The Analysis of Transformations for Prof t and Loss Data

Anthony C. Atkinson
The London School of Economics, London WC2A 2AE, UK,*
Marco Riani† and Aldo Corbellini‡, Dipartimento di Scienze Economiche e Aziendale and Interdepartmental Centre for Robust Statistics, Università di Parma,
43100 Parma, Italy

November 1, 2019

Abstract

We analyse data on the performance of investment funds, 99 out of 309 of which report a loss and on the prof tability of 1,405 f rms, 407 of which report losses. The problem in both cases is to use regression to predict performance from sets of explanatory variables. In one case, it is clear from scatterplots of the data that the negative responses have a lower variance than the positive ones and a different relationship with the explanatory variables. Because the data include negative responses, the Box-Cox transformation cannot be used. We develop a robust version of an extension to the Yeo-Johnson transformation which allows different transformations for positive and negative responses. Tests and graphical methods from our robust analysis allow the detection of outliers, the assessment of values of the two transformation parameters and the building of simple regression models. Performance comparisons are made with non-parametric transformations.

Keywords:

ACE; AVAS; Box-Cox transformation; constructed variable; fan plot; forward search; linked plots; robust methods; Yeo-Johnson transformation.

*e-mail: a.c.atkinson@lse.ac.uk
†e-mail: mriani@unipr.it
‡e-mail: aldo.corbellini@unipr.it

1 Introduction

In practice data are often transformed to approximate normality, leading to straightforward analyses and insightful conclusions. The first four pages of Box and Cox (1964) provide a thoughtful analysis of these advantages, to which may now be added the universal availability of flexible software for the analysis of normally distributed data with a variety of complex structures including time series and multivariate data. However, the widely-used and effective parametric family of power transformations introduced by Box and Cox for regression can only be applied to positive responses. Yeo and Johnson (2000) generalised this transformation to allow for the inclusion of zero and negative response values, which arise for example in economics (we analyse two examples in which losses have been made as well as profits) and in the analysis of difference data. However, like Box and Cox, Yeo and Johnson use likelihood methods based on aggregate statistics; their method is neither diagnostic nor robust. In this paper we:

- provide extensive analyses of data for which the Yeo-Johnson transformation is appropriate;
- detect the effect of individual observations on the estimated transformation parameter by the use of robust methods combined with insights on the correct transformation given by the "fan" plot;
- extend the methods to cases when there are different transformation parameters for positive and negative responses and provide a new test for the correctness of the parameter estimates, together with information on the distribution of the test statistic;
- use the test and novel graphical procedures to determine appropriate robust transformations of positive and negative responses when, as is the case in our examples, they are markedly different in the two response classes;
- provide brief comparisons with two nonparametric methods for response transformation.

We focus on the transformation of the response in linear models. Data on investment funds, some of which have negative returns, are introduced in §2 and the problem in data analysis presented. In §3 we give references to diagnostic and robust procedures for the Box-Cox transformation and define its extension, the Yeo-Johnson transformation. We introduce the normalized form of the transformation which includes the Jacobian, providing a convenient form for the analysis of data. In §4 we introduce an approximate score test for the value of the transformation parameter based on constructed variables from a Taylor series expansion

of the model. We thus avoid numerical maximization for estimation and tests of the transformation parameter λ. Constructed variables for the Yeo-Johnson transformation are derived in §5. As our two data examples show, it is not always true that positive and negative observations should be transformed with the same value of λ. A model for the extended Yeo-Johnson transformation for such data is presented in §6 and the constructed variables derived.

The resulting score tests are functions of aggregate statistics and so do not provide information on the effect of individual observations on the estimate of the transformation parameter. For this we employ a robust procedure, the forward search (Atkinson *et al.*, 2010), to order the observations from those closest to the model for the transformed data to those furthest from it. We thus obtain subsets of increasing sizes of the least outlying observations. The fan plot of §7 plots the score statistics from the search against subset size, indicating observations that do not agree with the proposed transformation. The procedure is extended to three tests: for the overall transformation and those for positive and negative observations, yielding an extended fan plot. Once the data have been transformed, the extended fan plot provides a robust method for testing the correctness of the transformation. Simulations show that the null distribution of these statistics can be very close to that of Student's t.

In §8 we give an extended analysis of the data on investment fund performance. Our method finds an overall transformation to normality. But the extended fan plot shows that the positive and negative observations should have different transformations, which we find. There is no evidence of any outlying observations. We compare these results with those from two nonparametric transformations: ACE (Breiman and Friedman, 1985) and AVAS (Tibshirani, 1988), which are based on smoothing and do not require the data to be positive. In the following §9 we analyse a more complicated set of data taken from balance sheets with 1,405 observations. Again, positive and negative observations require different transformations but now 42 outlying observations are detected. Despite deletion of these observations, the F statistic for the regression, given in Table 3, shows an increase equivalent to doubling of the sample size. Section 9 also illustrates the use of linked plots from the robust analyses of the data, which leads to exclusion of the outliers and the estimation of uncorrupted parameter estimates. Some remarks on the null distributions of the test statistics are in §10 with closing comments in §11. The two sets of data and Matlab programs for their analysis are available online.

2 Investment Funds

The purpose is to relate the medium term performance of 309 investment funds to two indicators. Of these funds 99 have negative performance. The data come from the Italian financial newspaper *Il Sole - 24 Ore* and the variables are:

y medium term (36 month) performance;
x_1 short term (12 month) performance;
x_2 short term volatility.

Performance is measured as the percentage change in the price of units of the fund.

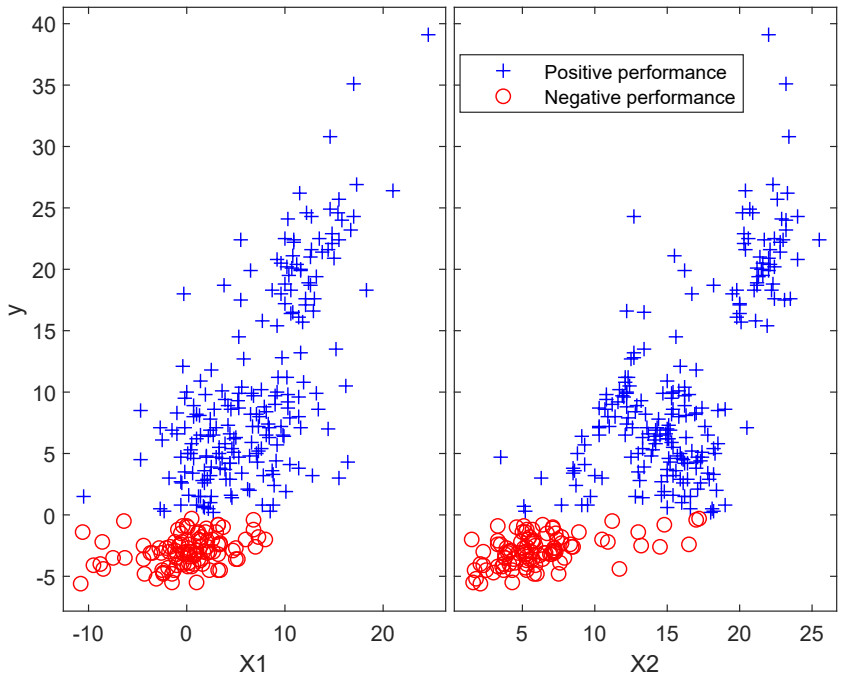

Figure 1: Investment fund data: scatterplots of y against x_1 and x_2

Scatterplots of y against the two explanatory variables are in Figure 1, with the negative responses shown as circles. It is clear that there is a strong, roughly linear, relationship between the response and both explanatory variables. It is also clear that the negative responses have a different behaviour from the positive ones: the variance is less and the slope of the relationship with both explanatory variables appears to be smaller. We employ an extended version of the Yeo-Johnson transformation to see whether we can achieve a transformation which satisfies the three requirements of homogeneity, additivity and approximate normality of

errors discussed in §3.

3 The Yeo-Johnson Transformation

In §8 we analyse the investment fund data using the Yeo-Johnson transformation, derived from the Box-Cox transformation for non-negative responses for which the transformed response is:

$$y_{\text{BC}}(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda & (\lambda \neq 0) \\ \log y & (\lambda = 0). \end{cases} \tag{1}$$

Box and Cox and Hinkley and Runger (1984) show that the interpretation of data analyses is improved by use of a normalized transformation that incorporates the Jacobian of the transformation, $J = \prod_{i=1}^{n} y_i^{\lambda-1} = \dot{y}_{\text{BC}}^{n(\lambda-1)}$, where $\dot{y}_{\text{BC}}$ is the geometric mean of the observations. The normalized transformation is $z(\lambda) = y(\lambda)/\dot{y}_{\text{BC}}^{(\lambda-1)}$, with $z(0) = \dot{y}_{\text{BC}} \log y$. The physical dimension of $z(\lambda)$ is that of y so that sums of squares of $z(\lambda)$ can be directly compared. Andrews *et al.* (1971) and Gnanadesikan (1977) extended the Box-Cox transformation to the analysis of multivariate data.

Box and Cox (1964) rely on complete-sample likelihood inference. Methods for the detection of outliers and the assessment of the inf uence of individual or groups of observations on the estimated transformation parameter are described by Cook and Weisberg (1999) and Atkinson and Riani (2000).

Although Box and Cox work with a normal theory likelihood, the aim of the transformation is to produce responses which have homogeneous variance, simple additive models and an approximately normal distribution of errors. All three aims are satisf ed in the examples given by Box and Cox (1964), as they are in the analyses of numerous other data sets, such as those in Atkinson and Riani (2000, Chapter 4). The aims of the Yeo-Johnson transformation are similar.

Box and Cox stress that they do not recommend transformation of the data by the maximum likelihood estimate $\hat{\lambda}$. Rather, they recommend the use of a value that lies within the conf dence region for λ while belonging to a grid of values with physical interpretability. In their two examples these are the log transformation $(\lambda = 0)$ and the reciprocal transformation $(\lambda = -1)$. The consequences for statistical inference are investigated by Bickel and Doksum (1981), Carroll (1982), Box and Cox (1982), Hinkley and Runger (1984) and by Chen *et al.* (2002) and discussants.

The Box-Cox transformation has two regimes, that for $\lambda \neq 0$ and that for $\lambda = 0$. In both cases, $y > 0$. To allow for y being either positive or negative, the Yeo-Johnson transformation of the response y requires four regimes, depending

both on the transformation parameter and on the response value. The forms of the transformed response are

$$y_{\text{YJ}}(\lambda) = \begin{cases} \dfrac{(y+1)^\lambda - 1}{\lambda} & y \geq 0 \quad \lambda \neq 0 \\ \log(y+1) & y \geq 0 \quad \lambda = 0 \\ -\dfrac{\{(-y+1)^{2-\lambda} - 1\}}{2-\lambda} & y < 0 \quad \lambda \neq 2 \\ -\log(-y+1) & y < 0 \quad \lambda = 2. \end{cases} \tag{2}$$

For $y \geq 0$ this is the generalized Box-Cox power transformation of $y + 1$. For negative y the transformation is of $-y + 1$ to the power $2 - \lambda$.

Analysis of data from this transformation also needs to include the Jacobian of the transformation to allow for changes of scale as λ varies. The required Jacobian, again for n observations, from equation (3.1) of Yeo and Johnson (2000), is

$$\log J_{\text{YJ}} = (\lambda - 1) \sum_{i=1}^{n} \text{sgn}(y_i) \log(|y_i| + 1). \tag{3}$$

We continue to work with a normalized transformation $z(\lambda) = y(\lambda)/J^{1/n}$ in which the Jacobian is spread over all observations. If $\dot{y}_{\text{YJ}}$ is the nth root of J_{YJ} in (3),

$$\dot{y}_{\text{YJ}} = \exp\left[\sum \{\text{sgn}(y_i) \log(|y_i| + 1)\}/n\right]. \tag{4}$$

The normalised versions of the transformations in (2) then become

$$z_{\text{YJ}}(\lambda) = \begin{cases} \dfrac{(y+1)^\lambda - 1}{\lambda \dot{y}_{\text{YJ}}^{\lambda-1}} & y \geq 0 \quad \lambda \neq 0 \\ \dot{y}_{\text{YJ}} \log(y+1) & y \geq 0 \quad \lambda = 0 \\ -\dfrac{\{(-y+1)^{2-\lambda} - 1\}}{(2-\lambda)\dot{y}_{\text{YJ}}^{\lambda-1}} & y < 0 \quad \lambda \neq 2 \\ -\log(-y+1)/\dot{y}_{\text{YJ}} & y < 0 \quad \lambda = 2. \end{cases} \tag{5}$$

4 An Approximate Score Test for the Transformation Parameter

For the linear regression model $z(\lambda) = x^T\beta + \epsilon$, where x is $p \times 1$ and the errors are independently normally distributed with variances σ^2, let $R(\lambda)$ be the residual

sum of squares of the $z(\lambda)$. Then, ignoring a constant, the prof le loglikelihood of the observations, maximized over β and σ^2, is

$$L_{\max}(\lambda) = -(n/2)\log\{R(\lambda)/n\}.$$

maximized by the value $\hat{\lambda}$ that minimizes $R(\lambda)$. To test the hypothesis that $\lambda = \lambda_0$, Box and Cox use the likelihood ratio test $n\log\{R(\lambda_0)/R(\hat{\lambda})\}$ which has an asymptotic chi-squared distribution. An alternative to the likelihood ratio test for the value of λ is the approximate score test (Atkinson, 1973) derived from a Taylor series expansion of $z(\lambda)$ around λ_0 as

$$z(\lambda) \doteq z(\lambda_0) + (\lambda - \lambda_0)w(\lambda_0), \tag{6}$$

where

$$w(\lambda_0) = \left.\frac{\partial z(\lambda)}{\partial \lambda}\right|_{\lambda=\lambda_0}.$$

The regression model can then be approximated as

$$z(\lambda) = x^T\beta - (\lambda - \lambda_0)w(\lambda_0) + \epsilon = x^T\beta + \gamma\, w(\lambda_0) + \epsilon. \tag{7}$$

The new variable $w(\lambda_0)$ is the constructed variable for the transformation. The approximate score statistic for testing the transformation $\lambda = \lambda_0$ is the t statistic for regression on $w(\lambda_0)$ in (7), that is the test for $\gamma = 0$ in the presence of all components of x.

An explicit form for the score test can be found by writing the extended model in matrix form as

$$E(Z) = X\beta + w\gamma,$$

where Z is $n \times 1$, X is $n \times p$ and γ is a scalar. The least squares estimate $\hat{\gamma}$ is found explicitly from the normal equations for this partitioned model using the algebra for added variables (Atkinson and Riani, 2000, §2.2). With $H = X(X^TX)^{-1}X^T$ (the hat matrix) and $A = I - H$

$$\hat{\gamma} = \frac{w^T(I-H)y}{w^T(I-H)w} = \frac{w^TAz}{w^TAw}, \tag{8}$$

where w is the $n \times 1$ vector of constructed variables.

Calculation of the statistic also requires s_w^2, the residual mean square estimate of σ^2 from regression of z on X and w,

$$(n-p-1)s_w^2 = z^TAz - (z^TAw)^2/(w^TAw).$$

The t statistic for testing that $\gamma = 0$ is then

$$t_\gamma = \frac{\hat{\gamma}}{\sqrt{\{s_w^2/(w^TAw)\}}},$$

on $n - p - 1$ degrees of freedom.

5 Constructed Variables

For the Box-Cox transformation there are two constructed variables, one for general λ and one for $\lambda = 0$, found by series expansion of the two forms of $z_{\text{BC}}(\lambda)$ coming from the two for $y_{\text{BC}}(\lambda)$ in (1), expressions for which are given by Atkinson and Riani (2000, p. 86). For the Yeo-Johnson transformation there is correspondingly a constructed variable for each of the four normalized transformations in (5). For $y \geq 0$ let $v_P = y + 1$ with $v_N = -y + 1$ when $y < 0$. The constructed variables are

$$w_{\text{YJ}}(\lambda) = \begin{cases} \{v_P^{\lambda}(\log v_P - k_P) + k_P\}/q_P, & y \geq 0 \quad \lambda \neq 0 \\ \dot{y}_{\text{YJ}} \log v_P(\log v_P/2 - \log \dot{y}_{\text{YJ}}) & y \geq 0 \quad \lambda = 0 \\ \{v_N^{2-\lambda}(\log v_N + k_N) - k_N\}/q_N & y < 0 \quad \lambda \neq 2 \\ \{\log v_N(\log v_N/2 + \log \dot{y})\}/\dot{y}_{\text{YJ}} & y < 0 \quad \lambda = 2. \end{cases} \tag{9}$$

In (9)

$$\begin{aligned} k_P &= \lambda^{-1} + \log \dot{y}_{\text{YJ}} \\ q_P &= \lambda \dot{y}_{\text{YJ}}^{\lambda-1} \\ k_N &= \log \dot{y}_{\text{YJ}} - (2-\lambda)^{-1} \\ q_N &= (2-\lambda)\dot{y}_{\text{YJ}}^{\lambda-1}. \end{aligned}$$

6 Homogeneity of Transformation

6.1 The Extended Yeo-Johnson Transformation

The transformations for negative and positive responses were determined by Yeo and Johnson (2000) by imposing the smoothness condition that the second derivative of $z_{\text{YJ}}(\lambda)$ with respect to y be smooth at $y = 0$. However some authors, for example Weisberg (2005), query the physical interpretability of this constraint which is indeed violated by the sets of data analysed in §§8 and 9. Accordingly, we extend the Yeo-Johnson transformation to allow two values of the transformations parameter: λ_N for negative observations and λ_P for non-negative ones.

To start we need the Jacobian of the transformation, but we now have separate calculations for positive and negative y. The Jacobian is def ned for all observations in (4) above. Let

$$S_N = \sum_{y_i<0} -\log(-y_i + 1) = \sum_{y_i<0} -\log v_{i,N} \quad \text{and} \quad \dot{y}_N = \exp(S_N/n). \tag{10}$$

Note division by n, not n_N (the number of negative y_i) as the Jacobian is spread over all observations.

Likewise, for the non-negative observations

$$S_P = \sum_{y_i \geq 0} \log(y_i + 1) = \sum_{y_i \geq 0} \log v_{i,P} \quad \text{and} \quad \dot{y}_P = \exp(S_P/n). \tag{11}$$

With these def nitions the normalised form of the extended Yeo-Johnson transformation is

$$z_{\text{EYJ}}(\lambda_N, \lambda_P) = \begin{cases} \dfrac{v_p^{\lambda_P} - 1}{\lambda_P \dot{y}_N^{\lambda_N - 1} \dot{y}_P^{\lambda_P - 1}} & y \geq 0 \quad \lambda_P \neq 0 \\ (\dot{y}_P / \dot{y}_N^{\lambda_N - 1}) \log v_P & y \geq 0 \quad \lambda_P = 0 \\ -\dfrac{v_N^{2-\lambda_N} - 1}{(2 - \lambda_N) \dot{y}_N^{\lambda_N - 1} \dot{y}_P^{\lambda_P - 1}} & y < 0 \quad \lambda_N \neq 2 \\ -\log v_N / \dot{y}_N \dot{y}_P^{\lambda_P - 1} & y < 0 \quad \lambda_N = 2. \end{cases} \tag{12}$$

Since $\dot{y} = \dot{y}_N \dot{y}_P$ this extended transformation reduces to the standard Yeo-Johnson transformation when $\lambda_N = \lambda_P$.

Determining whether a specif c extended Yeo-Johnson transformation is appropriate for a set of data requires testing whether some specif ed value λ_{N0} is appropriate for the negative observations and some λ_{P0} for the non-negative ones. We give below the constructed variables for testing the value of one transformation parameter, with the other held f xed. With data reasonably balanced over positive and negative responses both tests may be informative. We start with the variables for λ_P.

6.2 Test Transformation of Positive y

With an extension of the notation of §5 the constructed variables are

$$w_{\text{EYJ}}(\lambda_P) = \begin{cases} \{v_P^{\lambda_P}(\log v_P - k_P^*) + k_P^*\}/q_P^* & y \geq 0 \quad \lambda_P \neq 0 \\ \dot{y}_P \log v_P \{(\log v_P)/2 - \log y_P\} & y \geq 0 \quad \lambda = 0 \\ z_{\text{EYJ}}(\lambda_N, \lambda_P) \log \dot{y}_P & y < 0, \end{cases} \tag{13}$$

where $k_P^* = \lambda_P^{-1} + \log \dot{y}_P$ and $q_P^* = \lambda_P \dot{y}_P^{\lambda_P - 1}$. The structure is similar to that of the constructed variables in §5. The result for $y < 0$ arises because the transformation for $y < 0$ only depends on λ_P through the Jacobian.

6.3 Test Transformation of Negative y

$$w_{\text{EYJ}}(\lambda_N) = \begin{cases} -z_{\text{EYJ}}(\lambda_N, \lambda_P) \log \dot{y}_N & y \geq 0 \\ \{v_N^{2-\lambda_N}(\log v_N + k_N^*) - k_N^*\}/q_N^* & y < 0 \quad \lambda_N \neq 2 \\ \log v_N\{(\log v_N)/2 + \log y_N\}/y_N & y < 0 \quad \lambda_N = 2, \end{cases} \tag{14}$$

where $k_N^* = \lambda_N^{-1} + \log \dot{y}_N$ and $q_N^* = \lambda_N \dot{y}_N^{\lambda_N - 1}$.

6.4 Establishing Two Transformations

Finding robust values of the two transformation parameters is in two stages. We first use the score test derived from the constructed variables (9) for the Yeo-Johnson transformation to find an overall estimate of λ. Score tests based on the constructed variables for positive and negative observations (13) and (14), with a common value λ_0 for λ_N and λ_P, allow determination of whether separate values of λ_N and λ_P are required. Once the data have been satisfactorily transformed using (12), no further transformation is required. As a consequence, the score test for the standard Yeo-Johnson transformation with the null hypothesis that $\lambda_0 = 1$ will not lead to rejection. The second stage of determining robust values of λ_N and λ_P consists of transforming the data with various values of the two parameters until the hypothesis $\lambda_0 = 1$ for the transformed data is not rejected.

7 The Fan Plot and its Extension

Like the Box and Cox likelihood ratio test, the approximate score test of §4 is based on aggregate statistics and so will be sensitive to the presence of outliers and more general model mis-specification. If only a very few outliers are present they can sometimes be determined by the use of deletion diagnostics. However these methods, working back from a fit to all the data, rapidly become computationally infeasible. If there are outliers or other anomalous observations they can be more effectively exposed by robust methods. Marazzi *et al.* (2009) choose MM estimation (Yohai, 1987). Since they use a robust form of the likelihood ratio test for λ, they need to establish an approximation to $\hat{\lambda}$, to establish which they search over 201 values of λ. To detect the effect of individual observations on the estimated transformation we use the forward search (Atkinson *et al.*, 2010). Since we use a score test we do not need to find $\hat{\lambda}$ and typically need to search over only five or six values of λ. This robust method orders the data from those closest to the fitted model to those most remote: any outliers will enter at the end of the search. We thus obtain a series of subsets of the data of size $m, m_0 \leq m \leq n$ for

each of which we ref t the model and calculate the values of the score statistics for selected values of λ. These are then plotted against the number of observations m used for estimation to give the "fan plot". The subsets range in size from $m_0 = p + 2$ to the full sample, where $p + 1$ is the number of covariates including the constructed variable. The fan plot for the Box-Cox transformation was introduced in Riani and Atkinson (2000), which also includes a description of the forward search for regression. A more extended treatment is in Atkinson and Riani (2000), which additionally exhibits the failure of deletion diagnostics. This discussion demonstrates how seeming outliers may be generated by use of an incorrect value of λ_0. When the value of λ_0 is incorrect, these apparent outliers will enter towards the end of the search and give rise to increasing values of the score statistic. For a different value of λ_0 other observations may appear outlying or, for a value supported by the data, none may.

Here we use the constructed variables of §5 to introduce the fan plot for the Yeo-Johnson transformation and extend it, using the results of §6, to include testing for homogeneity of the transformation. The procedure involves one search and almost $3n$ calculations of the score statistic for each value of λ, typically f ve. There is thus an appreciable computational saving over procedures such as the likelihood ratio or Wald tests which required numerical optimization to f nd the value of $\hat{\lambda}$ at each step of the search. The fan plot was also used by Atkinson and Riani (2002a) for monitoring t-tests for regression coeff cients in a linear model; they employ orthogonality arguments to show that the forward test statistics have exactly a t distribution. However for transformations the constructed variables are functions of the response and so the statistics cannot exactly follow this t distribution. Comments on the distribution of the score tests for the Yeo-Johnson transformation are in §11.

Once the data have been correctly transformed, the extended fan plot for the transformed data with $\lambda_0 = 1$ should lie within the bounds for all values of m. We use this method at the end of our analysis of the investment fund data in §8 and of the balance sheet data in §9 to conf rm a transformation of the data which has different values of λ for positive and negative observations. Before analysing our sets of data, we use simulation to illustrate the null and non-null distributions of this score test.

Figure 2 shows the results of 10,000 simulations of 200 observations from a normal regression model in which no response transformation is needed; there are three explanatory variables and the parameters are chosen to give fairly strong regression with an average R^2 value of 0.8. This extended fan plot is a "forward plot" in which the value of the score test is plotted against the size of the subset for which it was calculated. There are trajectories of 5 empirical quantiles of the simulated distribution: 0.5%, 25%, 50%, 75% and 99.5%, the outer quantiles thus providing a 99% interval for values of the statistic. For each quantile, three

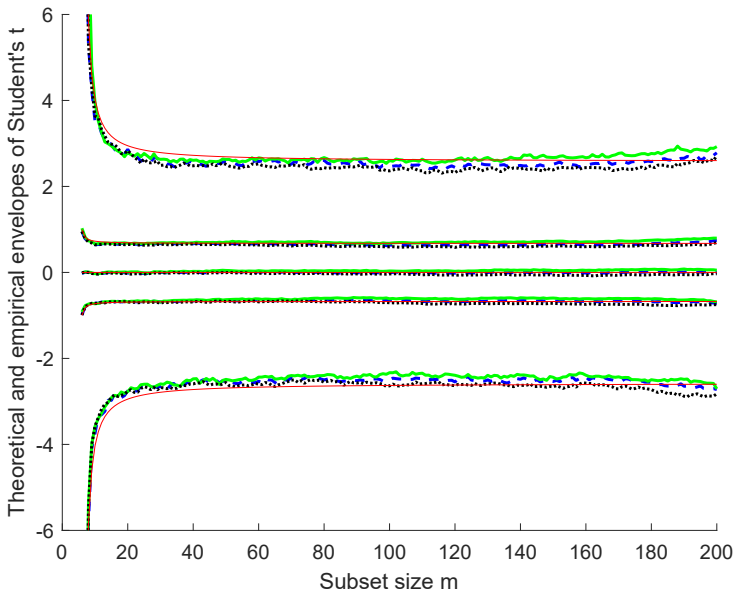

Figure 2: Extended fan plot showing null distribution of score statistics; 0.5%, 25%, 50%, 75% and 99.5% envelopes. For each quantile, reading down, the three statistics are for positive, all and negative observations (green, blue and black in the online version); continuous (red) line, t distribution on $m - 5$ d.f. 10,000 simulations of 200 observations. Null hypothesis $\lambda_0 = 1$

envelopes are plotted. In all cases, reading downwards, these are the statistics for the tests of positive, all and negative observations. Also included are percentage points of the t distribution on $m - 5$ degrees of freedom, which provide an excellent approximation to the values of the envelopes. The f gure shows how rapidly these envelopes approach the constant values of the normal quantiles. In §10 we discuss some points of the f ne structure of this plot which do not affect the interpretation given here.

Figure 3 shows a similar plot, but now when the null hypothesis of no transformation is false. The simulated positive observations were inversely transformed to require the transformation $\lambda_P = 1.5$ and the negative observations were inversely transformed to require $\lambda_N = 0$. For visual simplicity only the 25%, 50% and 75% quantiles of the empirical distribution of the statistic are shown. Now, at the end of the search, the intervals for the three statistics are separate with all lying below the 50% interval for the t distribution. The three uppermost trajectories are for the test for positive observations, with the trajectories for the overall transformation and the negative observations departing more from the bounds for the null distribution. This plot shows that, as the subset size increases, so does the power of the

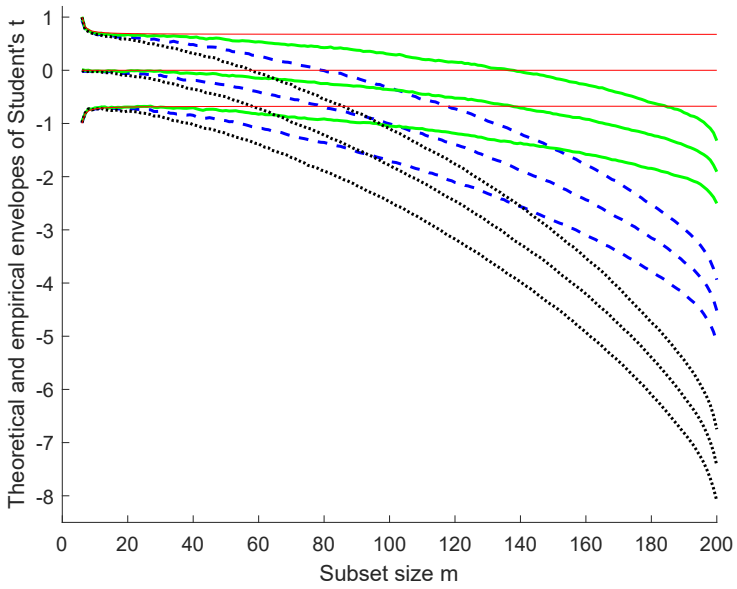

Figure 3: Extended fan plot showing non-null distribution of score statistics; 25%, 50% and 75% envelopes. For each set of three quantiles, reading down, the statistics are for positive, all and negative observations (green, blue and black in the online version); continuous (red) line, t distribution on $m - 5$ d.f. 10,000 simulations of 200 observations; $\lambda_P = 1.5$, $\lambda_N = 0$, null hypothesis $\lambda_0 = 1$

tests for positive and negative observations, as well as the power of the overall test. Further simulations, not included here, show how the power of the test increases with increasing the number of observations or reducing the error variance.

8 Analysis of the Investment Fund Data

8.1 Parametric Transformations

We start our analysis of the investment fund data using the transformation of §3 in which negative and positive observations are subject to the same transformation. The upper panel of Figure 4 shows the fan plot for values of λ_0 in the range 0.5 to 1. All curves are relatively smooth; there are no sudden changes towards the end of the trajectories which might indicate the presence of outliers. A value of 0.7 seems to be indicated for λ, although the trajectory for this value is outside the upper 1% bound around m = 200, returning inside from m around 240. This behaviour is typical of the masking caused by a sizable group of observations that differ systematically from the majority of the data. Indeed, the extended fan plot

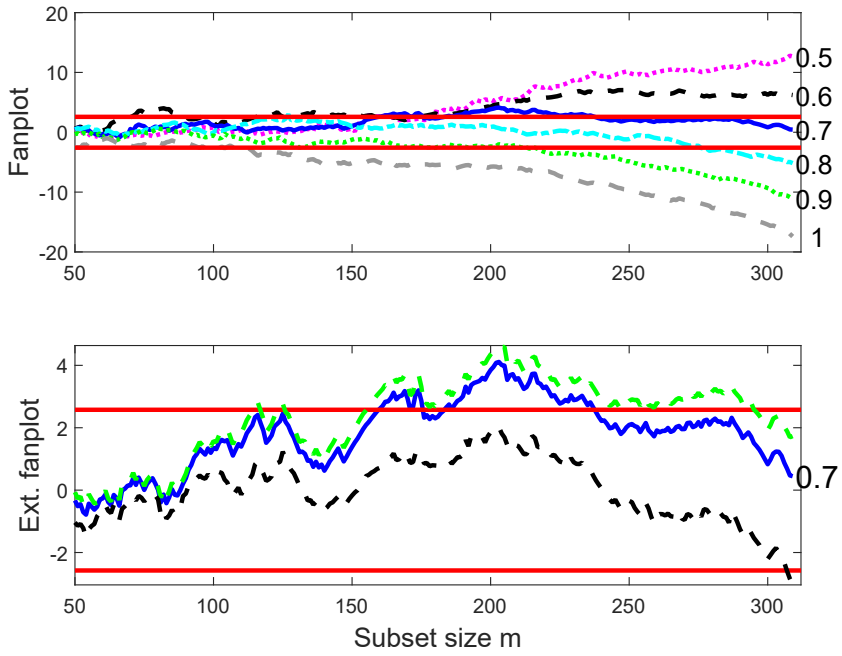

Figure 4: Investment fund data. Upper panel, fan plot indicating the overall transformation $\lambda = 0.7$; lower panel, extended fan plot for $\lambda_0 = 0.7$ suggesting different transformations for positive (upper green trajectory) and negative responses (lower black trajectory)

for $\lambda = 0.7$ in the lower panel of Figure 4 shows that the value of 0.7 is not satisfactory for all observations. The upper trajectory for the majority positive observations lies relatively close to that for an overall transformation of $\lambda = 0.7$, whereas that for the negative responses is outside the lower bound at the end of the search. This plot is a conf rmation of the suggestion of a different distribution for positive and negative observations suggested by the scatterplots of Figure 1.

The indication of the extended fan plot of Figure 4 is that the positive observations require a transformation higher than 0.7, since, from the upper panel, higher values of λ give a lower curve. Likewise, the negative observations require a value lower. Our strategy is to try sets of pairs of values. When we have found the correct transformation, the fan plot of the transformed data will indicate that no further transformation is required; that is we will accept the value $\lambda_0 = 1$ in this fan plot. For each tentative transformation we not only look at the fan plot of the transformed data but also look at the scatterplots and perform the full-sample regression on the two explanatory variables.

There are now two transformation parameters which, in line with the notation of §6, we call λ_P and λ_N. We start with two values straddling 0.7 and take $\lambda_P = 1$ and $\lambda_N = 0.5$. Comparison of the scatterplots in Figure 5 with those of the un-

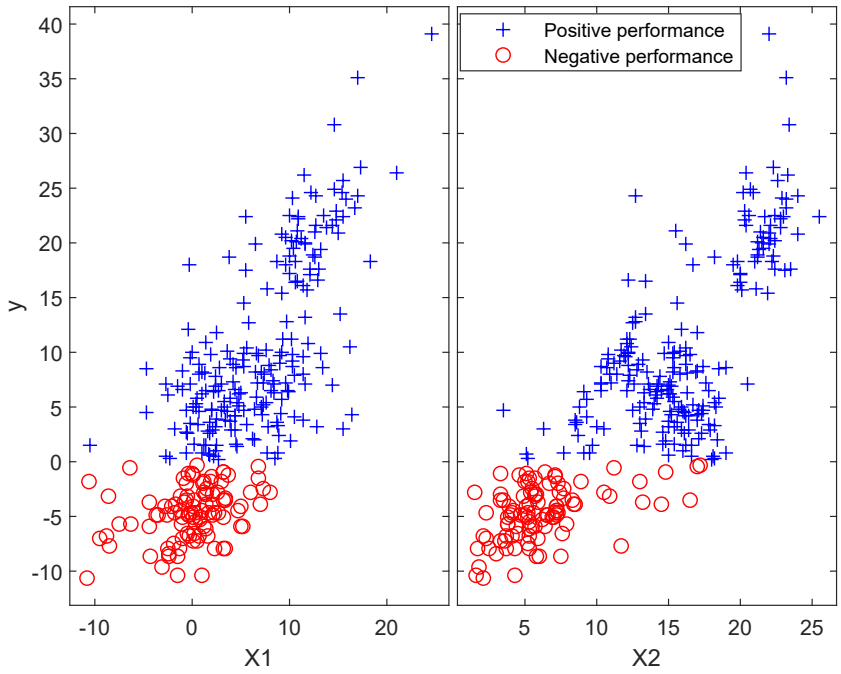

Figure 5: Investment fund data. Scatterplots of y against x_1 and x_2 for transformed data with $\lambda_P = 1$ and $\lambda_N = 0.5$

transformed data in Figure 1 shows that the variance of the negative observations is now closer to that of the positive ones. We check this impression with the extended fan plot for $\lambda_0 = 1$ for the transformed data in the upper panel of Figure 6. This plot provides a robust test that the transformation is $\lambda_P = 1$ and $\lambda_N = 0.5$. The correct transformations have not been found; although the trajectories for the positive and negative observations are far closer together than they are in the lower panel of Figure 4 for the overall transformation $\lambda = 0.7$, when all trajectories are well outside the lower bound at the end of the search.

The extended fan plot for $\lambda_P = 1$ and $\lambda_N = 0.25$ is in the centre panel of Figure 6 with that for $\lambda_P = 1$ and $\lambda_N = 0$ in the lower panel. These plots show improving properties; that for $\lambda_N = 0$ lies in the overall bounds throughout, with the trajectories for positive and negative observations virtually identical to the overall trajectory. The conclusion is that the positive observations should not be transformed, but that the negative observations should be transformed with $\lambda = 0$. From (2) it follows that this is not the log transformation for negative y.

The scatterplots of this final transformation are in Figure 7. The movement from the scatterplots of the original data in Figure 1 to a linear model with homogeneous scatter is clear. An interesting feature of the right-hand scatterplot in Figure 7 is that the observations appear to fall into three clusters, with funds with

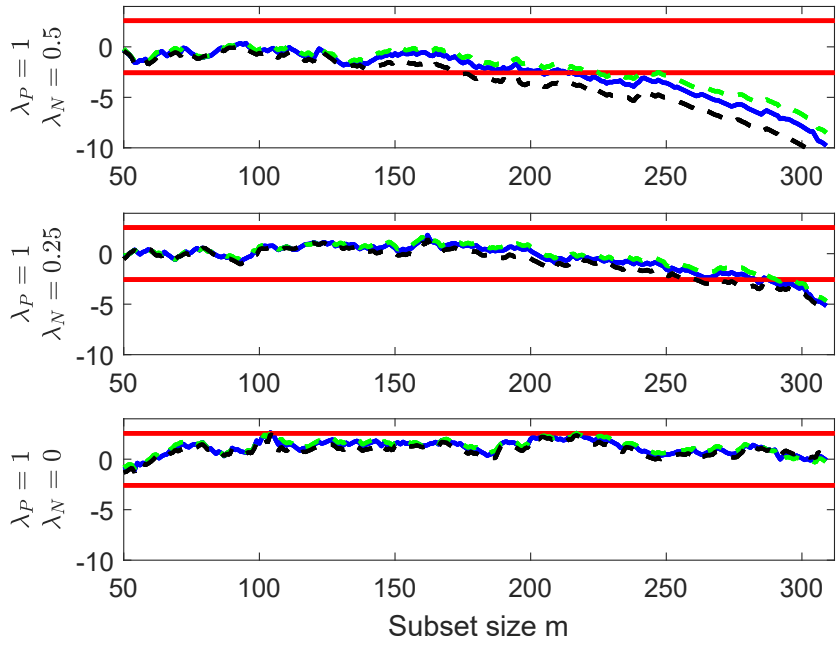

Figure 6: Investment fund data, checking the two transformation parameters. Extended fan plots for $\lambda = 1$ for the transformed data. Upper panel, $\lambda_P = 1$ and $\lambda_N = 0.5$; centre panel $\lambda_P = 1$ and $\lambda_N = 0.25$; lower panel $\lambda_P = 1$ and $\lambda_N = 0$, the preferred transformation. Upper (green) trajectory, positive observations, lower (black) trajectory, negative observations

high volatility being the most prof table over this time period.

We now consider two other statistical properties of the transformed and untransformed data. Table 1 shows summary properties of the regression on the two variables for $\lambda_P = 1$ and four values of λ_N. The line labelled F is the value of the F statistic for testing regression on a constant, x_1 and x_2 against regression on only a constant. As λ_N goes from 1 (the untransformed data) to 0 the value of the F statistic steadily increases from 556 to 685 in line with the results from the extended fan plots of Figure 6 which show the transformation becoming increasingly acceptable over this range. Likewise, the adjusted R^2 for the regression increases from 0.783 to 0.816. Interestingly, there is little change in the two tabulated values as λ_N goes from 0.25 to 0, although the centre panel of Figure 6 rejects the transformation with $\lambda_N = 0.25$.

The second assessment of the statistical properties of the transformation is given by the Normal QQ plots of Figure 8, with 90% pointwise intervals. The original data are in the left-hand panel. The transformation has made both tails more nearly normal and also produced a distribution which, for central values is

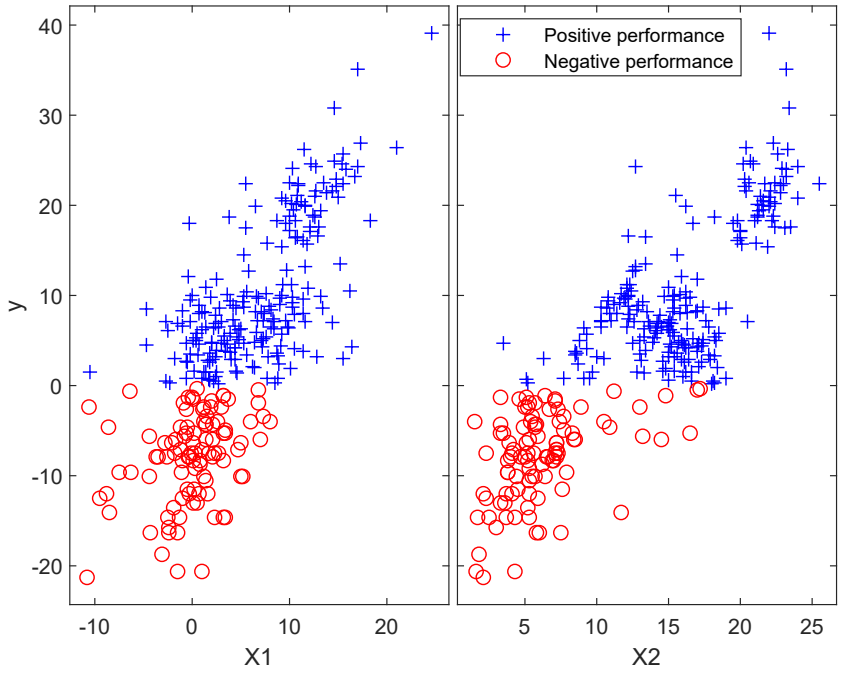

Figure 7: Investment fund data. Scatterplots of y against x_1 and x_2 for f nal transformation of the data with $\lambda_P = 1$ and $\lambda_N = 0$

slightly closer to the line of expected values for a normal sample. The differences between the two panels are clarif ed by the envelopes, especially in the tails of the distributions. The differences may not seem large, but slightly over 2/3rds of the observations are untransformed in both plots. The scatterplots show more forcibly the effect of transformation.

These results indicate that the Yeo-Johnson transformation and its extension to differing transformations for positive and negative values, has here achieved the three goals of the Box-Cox transformation. As the scatterplots of Figure 7 and the QQ plots of Figure 8 show, we have achieved a homogeneous and normal distribution of errors. The results summarized in Table 1 quantify the increasing

Table 1: Investment fund data: summary properties of regression for different transformations of positive (λ_P) and negative (λ_N) observations

λ_P	1	1	1	1
λ_N	1	0.5	0.25	0
$F_{2,306}$	556	643	681	685
R^2	0.783	0.807	0.815	0.816

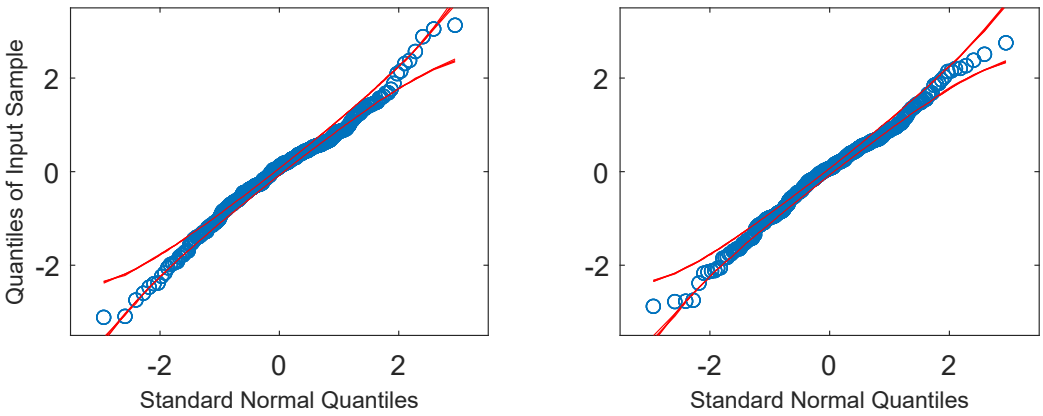

Figure 8: Investment fund data; normal QQ plots of residuals. Left-hand panel, original data; right-hand panel, final transformation of the data with $\lambda_P = 1$ and $\lambda_N = 0$

amount of the total variation in the data that is explained by regression using the correct transformation of the data. It is interesting that with 210 positive observations out of 309, linear interpolation in the estimates of λ suggests an overall transformation value $\tilde{\lambda} = (99 \times 0 + 210 \times 1)/309 = 0.680$, very close to the value of 0.7 arrived at in the overall fan plot of Figure 6. This satisfactory linear interpolation accords well with the Taylor series linearisation used to develop the approximate score statistic. These results could not have been achieved without the extension of the Yeo-Johnson transformation to test for and accommodate two transformation parameters.

8.2 Nonparametric Transformations

The Box-Cox and Yeo-Johnson transformations produce a smooth relationship between $y(\lambda)$ and the original y determined by the value of the parameters λ. A non-parametric alternative is to use smoothing to estimate this relationship. We use two such methods.

Both methods can transform explanatory variables and the response. The assumed model is a generalized additive model, that is one with transformations of both response and explanatory variables but without interactions. Both rely on repeated application of univariate smoothers. In ACE (alternating conditional expectations) Breiman and Friedman (1985) maximize a measure of correlation between all variables. As a consequence, the response variable in regression is not treated as being different from the explanatory variables. Tibshirani (1988) describes a related method in which the transformation for the response is intended

to yield additivity and variance stabilization (AVAS). The asymptotic variance stabilizing transformation is applied to the response.

Hastie and Tibshirani (1990, Chapter 7) provide a description of both ACE and AVAS with an emphasis on response transformation and the mathematical relationship to the Box-Cox transformation. Breiman and Friedman (1985) and their discussants, as well as Tibshirani, show examples of data analyses which raise questions about the performance of the two algorithms under certain conditions. In their contribution to the discussion of ACE, Buja and Kass (1985) express concern about inference, diagnostics and robustness. In the rejoinder to the discussion of their paper, Breiman and Friedman (1985) admit that ACE is not robust.

A difficulty in the comparison of parametric transformations with ACE and AVAS is that these are both model fitting techniques lacking many of the usual statistical justifications and machinery. In the Box-Cox and Yeo-Johnson transformations, the Jacobian of the transformation provides the basis for the comparison of analyses with various values of parameters λ. But the outputs of ACE and AVAS are a set of transformed responses, scaled to have unit variance. The aggregate statistic for comparison of models is the values of R^2.

The original progams for both and AVAS are written in 'classical' Fortran, without comments and with often non-informative variable names. This Fortran code also provides the basis of the R package Acepack. For our comparisons we have rewritten the packages in Matlab that has been fully compared and validated against the original Fortran.

Figure 9 shows the plot of parametrically transformed y against original y for the investment fund data when $\lambda_N = 0$ and $\lambda_P = 1$, with the transformed values scaled to have unit variance. Since there is no transformation for the positive observations, the right-hand part of the plot is a straight line. There is a clear change to a different relationship for negative data.

The two panels of Figure 10 compare the non-parametric transformations with the parametric transformation of Figure 9. In the left-hand panel the transformation from ACE is virtually straight for negative y, but more curved for positive y, with two points of inflection, the first around $y = 4$. There is also disagreement at the five or so smallest observations and the three largest. The transformation for AVAS in the right-hand panel is virtually the same as the parametric transformation for the more extreme observations. There are again two points of inflection, with the largest distance from the parametric transformation around $y = 20$.

We repeated the estimation of these transformations starting from $\lambda_N = 0$ and $\lambda_P = 1$ rather than from a value of one for each parameter. The changes were slight, the largest being that AVAS produced a transformation for the three largest

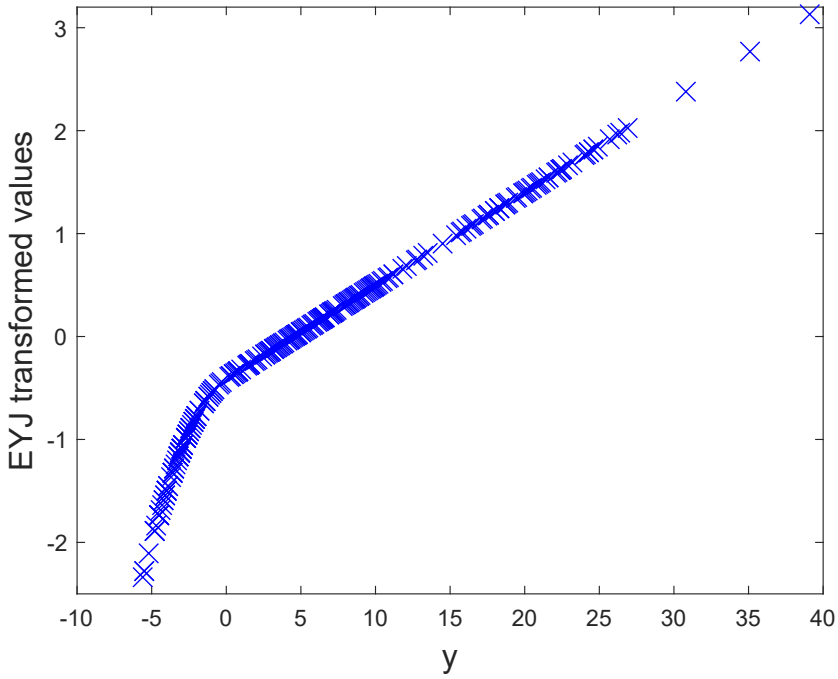

Figure 9: Investment fund data. Transformed responses from the extended Yeo-Johnson transformation with $\lambda_N = 0$ and $\lambda_P = 1$ against original response

observations closer to that found by ACE in the left-hand panel of Figure 10.

Some values of R^2 for regression on different transformations of the response are shown in Table 2. As Table 1 shows, the parametric transformation increases R^2 from 0.783 to 0.816. The comparable values for ACE and AVAS, starting from the untransformed data, are 0.836 and 0.817. ACE in this case performs better than the parametric transformation. The left-hand panel of Figure 10 suggests this may in part be because this transformation is not restricted to having only two transformation regimes which meet at zero. It is perhaps an indication that the cluster of observations for funds with medium volatility in the right-hand panel of Figure 7 would benef t from a different transformation. Although the AVAS transformation also curves in a similar way, it gives a value of 0.817 for R^2, very close to that for the parametric transformation, perhaps because there is appreciable divergence from the parametric transformation for values of y around 20, rather than close to zero for ACE, with agreement in the tails of the distribution. Starting the search for either nonparametric transformation from that found perimetrically results in values of R^2 very slightly less than those found when starting from the untransformed data.

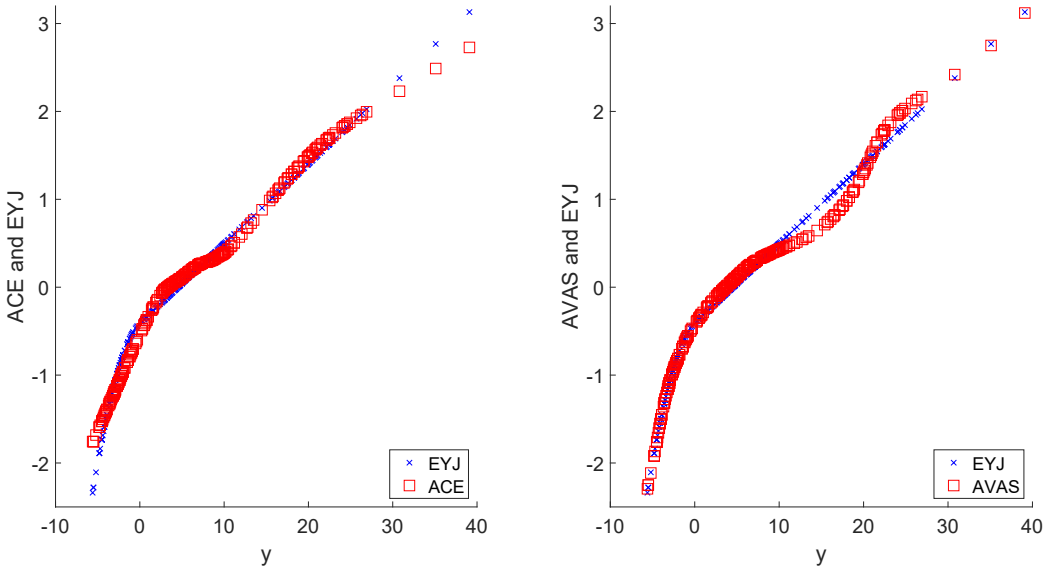

Figure 10: Investment fund data. Transformed responses from nonparametric transformations and from the extended Yeo-Johnson transformation with $\lambda_N = 0$ and $\lambda_P = 1$ against original response. Left-hand panel ACE, right-hand panel AVAS

9 A Robust Analysis of Balance Sheet Data

9.1 Parametric Transformation

After transformation, the investment fund data are surprisingly well behaved. In particular, there are no obvious indications of any outliers. We now analyse a larger data set taken from balance sheet information on limited liability companies, which does include outliers.

The response is profitability of individual firms in Italy. There are 998 observations with positive response and 407 with negative response, making 1,405 observations in all. The model variables are:

y	profitability, calculated as return over sales;
x_1	labour share; the ratio of labour cost to value added;
x_2	the ratio of tangible fixed assets to value added;
x_3	the ratio of intangible assets to total assets;
x_4	the ratio of industrial equipment to total assets;
x_5	the firm's interest burden; the ratio of the firm's total assets to net capital.

The aim is to explain the profitability by regression on the five explanatory

Table 2: Investment fund data: values of R^2 from regression for some parametric and non-parametric transformations: EYJ, the extended Yeo-Johnson transformation. For ACE and AVAS the values of λ_N and λ_P are the initial transformation of the data

λ_P	1	1
λ_N	1	0
EYJ	0.783	0.816
ACE	0.836	0.834
AVAS	0.817	0.814

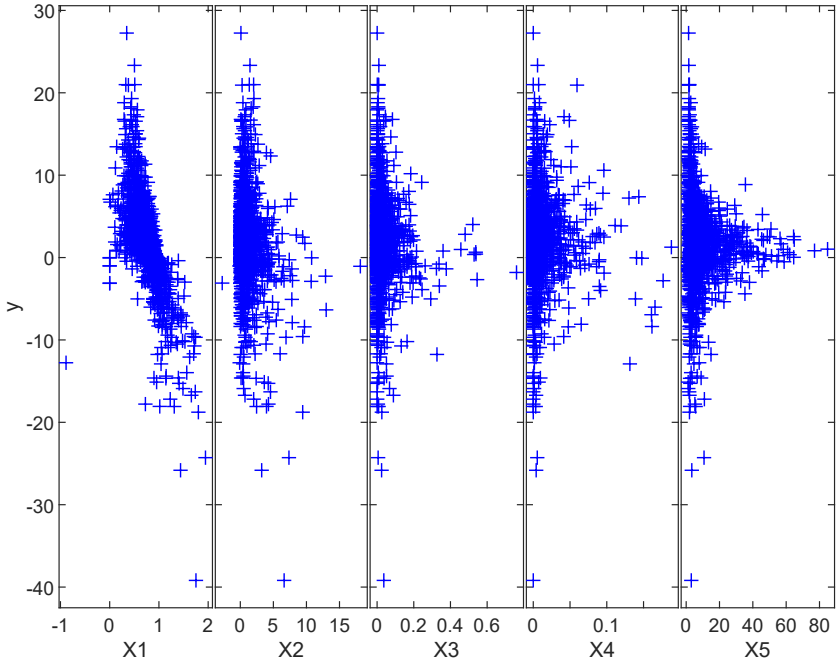

Figure 11: Balance sheet data: scatterplots of y against $x_1 - x_5$

variables. We start with the untransformed response. Scatterplots of y against the explanatory variables are in Figure 11. It is clear that there is a negative relationship between y and x_1. The relationships with the other variables are not so obvious. However, the values of the t-tests for the coeff cients in Table 3 show signif cant regression on all variables except x_4. Unlike the plot of the investment fund data in Figure 1, these plots do not convey an obvious message about the need for transformation.

We check the need for a transformation using fan plots. The top panel of Fig-

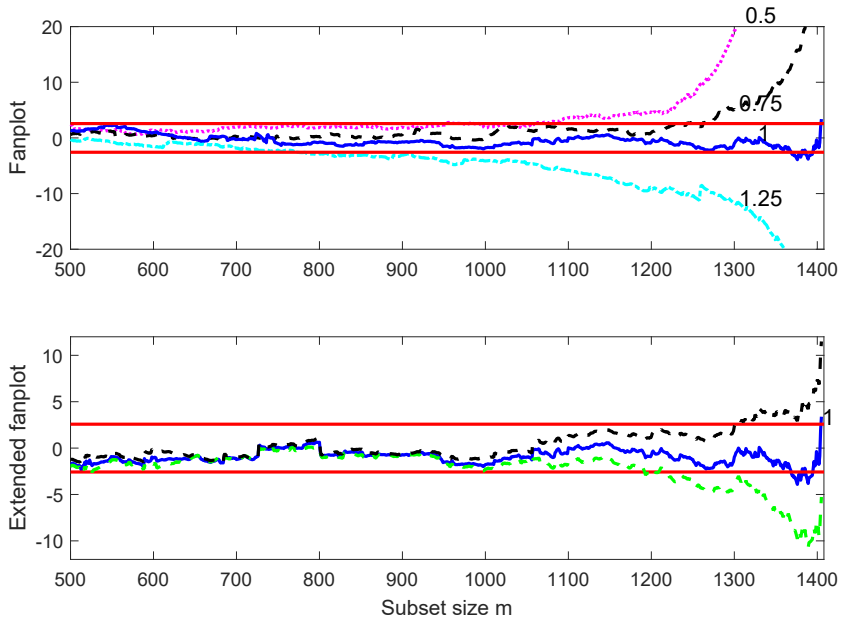

Figure 12: Balance sheet data. Upper panel, fan plot of overall statistic for $\lambda_0 = 0.5, 0.75, 1$ and 1.25, perhaps indicating the transformation $\lambda = 1$; lower panel, extended fan plot for $\lambda_0 = 1$; the need for different transformations for positive and negative observations is apparent, as is the effect of outliers. Upper (black) trajectory, negative observations, lower (green) trajectory, positive observations

ure 12 presents trajectories for the overall statistic for values of $\lambda_0 = 1.25, 1, 0.75$ and 0.5. Comparison with the fan plot for the investment fund data in Figure 4 shows the increase in power consequent on moving from a sample of 309 to one of 1,405. The hypothesis of no transformation $(\lambda_0 = 1)$ seems to be acceptable, although there is an abrupt increase in the value of the statistic towards the end of the search which might indicate the presence of outliers. The extended fan plot for testing $\lambda_0 = 1$ in the lower panel of the f gure clarif es this structure. This plot shows that the positive and negative observations apparently need different transformations. Unlike all the extended fan plots we have seen so far from Figure 2 onwards, the (black) trajectory for the negative observations is uppermost, indicating the need for a value of $\lambda_N > 1$. Likewise the lowest (green) trajectory indicates that λ_P should be less than one. (The upper panel of the f gure shows the effect of the value of λ_0 on the plot of the overall score statistic). The plot also shows a sharp increase in the values of all three statistics at the end of the search. This increase for the overall transformation is preceded by a decline in values, perhaps indicative of two different kinds of outlier.

Following the indication of Figure 12, we continue this analysis by f nding the

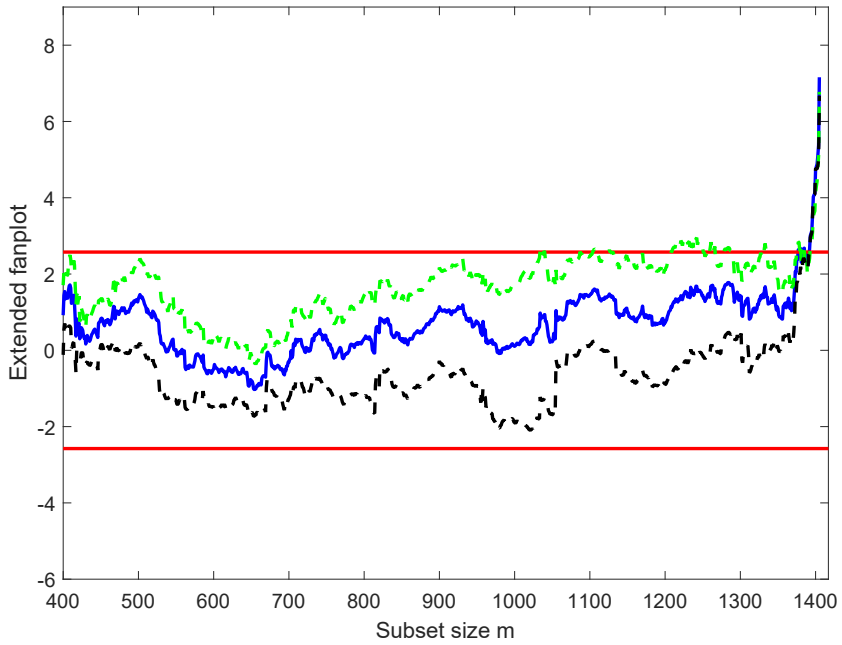

Figure 13: Balance sheet data, checking the two transformation parameters. Extended fan plots for $\lambda_0 = 1$ for the transformed data when $\lambda_P = 0.5$ and $\lambda_N = 1.5$ Upper (green) trajectory, positive observations, lower (black) trajectory, negative observations

best values of the two transformation parameters. The extended fan plot for testing the hypothesis that $\lambda_0 = 1$ for data transformed with $\lambda_P = 0.5$ and $\lambda_N = 1.5$ is in Figure 13. Although the trajectories for the three statistics do not overlap as they do for the well-behaved investment fund data in Figure 6, the three statistics lie virtually in the t-statistic boundaries until almost the end of the search, where the trajectories coincide when the suspected outliers enter. Other parameter values can be found for which the three trajectories are much closer, for example $\lambda_P = 0.75$ and $\lambda_N = 1$. However, the trajectories stray outside the t-statistic boundaries and indicate around 200 outliers.

The fan plot and its extension have been central to our modelling process. The fan plot depends on the forward search which orders the observations by closeness to the model f tted for each value of the subset size m. We now f nish with part of a forward search analysis of data with the recommended transformation $\lambda_P = 0.5$ and $\lambda_N = 1.5$ which leads to identif cation of outlying observations.

The top left-hand panel of Figure 14 shows a forward plot, that is a plot against subset size, of all 1,405 scaled residuals for values of m from 800. There is a broad upper band of residuals, in pale blue in the online version of the paper, which

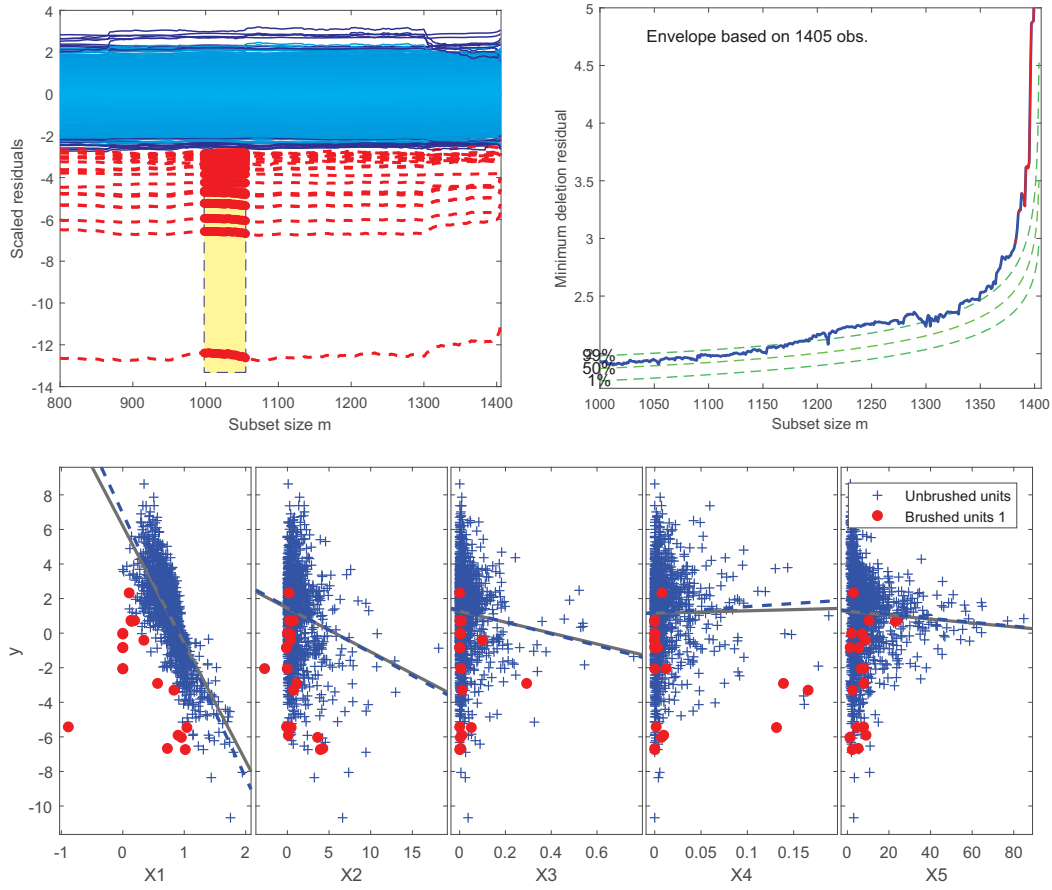

Figure 14: Balance sheet data; brushing linked plots from the forward search when $\lambda_P = 0.5$ and $\lambda_N = 1.5$. Upper left-hand panel, trajectories of residuals from the forward search with the residuals from 19 observations highlighted by brushing. Upper right-hand panel, linked forward plot of minimum deletion residuals during the search with the 19 brushed values shown in red. Lower panel, scatterplots of y against $x_1 - x_5$ showing the 19 brushed units: full line, regression with all data, dashed line regression after outlier deletion

lies between ± 2 throughout this part of the search. The slightly more extreme residuals are plotted in dark blue. The lower ones of these are slightly separated from a lower band of residuals, shown in red. What is remarkable is the stability of this pattern almost until the end of the search, indicative of a set of data lacking outliers having strong individual inf uence on the parameter estimates.

The upper right-hand panel of the f gure shows a forward plot of minimum deletion residuals, with a 99% envelope (dotted line) for these distances for the

full sample of 1,405 observations. Outlying observations cause the data values (in blue) to fall outside the envelopes. This is a first step in outlier detection; the automatic outlier detection procedure establishes the outlier free sample size against which outliers are judged. The lower panel of Figure 14 shows scatterplots of y against the explanatory variables.

To exhibit more information, these three plots have been linked. The highlighted, red, residuals in the top left-hand panel were subjectively identified by brushing the plot. We selected the trajectories, 19 in all, that lay within the brush in the centre of the figure. The trajectory of the 19 brushed observations in the forward plot of deletion residuals in the upper right-hand panel is shown in red. Sixteen of these observations are the last observations to enter the search, the other three entering from 21 units before the end of the search. These 19 observations are plotted as filled (red) circles in the scatterplot in the lower panel. The extreme negative trajectory in the top left-hand panel is clearly caused by the outlier in the bottom left of the scatterplot for x_1. This is a gross outlier, since x_1, the ratio of labour cost to value added, is rarely negative. The continuous lines in the panels come from multiple regression on all observations. The panels for x_1 and x_4, in particular, show the source of the many large negative residuals.

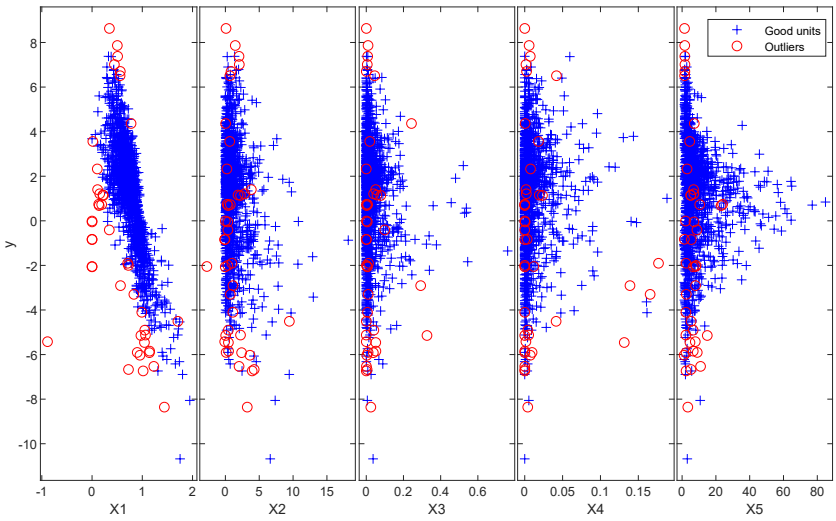

Figure 15: Balance sheet data: scatterplots of transformed y against $x_1 - x_5$ showing the 42 outliers found by the automatic procedure for outlier detection when $\lambda_P = 0.5$ and $\lambda_N = 1.5$

Brushing led to the subjective choice of 19 observations with large negative residuals. However, our automatic procedure for outlier detection (Riani *et al.*, 2009) identifies a total of 42 outliers. These observations are plotted in the panels

of Figure 15. On this transformed scale the outliers are evident, which they are not on the original scale in Figure 11.

The purpose of the analysis of these data was to build a regression model relating y to the five potential explanatory variables. Table 3 gives summary properties of the regressions for the untransformed data and the transformed data with 42 observations deleted. The F statistic for the regression, as against a constant model, increases on transformation and deletion from 293 to 588, a ratio of just over 2. Thus, despite the deletion of 42 observations, the transformation has led a doubling of the amount of information available, that is to an effective doubling of sample size; the t test values for the intercept and all variables except x_5 show appreciable increases in significance. The strengthening of evidence for multiple regression to include x_1, from -35.8 to -51.4, is in line with the pattern of outliers shown deleted in the first panel of Figure 15. However, the evidence for x_2 strengthens from -8.3 to -11.3, which is not particularly to be expected from inspection of the second panel of the figure.

The lower panel of Figure 14 also includes, as continuous lines, the fitted regression relationships before the outliers have been deleted. The dotted lines are for the relationships after transformation and outlier deletion. Although the significance of the fit has greatly increased due to outlier deletion, the figure shows that deletion of the outliers has a negligible effect on the estimated regression coefficients.

The negative sign for x_1 shows decreasing profitability for firms with high labour input. This is by far the most significant relationship. The negative signs for x_2 and x_3, ratios of fixed assets and intangibles such as software, are surprising. One possibility is that the investments that have been made are not yet yielding the intended advantages in, for example, labour reduction, whilst incurring capital costs (Bartoloni, 2013, §3). The fourth variable is not significant, whereas the negative effect of interest is to be anticipated. There is however a danger in over-interpreting the values of t-tests in multiple regression with correlated explanatory variables. Table 3 shows that the increase in the precision of estimation of the regression coefficients has been achieved by the transformation coupled with the deletion of only 3.9% of the total data.

9.2 Nonparametric Transformations

The results of the nonparametric analysis of the investment fund data in §8.2 showed that ACE provided a somewhat higher value of R^2 than the extended Yeo-Johnson transformation, although the shapes of the transformed responses in the two panels of Figure 10 do not show large differences between the parametric and non-parametric transformations. However, due to the presence of the outliers, there is a sharper difference between these two forms of transformation

Table 3: Balance sheet data: summary properties of regression for original data and for data with outliers removed with different transformations of positive (λ_P) and negative (λ_N) observations

λ_P	1	0.5
λ_N	1	1.5
Number of observations	1405	1363
Error d.f. ν	1399	1357
t_ν values		
Intercept	42.0	60.5
x_1	-35.8	-51.4
x_2	-8.3	-11.3
x_3	-3.6	-5.4
x_4	1.0	2.3
x_5	-3.4	-3.1
$F_{5,\nu}$ for regression	293	588
R^2	0.511	0.684

when analysing the balance sheet data.

A summary through the values of R^2 of the effect of outliers on the performance of the transformations is in Table 4. The outliers were identified in §9.1 on a scale found using the extended Yeo-Johnson transformation. Deletion of these observations without transformation increases R^2 from 0.511 to 0.638. For the extended Yeo-Johnson transformation in the presence of outliers the value is 0.559. The combination of transformation of the data and deletion of 42 outliers increases R^2 further to 0.684, which is the meaningful comparison. For ACE deletion of the same 42 outliers increases R^2 from 0.558 to 0.697; for AVAS the increase is from 0.526 to 0.646. Thus, after the outliers have been deleted, ACE produces a slightly large value of R^2 than the extended Yeo-Johnson transformation, whereas the value from AVAS is smaller.

10 Distribution of the Score Statistics

Mathematical analysis of the properties of the Box-Cox transformation is not straightforward, due to the need to find expectations of functions similar in structure to the constructed variables of §5. Examples are Draper and Cox (1969), Taylor (1986) and Cox and Reid (1987) who work with the un-normalized transformation $y(\lambda)$. Atkinson and Riani (2002b) provide some numerical results on the distribution in the fan plot of the score statistic for the Box-Cox transforma-

Table 4: Balance sheet data: values of R^2 from regression for parametric and non-parametric transformations on complete data and data with 42 outliers deleted: EYJ, the extended Yeo-Johnson transformation

Data	Complete	Outliers deleted
Untransformed	0.511	0.638
EYJ	0.559	0.684
ACE	0.558	0.697
AVAS	0.526	0.646

tion. Following the results on forward t-tests mentioned in §7, they examine the departures from the null distribution, which are most extreme towards the end of the search, where the statistic has too large a variance. They show that increasingly strong regression relationships lead to distributions that are closer to t, a f nding that explains the results of Atkinson and Lawrance (1989) on the comparison of tests for the Box-Cox transformation. Careful inspection of the simulated null distribution of the statistics in Figure 2 shows that towards the end of the search, all three envelopes for the extreme quantiles likewise indicate a slightly longer-tailed distribution than t. Furthermore, they are slightly too tight in the middle of the search.

The simulation of the null distributions of the statistics in the extended fan plots of Figure 2 also shows that, for all quantiles, the statistic for positive y lies slightly above the overall statistic and that for the negative observations lies slightly below. Some insight into this result comes from the structure of the score statistic (8), which depends on quadratic forms in w and z. In the Yeo-Johnson transformation the quadratic forms contain two terms, one for the positive observations and one for the negative. When testing for positive y the constructed variable for the negative observations for general λ is $z(\lambda) \log \dot{y}_P$. From the definition in (11), $\dot{y}_P$ is always positive, so that the contribution to the numerator of $\hat{\gamma}$ is a multiple of the residual sum of squares of the negative transformed observations. Similarly, for testing the negative observations, with $\dot{y}_N$ likewise positive (10), the contribution is minus a multiple of the residual sum of squares of the positive transformed observations.

11 Discussion

The Yeo-Johnson transformation has been widely cited, particularly in hydrology, perhaps because zero values of rainfall are common, combined with a skew positive distribution of values. However many authors, for example Su *et al.* (2009)

in a medical context, argue that such semi-continuous data are better analysed by treating separately the continuous and zero parts. QQ plots of residuals after transformation would reveal the effect of a spike of observations at zero by giving a horizontal part of the plot. The (un-normalized) transformation has also been applied to the distribution of explanatory variables (Peterson, 2018), much in the spirit of Whittaker *et al.* (2005) who use the related neglog transformation.

Our extended Yeo-Johnson transformation allows different transformations for positive and negative observations. An extension is to two transformation regimes for observations above and below some threshold, for example medical subjects with high or normal measurements of a response. The two non-parametric transformations we have discussed should be helpful in indicating the value of the threshold. Indeed, there is an indication of a threshold higher than zero for the ACE transformation of the investment fund data in the left-hand panel of Figure 10. Such transformations can also be expected to have advantages over parametric transformations for large data sets, where there may be more than two groups divided by response value that require distinct transformations. Comments on the forward search for large data sets are given by Riani *et al.* (2019). Computation times can be appreciably reduced by moving forward in blocks of $K > 1$ observations, rather than incrementing the subset one observation at a time as we have done here.

Both ACE and AVAS allow the transformation of explanatory variables as well as of the response. These methods thus provide a generalization of the transformation of explanatory variables (Box and Tidwell, 1962). The approach is distinct from the 'transform both sides' method described in Carroll and Ruppert (1988, Cap.4), in which the relationship between the mean response and the model is known. The purpose of the transformation is to achieve approximate normality and constancy of variance for the response. In any such developments of transformation methodology, robustness, such as we have achieved here, combining the forward search and a fan plot, will be important.

The calculations in this paper used Matlab routines from the FSDA toolbox, available as a Matlab add-on from the Mathworks f le exchange `https://www.mathworks.com/matlabcentral/fileexchange/`. The data, the code used to reproduce all results including plots, and links to FSDA routines are available at `http://www.riani.it/ARC2019`.

Acknowledgements

This research benef ts from the HPC (High Performance Computing) facility of the University of Parma. M.R. gratefully acknowledges support from the CRoNoS project, reference CRoNoS COST Action IC1408. A.C.A. would like to thank the European Union's Horizon 2020 Research and Innovation Programme

for its f nancial support of the Prime Fish project, Grant Agreement No. 635761.

References

Andrews, D. F., Gnanadesikan, R., and Warner, J. L. (1971). Transformations of multivariate data. *Biometrics*, **27**, 825–840.

Atkinson, A. C. (1973). Testing transformations to normality. *Journal of the Royal Statistical Society, Series B*, **35**, 473–479.

Atkinson, A. C. and Lawrance, A. J. (1989). A comparison of asymptotically equivalent tests of regression transformation. *Biometrika*, **76**, 223–229.

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

Atkinson, A. C. and Riani, M. (2002a). Forward search added-variable t tests and the effect of masked outliers on model selection. *Biometrika*, **89**, 939–946.

Atkinson, A. C. and Riani, M. (2002b). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemometrics and Intelligent Laboratory Systems*, **60**, 87–100.

Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: theory and data analysis (with discussion). *Journal of the Korean Statistical Society*, **39**, 117–134. doi:10.1016/j.jkss.2010.02.007.

Bartoloni, E. (2013). Prof tability and innovation: new empirical f ndings based on Italian data 1996-2003. *Rivista Internazionale di Scienze Sociali*, **121**, 137–170.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, **76**, 296–311.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211–246.

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, **77**, 209–210.

Box, G. E. P. and Tidwell, P. W. (1962). Transformations of the independent variables. *Technometrics*, **4**, 531–550.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and transformation (with discussion). *Journal of the American Statistical Association*, **80**, 580–619.

Buja, A. and Kass, R. E. (1985). Comment on "Estimating optimal transformations for multiple regression and transformation" by Breiman and Friedman. *Journal of the American Statistical Association*, **80**, 602–607.

Carroll, R. J. (1982). Prediction and power transformations when the choice of power is restricted to a f nite set. *Journal of the American Statistical Association*, **77**, 908–915.

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, New York.

Chen, G., Lockhart, R. A., and Stephens, M. A. (2002). Box-Cox transformations in linear models: large sample theory and tests of normality (with discussion). *The Canadian Journal of Statistics*, **30**, 177–234.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **49**, 1–39.

Draper, N. R. and Cox, D. R. (1969). On distributions and their transformation to normality. *Journal of the Royal Statistical Society, Series B*, **31**, 472–476.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, **79**, 302–309.

Marazzi, A., Villar, A. J., and Yohai, V. J. (2009). Robust response transformations based on optimal prediction. *Journal of the American Statistical Association*, **104**, 360–370. DOI: 10.1198/jasa.2009.0109.

Peterson, R. A. (2018). bestnormalize: Normalizing transformation functions. URL https://CRAN.R-project.org/package=bestNormalize.

Riani, M. and Atkinson, A. C. (2000). Robust diagnostic data analysis: Transformations in regression (with discussion). *Technometrics*, **42**, 384–398.

Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, **71**, 447–466.

Riani, M., Atkinson, A. C., Cerioli, C., and Corbellini, A. (2019). Discussion on the paper: Data Science, Big Data and Statistics, by Pedro Galeano and Daniel Peñ˜a. *TEST*. DOI: https://doi.org/10.1007/s11749-019-00647-5.

Su, L., Tom, B. D. M., and Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics*, **10**, 374–389.

Taylor, J. M. G. (1986). The retransformed mean after a f tted power transformation. *Journal of the American Statistical Association*, **81**, 114–118.

Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, **83**, 394–405.

Weisberg, S. (2005). Yeo-Johnson power transformations. https://www.stat.umn.edu/arc/yjpower.pdf.

Whittaker, J., Whitehead, C., and Somers, M. (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Applied Statistics*, **54**, 863–878.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959.

Yohai, V. J. (1987). High breakdown-point and high eff ciency estimates for regression. *The Annals of Statistics*, **15**, 642–656.