A Semi-Parametric Method for Robust Multivariate Error Detection in Skewed Functional Data with Application to Historical Radiosonde Winds

Ying Sun[1], Amanda S. Hering[2,*] and Doug Nychka[3]

[1] *Division of Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia; ying.sun@kaust.edu.sa*
[2] *Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO 80401, USA. 303.273.3880; ahering@mines.edu*
[3] *Computational Information Systems Laboratory, National Center for Atmospheric Research, Boulder, CO 80305, USA. 303.497.1000; nychka@ucar.edu*
[*] *Corresponding author*

Abstract. *Quality control methods for multivariate data are generally based on using robust estimates of parameters for the multivariate normal (MVN) distribution. However, many multivariate data generating processes do not produce elliptical contours, and in such cases, error detection using the MVN distribution would lead to many legitimate observations being erroneously flagged. In this work, we develop a semi-parametric method for identifying errors in skewed multivariate data that also has a functional component. In the first step, we remove potential outliers by assigning each multivariate observation or function a depth score and remove those observations that fall beyond a given threshold. The remaining observations are used to estimate the parameters in a multivariate skew-t (MVST) distribution, and this estimated distribution is used in assigning all observations a probability of having been generated from this MVST. We test the performance of this two-step approach in simulation against a more common MVN method adapted for functional data. When the observations are skewed, our approach has a higher percentage of correctly identified outliers and a lower percentage false positives. Finally, we show how our method can be used in practice with radiosonde launches at a Denver, Colorado station of horizontal and vertical wind components measured at 8 vertical pressure levels.*

Keywords. *Multivariate skew-t distribution; Outlier detection; Radiosondes; Wind profiles*

1 Introduction

Detecting multivariate outliers is an inherently difficult problem since an observation may not be considered an outlier in any one given dimension, but it could be unusual when considered jointly across all of its dimensions. The most common approach to detecting multivariate outliers is to use robust estimation of the parameters of a multivariate normal (MVN) distribution [1, 2], since the outliers themselves can influence the parameter estimates. Then, a MVN based algorithm to detect outliers is applied using the robust estimates of location and scale in Mahalanobis distance [3]. However, many processes do not fit the MVN distribution profile and may have heavy tailed and/or skewed distributions. The multivariate skew-*t* (MVST) distribution is flexible enough to fit such variations in the third and fourth moments of a

distribution, which accommodate skewness and kurtosis, respectively, and has the MVN distribution as a special, central case [4].

The wind vector with horizontal, u, and vertical, v, components does not typically follow a multivariate normal distribution, an example of which is given in Figure 1. Robustly estimated MVN contours are in the left plot ([1, 5]), and a method based on robust Mahalanobis distances and the MVN distribution [3] flags all of the red dots as outliers when they are not. The contours based on our MVST maximum likelihood estimation are given in the right-hand plot, and only one potential outlier is flagged after an initial non-parametric sweep of the observations. In particular, we are motivated by the problem of error detection in winds recorded in historical radiosonde launches in which extreme values are still of interest, but errors are prevalent throughout the record. Radiosondes are instruments attached to weather balloons that measure atmospheric variables, including winds, and are released daily at stations around the globe.

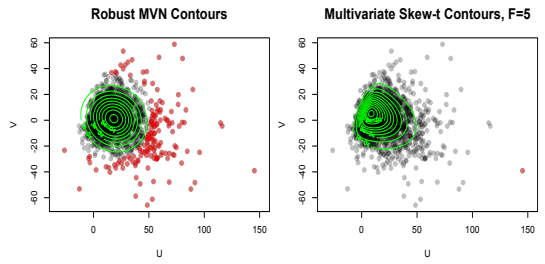

Figure 1: Scatterplots of u and v components simulated from a MVST distribution with robust MVN (left) and MVST (right) contours overlaid. Flagged observations are in red.

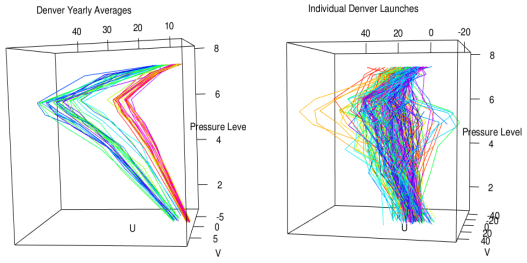

Figure 2: On the left are yearly averages from 1962 to 2011 of the wind profiles at Denver. On the right are 230 launches in 1962 at the Denver site.

Each launch of a radiosonde produces a function of wind over pressure level. Figure 2 shows the u and v components of wind plotted as a function of pressure for yearly averaged launches over 50 years from 1962 to 2011 (left) and over 200 individual launches in 1962 (right) at the Denver station. In the yearly averages, there is a clear shift in observed winds, which indicates a change in instrumentation or slight shift in station location, representing a systematic change. In the individual launches, there is a substantial amount of variability with an increase in spread in the mid-pressure levels where the jet-stream is located, and we focus this work on identifying random errors in the historical archive as statistical methods are less applicable for dealing with systematic errors.

We design a method for handling the multivariate, skewed, and functional aspects of this data. In the remainder of this paper, we describe three approaches for multivariate error detection: a parametric MVN approach [3]; a non-parametric depth approach; and a semi-parametric approach using depth and the MVST distribution. These methods are evaluated through simulation, and we record the number of correctly classified errors (true positives) and incorrectly identified true observations (false negatives), with the goal of having high values of the former and low values of the latter. The effect of skewness on the methods is also tested as well as the effect of applying the methods by launch or by pressure level.

2 Error Detection Methods

The methods described here are applied in our simulation study to evaluate their properties. We simulate the u and v wind components at all 8 pressure levels simultaneously. To do this, we use a 16-dimensional MVST distribution, fitted based on 10,000 launches at the Denver, Colorado station. We use these esti-

mated parameters as the observed level of skewness and kurtosis. Then, we also simulate data assuming a MVN distribution but with the same center and spread or assuming more extremely skewed and heavier tails than what was observed. For each data type, we either (i) insert no outliers; (ii) contaminate an entire launch; (iii) contaminate the higher levels of a launch; or (iv) contaminate random levels.

Multivariate Normal Method: The traditional approach in [3] for detecting errors assumes that the primary underlying p-dimensional process follows a MVN distribution and uses Mahalanobis distance to measure the distance of observations from the center of the distribution, as follows:

$$MD_i^r = ((\mathbf{x}_i - \boldsymbol{\mu}_r)' \boldsymbol{\Sigma}_r^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_r))^{1/2},$$

but in place of the classical estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, uses the robust estimators, μ_r and Σ_r [5]. Under the assumption of MVN, $(MD_i^r)^2 \sim \chi_p^2$. The threshold for determining when an observation is an outlier could be $\chi^2_{1-\alpha,p}$ for an appropriate choice of α, but this ignores the fact that as the sample size gets large, the likelihood of observing extreme values increases. Thus, the authors introduce an adjusted threshold for classifying observations as errors obtained through simulation under $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}_r$. This method can be applied for an individual pressure level, in which case $p = 2$, or for an entire launch wherein $p = 16$.

Depth Method: In the depth method, we can take advantage of the functional component of the launches and do not assume MVN. We take the following steps:

1. Let $\mathbf{x}_j(h_i) = (u, v)_j^T$ be the bivariate observations in $\mathbb{R}^2$ given $h_i \in \{700, 500, 400, 300, 250, 200, 100, 70\}$ for q levels of launches. Let $j = 1, \ldots, m_i$ be the number of observations for h_i. If there are no missing values, then $m_1 = \cdots = m_q = n$.
2. Given h_i, compute Tukey's depth values for each observation, $\mathbf{x}_j$, denoted $HD_i(\mathbf{x}_j)$. The Tukey's depth of a point $\mathbf{x}$ relative to a bivariate dataset is given by the smallest number of data points contained in a closed half-plane of which the boundary line passes through $\mathbf{x}$.
3. For each launch, define the bivariate functional data as $\mathbf{y}_j = (\mathbf{x}_j(h_1), \ldots, \mathbf{x}_j(h_q))$.
4. Define the Bivariate Functional Depth (BFD) value of the jth launch as a weighted average of the Tukey's depth at each pressure level:
$$BFD(\mathbf{y}_j) = \sum_{i=1}^{q} w_i HD_i(\mathbf{x}_j),$$
where the weights are proportional to the number of observations on each level.
5. Each launch has one depth value $BFD(\mathbf{y}_j)$ indicating its location in the sample. The functional median is the launch with the largest BFD. A convex hull is constructed based on the BFD representing the 50% central region.
6. Fences are determined by inflating the central convex hull by F times the distance from the median to the central convex hull. An appropriate value of F can be found through simulation using the same sample size and data dimension when there are no outliers.

An alternative approach by pressure level that classifies observations using only $HD_i(\mathbf{x}_j)$ as opposed to $BFD(\mathbf{y}_j)$ can also be applied.

Multivariate Skew-t Method: A p-dimensional random vector, $\mathbf{Y}$, following a MVST distribution can be described as follows [4]:

$$\mathbf{Y} \sim MVST_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu),$$

where $\boldsymbol{\xi}$ is $p \times 1$ and is like a measure of center; $\boldsymbol{\Omega}$, $p \times p$, is like a variance-covariance matrix; $\boldsymbol{\alpha}$ is $p \times 1$ and measures the skewness in each dimension; and ν is scalar and represents the heaviness of the tails. The quadratic form, $(Y-\boldsymbol{\xi})^T \boldsymbol{\Omega}^{-1}(Y-\boldsymbol{\xi}) \sim p \cdot F(p,\nu)$. We follow a similar approach as [3], but we first sweep the data with either the bivariate depth or functional data depth methods to initially remove very large observations before estimating the parameters of the distribution. We also adjust the threshold for flagging observations, but we base this adjustment on the MVST distribution and its robustly estimated parameters. We have found that the estimation of ν, in particular, is very sensitive to the presence of errors. Again, we can apply this approach for each pressure level or for an entire launch.

3 Results

In simulations, we have found the MVST approach applied at each pressure level performs the best in terms of the lowest number of false positives (incorrectly flagged true observations) and the highest number of true negatives (correctly flagged errors). The MVN method is very sensitive to the skewness and tends to flag a higher proportion of true values as outliers, on average, as the skewness increases. The bivariate depth method does not perform as well as either of the MVST or the MVN methods, but in using it as a first sweep of the observations before applying the MVST method works very well. Sweeping the observations with the functional data depth also performs well but not quite as well as removing observations by pressure level. Because of space limitations, full results are not shown here.

4 Conclusion

The methods developed and experiments performed thus far have demonstrated that the MVST method with outlying observations first swept by the functional or bivariate depth method is a very good candidate for application to the entire wind radiosonde archive. In practice, more understanding of the archive is required before we should use the MVST approach for fast, statistical quality control. For example, there is much missing data in the archive, and finding the optimal number of observations to use in a moving window through a station's launches over time will be necessary.

References

[1] Rousseeuw, P. J. and Van Driessen, K. (1999) "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, 41: 212–223.

[2] Peña, D. and Prieto, F. J. (2001) "Multivariate outlier detection and robust covariance matrix estimation," *Technometrics*, 43: 286–300.

[3] Filzmoser, P., Reimann C., and Garrett R. G. (2005) "Multivariate outlier detection in exploration geochemistry," *Computers and Geosciences*, 31: 579–587.

[4] Azzalini, A. and Capitanio A. (2003) "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t* distribution," *Journal of the Royal Statistical Society, Series B*, 65: 367–389.

[5] Rousseeuw, P. J. (1985) "Multivariate estimation with high breakdown point," in Grossmann, W., Pflug, G., Vincze, I. Wertz, W. (Eds.), *Mathematical Statistics and Applications, Vol. B*, Akadémiai Kiadó: Budapest, Hungary, pp. 283–297.

[6] Rousseuw, P.J. and Ruts, I. (1996) "Bivariate location depth," *Applied Statistics–Journal of the Royal Statistical Society, Series C*, 45: 516–526.