5th International Workshop on Climate Informatics
September 24-25, 2015
2015
Hosted by the National Center for Atmospheric Research in Boulder, CO

INFLUENCE RANKING BASED ON TILTING WITH APPLICATION TO CLIMATE DATA

Yuan Yan[1], Marc G. Genton[1]

Abstract—**The tilting method ranks each observation in terms of its influence on a general statistic, which can be mean, covariance, etc. This approach is based on 'tilting' or re-weighting each data value to achieve a given small change of the statistic, while minimizing the total amount of tilt. Then the influence ranking for each data corresponds to the rank of the tilted data weights. The tilting method can be applied to univariate, multivariate, functional or multivariate functional data. It allows for robust analysis and outlier detection. Climate data are intrinsically functional, either temporally or spatially or both. We illustrate the use of the tilting method by applying it to sea surface temperature data and bivariate data of mean monthly temperatures and precipitations recorded at Canadian weather stations.**

I. Method

The concept of order statistics is essential in robust inference, for example, to define the median and trimmed mean. Unlike on the real line, there exists no intuitive total order in multi-dimensional or functional spaces. The notion of depth is an extension from univariate to multivariate and functional settings to provide a center-outward ordering. There are many existing notions of depth for multivariate and functional data, refer to [1] and [2] for a review and properties related to data depth.

Genton and Hall [3] suggested a new interpretable approach to ranking data based on tilting. The tilting approach ranks data according to their influence on a statistic $\hat{\omega}(t)$. The idea is to first assign a tilted weight (p_i), instead of equal weights ($p_0 = 1/n$), to each observation, and compute the tilted statistic $\hat{\omega}_p(t)$. Then find the p_i such that the distance between p_i and p_0 is minimized while keeping the difference between $\hat{\omega}(t)$ and $\hat{\omega}_p(t)$ small. By doing so, we get a sequence of p_i on which the influence rank of an observation depends.

$^{t}hatwehaveadatamatrix\mathrm{X}_{n\times m}$ with sample size n and m variables. Let $\boldsymbol{p} = (p_1, \ldots, p_n)^\top$ be a multinomial distribution on n points and $\boldsymbol{p}_0 = (\frac{1}{n}, \ldots, \frac{1}{n})^\top$. The statistics of interest for the tilting method is the sample mean $\hat{\boldsymbol{\omega}}_{\boldsymbol{p}_0} = X^\top \boldsymbol{p}_0$ and the tilted version is $\hat{\boldsymbol{\omega}}_{\boldsymbol{p}} = X^\top \boldsymbol{p}$. The tilting approach to ranking influence for the functional mean can be formulated by the following constrained optimization problem:

$$\begin{aligned} \min_{\boldsymbol{p}} \quad & \sum_{i=1}^{n} p_i \log\,(np_i) \\ \text{subject to} \quad & \|\boldsymbol{r}(\boldsymbol{p})\|^2 = \epsilon^2,\ \sum_{i=1}^{n} p_i = 1,\ \boldsymbol{p} \geq \boldsymbol{0}, \end{aligned}$$

where $\sum_{i=1}^{n} p_i \log\,(np_i)$ is the Kullback-Leibler divergence as a measure of distance between $\boldsymbol{p}$ and $\boldsymbol{p}_0$, and $\boldsymbol{r}(\boldsymbol{p}) = \hat{\boldsymbol{\omega}}_{\boldsymbol{p}} - \hat{\boldsymbol{\omega}}_{\boldsymbol{p}_0} = X^\top(\boldsymbol{p} - \boldsymbol{p}_0)$.

After solving the problem and getting the vector $\boldsymbol{p}$, we define the tilting depth as the value

$$TD(\boldsymbol{x}_i) = |p_i - \text{median}(\boldsymbol{p}_0)|$$

and we can use this tilting depth to make robust inference. As proved in [3], as $\epsilon \to 0$, the ranks of $\boldsymbol{p}$ correspond to the ranks of the lengths of each observation after projection to the univariate space spanned by the first empirical orthogonal function. From this perspective, the tilting depth is related to the random projection depth defined in [5]. Rather than projecting to a randomly selected direction, the tilting approach is equivalent to projecting to the direction of the first empirical orthogonal function.

This depth has the advantages of being distance based and possessing an intuitive interpretation of the ranks based on projection to the principal component function.

II. Outlier Detection

Sun and Genton [4] proposed the functional boxplot as an extension of the univariate boxplot to the functional setting. Depending on the depth it is based on, the functional boxplot defines the sample 50% central region as the band delimited by half of the deepest curves from the sample. Then the fences are obtained by expanding the 50% central region 1.5 times and curves outside the fences are detected as potential outliers. Sun and Genton

Corresponding author: Yuan Yan, yuan.yan@kaust.edu.sa
[1]CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

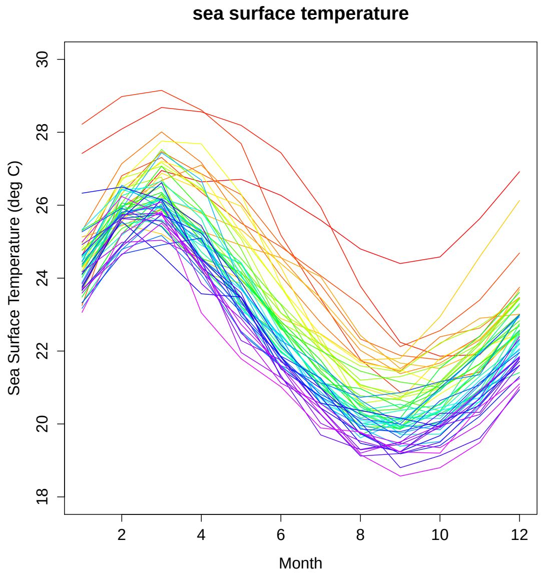

Fig. 1. Rainbow plot for the SST data based on tilted weight $\boldsymbol{p}$

[7] also proposed the adjusted functional boxplot, which modified the expanding factor 1.5 to be chosen by a simulation based method. We use the tilting depth in conjunction with the (adjusted) functional boxplot to detect outliers.

III. Applications

To illustrate the use of the tilting method, we first apply it to sea surface temperature data. This dataset consists of average monthly sea surface temperatures from January 1951 to December 2007 measured for the Niño 1+2 region (i.e., $0 - 10$°S, 90°W-80°W).

We can draw a rainbow plot [6] according to the tilted p_i for each observation. Since the tilted value $\boldsymbol{p}$ accepts sign, this is an ordering from 'positive' to 'negative' (See Figure 1). It turns out that the red and purple curves are the strongest El Niño and La Niña years respectively.

Then we illustrate the outlier detection performance by applying our method to a bivariate data of mean monthly temperature and precipitation at 35 different locations in Canada averaged over 1960 to 1994 [8].

By using the functional boxplot combined with the tilting depth, we detect 8 outliers for the Canadian weather data. Seven of the outlying stations lie in the northern part of Canada and possess overall cold temperature. The other outlying station is Pr. Rupert which is a port city and has a high level of precipitation.

IV. Further Studies

Improvement of this method can be made by using a different distance measure of the difference between the statistic $\hat{\boldsymbol{\omega}}_{\boldsymbol{p}_0}$ and its tilted version $\hat{\boldsymbol{\omega}}_{\boldsymbol{p}}$. We may first apply robust scaling to standardize the variables or use the Mahalanobis distance. We plan to also further explore this method by using other statistics of interest (e.g. covariance, skewness or kurtosis).

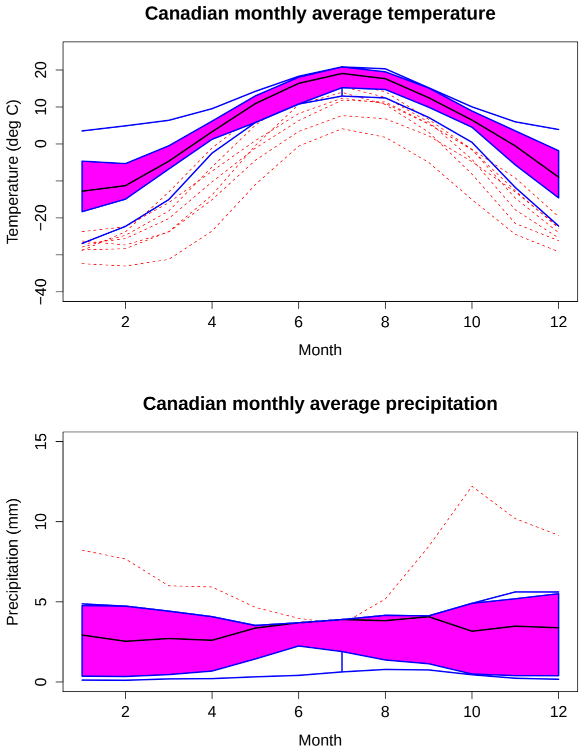

Fig. 2. Outlier detection by functional boxplot combined with the tilting depth for Canadian weather data

References

[1] Zuo, Y., Serfling, R., "General notions of statistical depth function," *Annals of Statistics*, **28**: 461-482, 2000.

[2] Mosler, K., "Depth statistics," In Becker, C., Fried, R., Kuhnt, S. (eds) *Robustness and complex data structures: Festschrift in Honour of Ursula Gather*. Springer, Berlin, pp.17-34, 2013.

[3] Genton, M. G., Hall, P., "A tilting approach to ranking influence," *Journal of the Royal Statistical Society, Series B*, to appear, 2016.

[4] Sun, Y., Genton, M. G., "Functional boxplots," *Journal of Computational and Graphical Statistics*, **20**: 316-334, 2011.

[5] Cuevas, A., Febrero, M., Fraiman, R., "Robust estimation and classification for functional data via projection-based depth notions," *Computational Statistics*, **22**: 481-496, 2007.

[6] Hyndman, R., Shang, H., "Rainbow plots, bagplots, and boxplots for functional data," *Journal of Computational and Graphical Statistics*, **19**(1): 29-45, 2010.

[7] Sun, Y., Genton, M. G., "Adjusted functional boxplots for spatio-temporal data visualization and outlier detection," *Environmetrics*, **23**: 54-64, 2012.

[8] Ramsay, J., Silverman, B. W., *Functional data analysis*, 2nd edn. Springer, New York, 2005.