A Note on Automatic Data Transformation

Qing Feng, Jan Hannig and J.S.Marron

Department of Statistics and Operations Research

The University of North Carolina at Chapel Hill

Abstract

Modern data analysis frequently involves variables with highly non-Gaussian marginal distributions. However, commonly used analysis methods are most effective with roughly Gaussian data. This paper introduces an automatic transformation that improves the closeness of distributions to normality. For each variable, a new family of parametrizations of the shifted logarithm transformation is proposed, which is unique in treating the data as real-valued, and in allowing transformation for both left and right skewness within the single family. This also allows an automatic selection of the parameter value (which is crucial for high dimensional data with many variables to transform) by minimizing the Anderson-Darling test statistic of the transformed data. An application to image features extracted from melanoma microscopy slides demonstrate the utility of the proposed transformation in addressing data with excessive skewness, heteroscedasticity and influential observations.

keywords: Automatic Transformation, Shifted Logarithm Transformation, Anderson-Darling Test Statistics, Heteroscedasticity.

1 Introduction

Technological developments have led to methods for generating complex data objects such as DNA chip data and digital images of tumors. These new types of data objects frequently strongly violate the approximate normality assumption which is commonly made in statistical techniques. Therefore, an appropriate data transformation can be very useful for improving the closeness of the data distribution to normality.

Many transformation techniques have been proposed. Sakia (1992) provided a comprehensive review of the Box-Cox (Box and Cox, 1964) and related transformations. Various methods have been developed for selecting the transformation parameters, including the maximum likelihood

method (Box and Cox, 1964), robust adaptive method (Carroll, 1980), Kullback-Leibler information based method (Hernandez and Johnson, 1980), and Kendall's rank correlation based method (Han, 1987).

A commonly used member of the Box-Cox family is the logarithm transformation, which is useful for tackling data sets generated by a multiplicative process. Furthermore, the logarithm transformation can stabilize the asymptotic variance of data. One important application is to transform some types of microarray data. A shift parameter was further introduced to make the logarithm transformations more flexible and useful. See Section 3 of Yang (1995) for a good overview of the shifted logarithm transformation. The parameterizations of the shift parameter strongly depend on knowledge of the data e.g. data range, data distribution, so user intervention is usually required. However, modern high-output data sets usually have a very large number of variables, i.e. features, so there is a strong need to automate the selection of shift parameter, which is an important contribution of this paper.

We propose a new automatic data transformation scheme for making various types of marginal distributions close to being normally distributed. In particular, we aim at addressing certain types of departures from normality e.g. strong skewness. Our proposed method focuses on the family of shifted logarithm transformations and introduces a new parametrization which treats the data as lying on the entire real line. Besides, our parametrization makes the selection of shift tuning parameter independent of data magnitude which is an advantage for automation. This algorithm is designed to automatically select a parameter value such that the transformed data has the smallest *Anderson-Darling test statistic*. Furthermore, this transformation scheme includes a winsorisation of influential observations based on the extreme value theorem.

1.1 Data Example

A motivating data example is digital image analysis in a study of mutant types of melanocytic lesions (Miedema et al., 2012). Many of the raw features extracted from digital images contain excessive skewness. For example, the marginal distributions of two of the image features are visualized by the *kernel density estimated plots* (KDE plots) in the top row of Figure 1. The blue curves are the Gaussian kernel density estimate i.e. smoothed histograms, using Sheather-Jones plug-in bandwidths. (See Chapter 3 of Wand and Jones (1994) for the comparison of bandwidth selection methods.) The green dots are jitter plots of the data. Each symbol is a data point whose horizontal coordinate is the value and vertical coordinate is based on data ordering for

visual separation. As can be seen, these distributions are highly skewed. For such data sets with substantial skewness, an analysis based on a Gaussian assumption would tend to generate poor results. The bottom plots of Figure 1 display the KDE plots of each feature vector after our automatic transformation. The kernel density estimates (blue curves) are approximately symmetric.
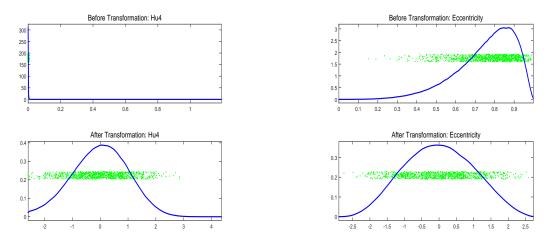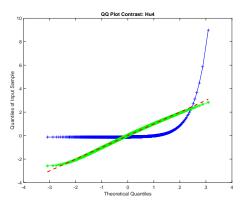

Figure 1: Comparison of the KDE-plots of two image feature vectors before (top row) and after (bottom row) transformation. This shows that the transformed distributions are much closer to Gaussian for data with both positive (Hu4, left column) and negative (Eccentricity, right column) skewness.

The Q-Q Plot in Figure 2 gives a more precise measure of closeness to the standard normal distribution. The left panel shows the Q-Q plots for Hu4 applied with standardization only (blue plus signs) and for Hu4 after automatic transformation (green stars). The symbols are the quantiles of 1000 randomly selected data points against the theoretical quantiles of the standard normal distribution. For comparison, we also show the 45 degree red dashed line. The blue plus signs clearly depart from this line, while the green stars approximately lie on the line. This contrast suggests a dramatic improvement in normality by our automatic transformation of Hu4. A similar improvement in normality of Eccentricity is also shown in the right panel. Although there are slight departures at each tail of the transformed data, an overall improvement can be seen as the majority of the quantiles approach the theoretical quantiles of the standard normal distribution.

Even though our transformation acts only on the marginal distributions, it often results in major improvement of the joint distribution of the features. In Figure 3, the scatter plot on the left shows a strong non-linear relationship between the Hu4 and Eccentricity that were studied in Figures 1 and 2. After transformation, the scatter plot on the right shows a bivariate Gaussian
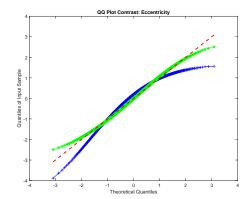
3

Figure 2: The QQ-plots of Hu4 (left) and Eccentricity (right). The comparison between before (blue plus signs) and after (green stars) indicates a major overall improvement in closeness to normality made by transformation.

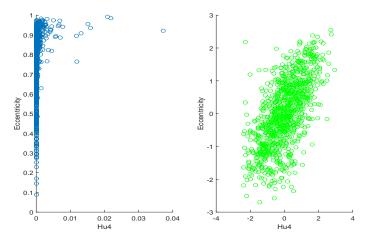relationship which is much more amenable to analysis using standard statistical tools.

Figure 3: Comparison of the scatter plots, showing the joint distributions from Figure 1, before (left) and after (right) transformation. Relationship after transformation is much closer to linear.

2   Methodology

In this section, a novel automatic data transformation scheme is proposed for general data sets to achieve approximate normality. For any given data set, the transformation works feature by feature. In other words, for a data matrix with columns considered as data objects and rows considered as features, the transformation is applied to each row.

The transformation scheme consists of three components: a family of shifted logarithm trans-

formation functions indexed by a parameter β, standardization with an option for winsorisation of extreme observations and an evaluation of the transformation with a given parameter value. The key steps will be introduced in the following subsections.

The transformation scheme is a grid search based on three components to determine the optimal value of β for each feature, which is outlined as

- *Initialization:* Construct a grid of parameter values $\beta = \{\beta_k, k = 1, \cdots, n\}$

- *Step 1:* Apply the transformation function to the feature vector for each parameter value β_k.

- *Step 2:* Standardize the transformed feature vector and winsorise any existing extreme observations. Re-standardize the feature vector if winsorisation has been done.

- *Step 3:* Calculate the Anderson-Darling test statistic.

Lastly, select β to minimize the Anderson-Darling test statistic for normality.

2.1 Transformation Function

A new parametrization of the family of shifted logarithm functions, $\{\phi_\beta, \beta \in \mathbb{R}\}$, is proposed for addressing both left and right skewness in features. For a feature vector $X = (X_1, X_2, \cdots, X_n)$, the sample skewness is $g(X) = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2)^{\frac{3}{2}}}$ where $\bar{X}$ is the sample mean of the vector.

As convex transformation functions tend to increase the skewness of data while concave transformations reduce it (van Zwet, 1964), the transformation functions are chosen to be concave for $g(X) > 0$ and convex for $g(X) < 0$. As logarithm functions are concave, the transformation function can be a logarithm for the case $g(X) > 0$ i.e. $\log(X_i)$. While for the other case $g(X) < 0$, the transformation function should be made convex by inserting negative signs within and before a logarithm function i.e. $-\log(-X_i)$.

Since logarithm functions require positive inputs, it is important to modify the functions for both cases to be valid for any X_i. For example, in the case $g(X) > 0$, this concern can be resolved by subtracting the minimal value of the feature vectors from X_i and adding a positive shift parameter α. That is,

$$\log(X_i - min(X_1, X_2, \cdots, X_n) + \alpha), \tag{1}$$

Similarly for the negative skewness $g(X) < 0$, the function is

$$-\log(max(X_1, X_2, \cdots, X_n) - X_i + \alpha), \tag{2}$$

5

The shift parameter α is further parameterized in terms of the multiples of the range of the feature vectors i.e. $R = max(X_1, X_2, \cdots, X_n)$ - $min(X_1, X_2, \cdots, X_n)$. This makes the selection of parameter values independent of the data magnitude. In particular, set

$$\alpha = |\frac{1}{\beta}|R. \tag{3}$$

By tuning the value of β, the effect of the transformation varies. In particular, the transformation together with standardization is equivalent to standardization only, when the parameter β approaches 0. In order to make the resulting transformation function $\phi_\beta(X_i)$ continuous over $\beta \in \mathbb{R}$, we define our transformation to be standardization only for $\beta = 0$.

Incorporating all these elements, the formal representation of the family of transformation functions is

$$\phi_\beta(X_i) = \begin{cases} log(X_i - min(X_1, X_2, \cdots, X_n) + |\frac{1}{\beta}|R), & \beta > 0 \\ -log(max(X_1, X_2, \cdots, X_n) - X_i + |\frac{1}{\beta}|R), & \beta < 0 \end{cases} \tag{4}$$

in which $\beta \in \mathbb{R}$, R and $g(X)$ are as defined above.

2.2 Standardization and Winsorisation

Standardization is applied to the transformed feature vector i.e.$[\phi_\beta(X_1), \cdots, \phi_\beta(X_n)]$, by subtracting its median and dividing by the mean absolute deviation from the median[1]. Denote the vector after standardization as $X^\dagger$. A winsorisation of $X^\dagger$ at an appropriate threshold is further applied to reduce the influence of extreme observations.

2.2.1 Winsorisation

Extreme value theory provides reasonable choices of thresholds for winsorisation. A fundamental result of that area is the *Three Types Theorem*. See Leadbetter et al. (2011) for detailed discussion.

Theorem 1 (The Extremal Types Theorem) *Suppose* $X = (X_1, X_2, \cdots, X_n)$ *are independent, identically distributed standard normal random variables, there exist real constants* $a_n > 0$ *and* b_n *such that*

$$P(\frac{M_n - b_n}{a_n} \leq x) \rightarrow G(x) \tag{5}$$

where $G(x)$ *is the cumulative distribution function of the standard Gumbel distribution i.e.* $G(x) = e^{-e^{-x}}$ *and*

$$b_n = (2\log n)^{\frac{1}{2}} - \frac{\log\log n + \log(4\pi)}{(2\log n)^{\frac{1}{2}}} \tag{6}$$

---

[1]If the mean absolute deviation is zero, return a vector of zeros.

$$a_n = (2 \log n)^{-\frac{1}{2}} \tag{7}$$

From this extreme value theory, the threshold of the standardized vector $X^\dagger$ is computed based on the 95th percentile of the standard Gumbel distribution (p_{95}), that is

$$L = p_{95} a_n + b_n. \tag{8}$$

When the absolute value of the element in $X^\dagger$ is greater than L i.e. $|X_i^\dagger| > L$, the element value is winsorized (i.e pulled back) to the value $sign(X_i^\dagger)L$. After the winsorisation, the feature vector will be standardized again, by substracting the sample mean and dividing by the standard deviation.

2.3 Evaluation

The evaluation of the stated transformation procedure is based on measuring the distance between the empirical distribution function (EDF) of the transformed data and the cumulative distribution function (CDF) of the standard normal. Commonly used EDF statistics are the Kolmogorov-Smirnov test statistic, the Cramér-von Mises test statistic, the Watson statistic and the Anderson Darling test statistic. Stephens (1974) conducted power studies of these statistics under different specifications of hypothesized distributions. Based on this study, the Anderson Darling test statistic is considered as powerful for detecting most common departures from normality. Therefore, that is used here as the criterion for evaluation. A computable form of the Anderson Darling test statistic is defined in terms of the order statistics i.e.

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} [\log \Phi(X_{(i)}) + \log(1 - \Phi(X_{(n+1-i)}))] \tag{9}$$

Larger values of this indicate stronger departures from Gaussianity. Thus, by searching for a parameter value minimizing this statistic, an optimal transformation for improving the closeness of the distributions of features to normality is obtained.

References

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

Raymond J Carroll. A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 71–78, 1980.

Aaron K Han. A non-parametric analysis of transformations. *Journal of Econometrics*, 35(2): 191–209, 1987.

Fabian Hernandez and Richard A Johnson. The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, 75(372):855–861, 1980.

R. Leadbetter, G. Lindgren, and H. Rootzen. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer New York, 2011. ISBN 9781461254515.

Jayson Miedema, James Stephen Marron, Marc Niethammer, David Borland, John Woosley, Jason Coposky, Susan Wei, Howard Reisner, and Nancy E Thomas. Image and statistical analysis of melanocytic histology. *Histopathology*, 61(3):436–444, 2012.

RM Sakia. The box-cox transformation technique: a review. *The statistician*, pages 169–178, 1992.

Michael A Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737, 1974.

Willem Rutger van Zwet. *Convex transformations of random variables*, volume 7. Mathematisch centrum, 1964.

Matt P Wand and M Chris Jones. *Kernel smoothing*. Crc Press, 1994.

L. Yang. *Transformation-Density Estimation*. PhD thesis, University of North Carolina, Chapel Hill, 1995.