ACE, AVAS and Robust Data Transformations

Anthony C. Atkinson, Marco Riani, Aldo Corbellini, and Gianluca Morelli

Abstract Unlike the Box-Cox transformation, that of Yeo and Johnson for the response of a linear model can be applied when the observations are not constrained to be positive. We study the extended Yeo–Johnson transformation in which positive and negative observations can be transformed with different parameter values. The procedure is illustrated for data with many outliers. The data are cleaned with a robust method, the forward search, and the obtained transformations compared with the results from two nonparametric transformation methods based on data smoothing.

Keywords Box-Cox transformation · Extended Yeo–Johnson transformation · Fan plot · Forward search · Nonparametric transformation · Smoothing

1 Introduction

The widely used parametric family of power transformations introduced by [3] is only applicable to positive observations. Yeo and Johnson [7] spliced together two Box-Cox transformations to provide a one-parameter family of transformations for data that can be positive or negative. Atkinson et al. [2] extended the Yeo–Johnson

A. C. Atkinson
Department of Statistics, London School of Economics, London, UK
e-mail: a.c.atkinson@lse.ac.uk

M. Riani (✉) · A. Corbellini · G. Morelli
Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma, Parma, Italy
e-mail: mriani@unipr.it

A. Corbellini
e-mail: aldo.corbellini@unipr.it

G. Morelli
e-mail: gianluca.morelli@unipr.it

S. Balzano et al. (eds.), *Statistical Learning and Modeling in Data Analysis*,
Studies in Classification, Data Analysis, and Knowledge Organization,
https://doi.org/10.1007/978-3-030-69944-4_2

transformation to allow different parameter values for the transformation of positive and negative observations; they illustrate the usefulness of this procedure through the analysis of two sets of data.

Nonparametric transformations provide an alternative to such families of parametric transformations. The purpose of this short paper is to compare parametric and nonparametric transformations for a set of data that contain outliers, to illustrate a robust method for cleaning the data of outliers and to compare transformations on the cleaned data.

2 Extended Parametric Transformations

The purpose of these parametric transformations is to achieve a response which is approximately normally distributed with errors of constant variance and a linear model of simple form. For comparisons of estimates of parameters for different values of λ, many authors, starting with [3], stress the importance of working with a normalized transformation allowing for the change of scale of the observations with transformation. For the Box-Cox transformation, the normalized transformation is

$$z(\lambda) = (y^{\lambda} - 1)(z^{\lambda} - 1)/(\lambda \dot{y}^{\lambda-1}) \ (\lambda \neq 0); \ \dot{y} \log y \ (\lambda = 0), \tag{1}$$

where $\dot{y}$ is the geometric mean of y and J, the Jacobian of the transformation is given by $\log J = n(\lambda - 1) \log \dot{y}$. The linear model to be fitted is $z(\lambda) = X\beta(\lambda) + \varepsilon$, where X is $n \times p$, β is a $p \times 1$ vector of unknown parameters and the variance of ε is σ^2.

The normalized transformation for the two-parameter extended Yeo–Johnson (EYJ) transformation is given by [2]. There are now four regions of y, rather than two, with distinct forms of $z(\lambda)$ and the Jacobian is now a more complicated function of the observations.

3 Robustness and the Fan Plot

We use a robust procedure, the Forward Search [1] to order the data by closeness to the fitted model. The procedure starts from a carefully chosen subset of $m_0 = p + 1$ observations and moves forward increasing the subset size m by introducing the observation, not used in fitting, that is closest to the fitted model, until all observations have been fitted. Outliers, if any, enter at the end of the search.

Outliers in one value of λ may not be so for some other values. We, therefore, need to repeat the forward search for a grid of values of λ. For each resultant ordering of the data, we monitor evidence for the correctness of the transformation as m increases. We include the constructed variable $w(\lambda) = \partial z(\lambda)/\partial\lambda$ in the linear model for the EYJ transformation. The approximate score statistic for the value λ_0 is the t-test

for the significance of $w(\lambda_0)$ in the regression. The plot of trajectories of the score statistic against subset size for a set of values of λ is called a fan plot.

Constructed variables for the one-parameter Yeo–Johnson transformation are given by [2]. They further derive constructed variables for testing whether positive and negative observations require the same transformation. These come from the extended transformation in which one kind of response has the parameter $\lambda + \alpha$ and the other λ. The test is for $\alpha = 0$.

4 Augmented Investment Fund Data

As our example, we analyze data on the relationship between the medium term performance of 309 investment funds and two indicators. Of these funds, 99 have negative performance. To examine the properties of transformation procedures in the presence of outliers, we augmented the data with 40 outliers, to produce a data set in which the outliers are evident after transformation, but not before. The analysis of the uncontaminated data [2] concludes that the negative observations need transformation with parameter $\lambda_N = 0$, which for the EYJ is not the log transformation. The positive observations need no transformation ($\lambda_P = 1$). The data are well behaved, with no evidence of any outliers.

The fan plot for the augmented data indicates that the majority of the outliers enters the subset at the end of the search; the structure of the plot changes for $m > 310$. The extended fan plot with separate trajectories for positive and negative observations shows that different transformations are required for the two parts of the data. The best values are $\lambda_N = 0$ and $\lambda_P = 1$ when the trajectories of the score statistics for positive and negative observations are similar to those for the overall data until m is around 320. This is the transformation found for the uncontaminated data.

5 Robust Analysis

We now identify the outlying observations by a forward search analysis of the data with the recommended transformation $\lambda_P = 1$ and $\lambda_N = 0$.

The left-hand panel of Fig. 1 shows a forward plot of all 349 scaled residuals of the augmented data for a wide range of values of m. There is an upper band of residuals, in blue in the online version of the paper, separated from a lower band of 37 residuals, shown in red. What is remarkable is the stability of this pattern, until $m = 310$, indicative of a set of data without outliers and with normally distributed errors and the second group of observations, not included in the subsets used for fitting.

The highlighted, red, residuals were identified by brushing the plot. That is, we selected all the trajectories that lie within the brush in the centre of the figure. The right-hand panel of the figure shows a linked forward plot of minimum Mahalanobis

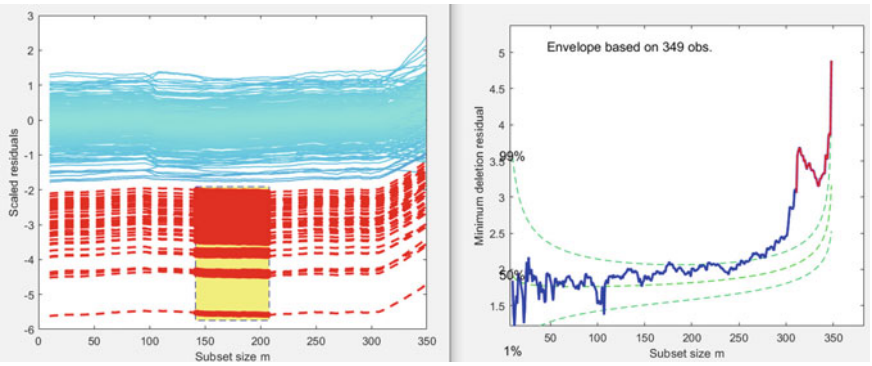

Fig. 1 Augmented investment fund data; brushing linked plots from the forward search when $\lambda_P = 1$ and $\lambda_N = 0$. Left-hand panel, trajectories of residuals from the forward search with the residuals from 37 observations highlighted by brushing. Right-hand panel, linked forward plot of minimum deletion residuals during the search with the 37 brushed values shown in red

distances, with the trajectory of the 37 brushed observations shown in red. These are indeed the last observations to enter the search. Our automatic procedure for outlier detection [5] in fact identifies 35 outliers.

6 Nonparametric Transformations

It is clear from the results of the previous sections that the contaminated data need both cleaning and transformation. The purpose of this section is to determine what information nonparametric transformations provide on the presence of outliers, the transformation of the data and whether the parametric extended Yeo–Johnson transformation can be improved by further transformation. The parametric transformations produce a smooth relationship between $z(\lambda)$ and the original y. A nonparametric alternative is to use smoothing to estimate this relationship. We use two such methods, ACE—Alternating Conditional Expectations—[4] and AVAS [6] in which the transformation for the response is intended to yield additivity and variance stabilization. We consider only response transformation, comparing models through the value of unadjusted R^2.

We start with the extended Yeo–Johnson transformation, using the parameter estimates $\lambda_P = 1$ and $\lambda_N = 0$ for all comparisons. First we look at the contaminated data before and after cleaning. The left-hand panel of Fig. 2 shows the QQ plot of the residuals of all 349 observations from regression with the original response. The sigmoid shape of this plot indicates that the observations are not normally distributed. The right-hand panel is the QQ plot for residuals of the transformed cleaned data. The distribution of residuals is much closer to normality, although the centre of the curve indicates that many small residuals are slightly too large in absolute value.

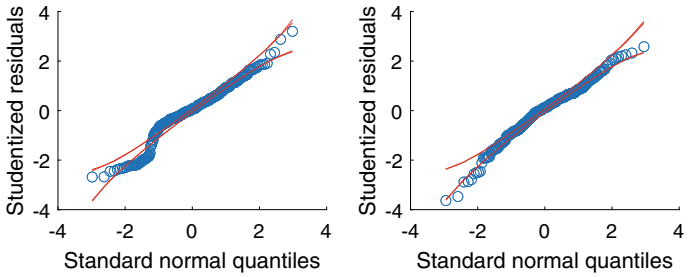

Fig. 2 Comparison of normal QQ plots of residuals. Left-hand panel, untransformed contaminated data. Right-hand panel, transformed cleaned data

Table 1 Investment fund data: summary properties of regression for parametric and nonparametric transformations of contaminated and cleaned data

	Contaminated	Cleaned and transformed
Untransformed	0.399	–
EYJ	0.356	0.783
AVAS	0.241	0.778
ACE	0.421	0.806
ACE (monotonic)	0.417	0.805

The value of R^2 for regression on the untransformed contaminated data is 0.399. For the cleaned transformed data it is 0.783 and for the uncontaminated data 0.816. The left-hand column of Table 1 lists the values of R^2 achieved by regression on parametric and nonparametric transformations of the contaminated data. The largest value is 0.421 for unconstrained ACE. The monotonicity constraint on ACE comes from isotonic regression on the unconstrained transformation and yields a slightly reduced value of 0.417. AVAS produces a value of 0.241, less than that for EYJ. Figure 3 provides plots of transformed against untransformed response for these four transformations.

The top left-hand panel of the figure shows that the EYJ transformation for $y > 0$ is linear (no transformation), whereas for negative y, the transformation is concave, transforming the more negative observations to be more extreme. AVAS, in the top right-hand panel, provides a more smooth concave curve, which not only makes the more negative values more extreme but makes the more positive values less extreme. Unconstrained ACE is virtually linear for $y > 12$, but shrinks in the most negative observations, some of which are outliers. Constrained ACE is formed by isotonic regression on the unconstrained version, and as the figure shows, is similar in structure to ACE. Both transformations show several points of inflection for $y < 12$, especially just above zero.

If the errors are approximately normally distributed and the model is correct, the plot of residuals against fitted values should be without any features, apart from those

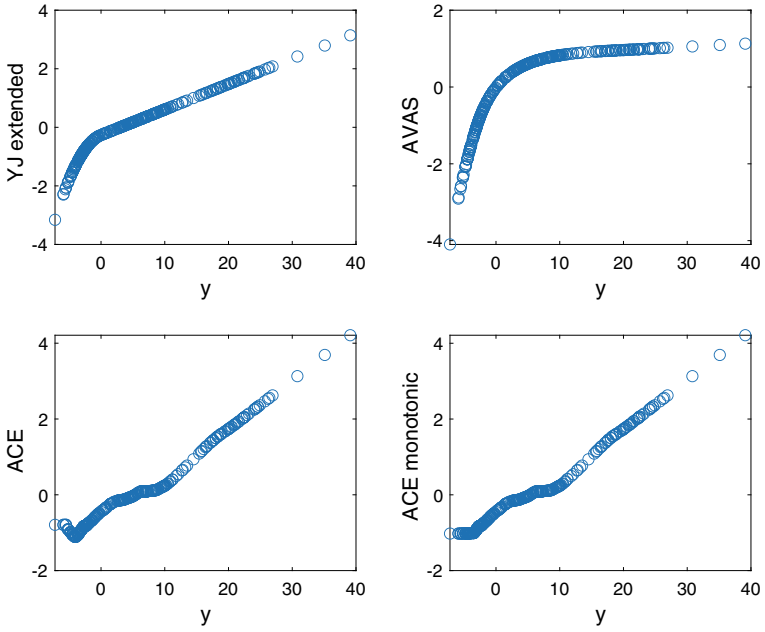

Fig. 3 Contaminated data: transformed responses against untransformed responses. Top row, EYJ and AVAS. Bottom row, ACE, constrained and unconstrained

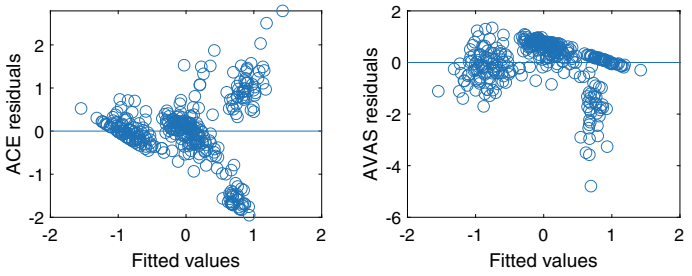

Fig. 4 Contaminated data; residuals against fitted values. Left-hand panel, constrained ACE. Right-hand panel AVAS (note the scale of these residuals)

from the distribution of fitted values. The left-hand panel of Fig. 4 shows such a plot of residuals from constrained ACE. The plot is wedge shaped, with a sharp lower diagonal bound. The other panel, for AVAS, also has some structure, in this case, a cloud of large negative residuals for fitted values around 0.5; the nonparametric transformations indicate faults in the model or data.

We now look at the transformation of the cleaned data after it has been subjected to the extended Yeo–Johnson transformation to check whether the properties can be improved by a further nonparametric transformation. Values of R^2 for such transformations are in the right-hand column of Table 1. The value for EYJ is 0.783.

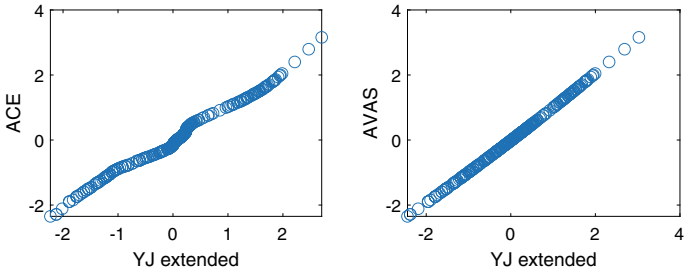

Fig. 5 Nonparametric transformations of cleaned transformed data against EYJ. Left-hand panel, constrained ACE, right-hand panel, AVAS

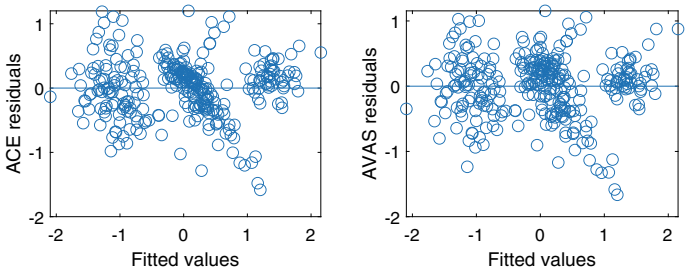

Fig. 6 Residuals against fitted values from nonparametric transformations of cleaned transformed data. Left-hand panel, constrained ACE, right-hand panel, AVAS

AVAS is slightly less than this at 0.778, whereas constrained ACE is 0.805, with the unconstrained version giving a value of 0.806.

The left-hand panel of Fig. 5 shows the plot of transformed y from constrained ACE against the values from EYJ. Some of the points of inflection shown in Fig. 3 remain and correspond to original values of y that were just positive. The indication is that the two-parameter EYJ transformation with one transition point can be improved by using a distinct transformation for the observations just above zero, leading to a slight increase in R^2. The right-hand panel for AVAS shows a virtually straight line and the transformation is very close to that for EYJ.

The plots in Fig. 6 are of residuals against fitted values for the two transformations featured in Fig. 5; both indicate the presence of three groups of funds, which are also surprisingly well transformed by the two-parameter EYJ procedure. Although the two plots are similar, some of the details of the central group are different, which is where the two transformations diverge. The QQ plots for the nonparametric transformations are close to that for EYJ shown in Fig. 2.

The results of this section indicate that the nonparametric transformations do not provide a robust procedure. But they can provide insight when used to check a suggested parametric transformation. For the EYJ it is possible that the two transformation regions may not separate at zero, but at some value to be determined. A second aspect is whether two regions of transformation are enough. It may be that

for some data structures the flexibility of the nonparametric transformation will lead to improved data modelling.

Acknowledgements This research benefits from the HPC (High Performance Computing) facility of the University of Parma. We acknowledge financial support from the University of Parma project "Statistics for fraud detection, with applications to trade data and financial statements" and from the Department of Statistics, London School of Economics.

References

1. Atkinson, A.C., Riani, M., Cerioli, A.: The Forward Search: theory and data analysis (with discussion). J. Korean Stat. Soc. **39**, 117–134 (2010). https://doi.org/10.1016/j.jkss.2010.02.007
2. Atkinson, A.C., Riani, M., Corbellini, A.: The analysis of transformations for profit-and-loss data. J. R. Stat. Soc. C **69**, 251–275 (2020)
3. Box, G.E.P., Cox, D.R.: An analysis of transformations (with discussion). J. R. Stat. Soc., Ser. B **26**, 211–246 (1964)
4. Breiman, L., Friedman, J.H.: Estimating optimal transformations for multiple regression and transformation (with discussion). J. Am. Stat. Assoc. **80**, 580–619 (1985)
5. Riani, M., Atkinson, A.C., Cerioli, A.: Finding an unknown number of multivariate outliers. J. R. Stat. Soc., Ser. B **71**, 447–466 (2009)
6. Tibshirani, R.: Estimating transformations for regression via additivity and variance stabilization. J. Am. Stat. Assoc. **83**, 394–405 (1988)
7. Yeo, I.-K., Johnson, R.A.: A new family of power transformations to improve normality or symmetry. Biometrika **87**, 954–959 (2000)