 Analysis of DNA alteration data: combining outlier detection with classification

References

[1] M.M. Breunig, H.-P. Kriegel, R.T. Ng, and J. Sander. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104. SIGMOD, 2000.

[2] K.A. Brownlee. *Statistical Theory and Methodology in Science and Engineering*. Wiley, New York, NY, 1960, 2nd ed. 1965.

[3] Y. Chen, X. Dang, H. Peng, and H.L. Bart Jr. Outlier detection with the kernelized spatial depth function. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31-2:288–305, 2009.

[4] X. Dang and R. Serfling. Nonparametric Depth-Based Multivariate Outlier Identifiers, and Masking Robustness Properties. *J. Statist. Plan. and Inf.*, 140:782–801, 2010.

[5] A.S. Hadi. Identifying Multiple Outliers in Multivariate Data. *J. Royal Stat. Soc. Ser. B*, 54-3:761–771, 1992.

[6] D.M. Hawkins, D. Bradu, and G.V. Kass. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 26:197–208, 1984.

[7] J. Hoeting, A.E. Raftery, and D. Madigan. A method for simultaneous variable selection and outlier identification in linear regression. *Comput. Statist. and Data Anal.*, 22:251–270, 1996.

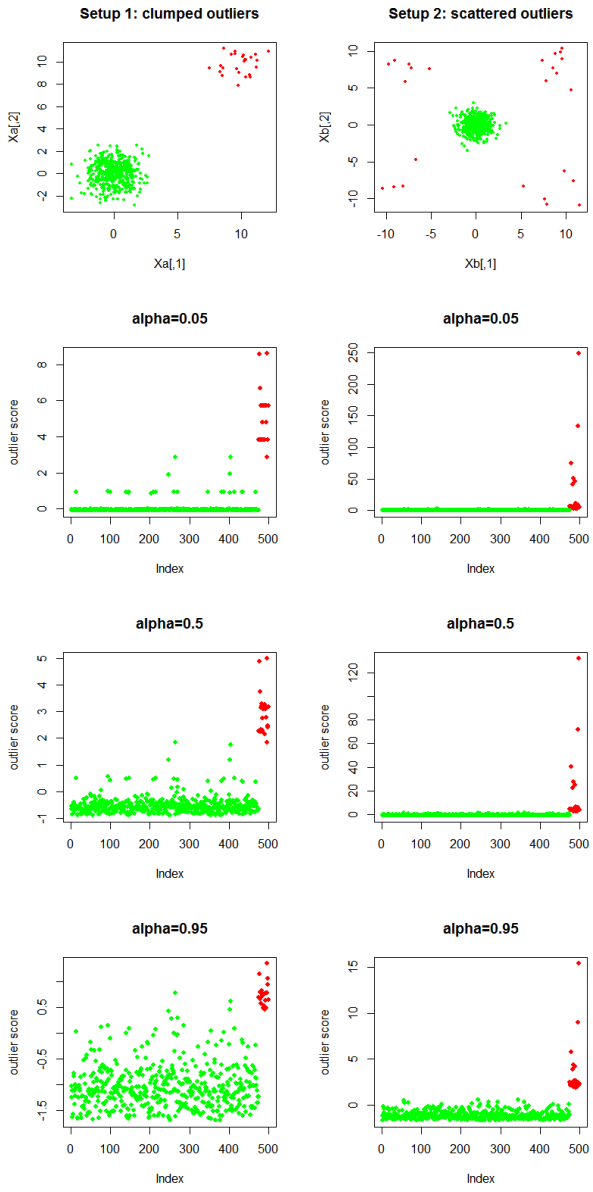

Figure 1: Scatter plot (top row) nd index plots (rows 2-4) for simulation setup 1 (left) and 2 (right)) $\alpha = 0.05, 0.5, 0.95$

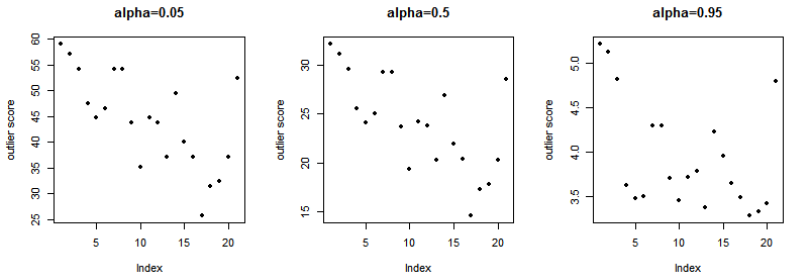

Figure 2: Outlier scores for stackloss data

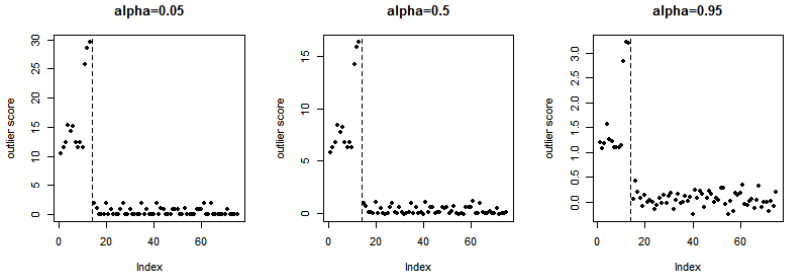

Figure 3: Outlier scores for Hawkins-Bradu-Kass data
(Dotted line at index = 14)