



# Ordered quantile normalization: a semiparametric transformation built for the cross-validation era

Ryan A. Peterson & Joseph E. Cavanaugh

To cite this article: Ryan A. Peterson & Joseph E. Cavanaugh (2019): Ordered quantile normalization: a semiparametric transformation built for the cross-validation era, *Journal of Applied Statistics*, DOI: [10.1080/02664763.2019.1630372](https://doi.org/10.1080/02664763.2019.1630372)

To link to this article: <https://doi.org/10.1080/02664763.2019.1630372>



Published online: 15 Jun 2019.



Submit your article to this journal 



View Crossmark data 



# Ordered quantile normalization: a semiparametric transformation built for the cross-validation era

Ryan A. Peterson <sup>a,b</sup> and Joseph E. Cavanaugh <sup>a</sup>

<sup>a</sup>Department of Biostatistics, University of Iowa College of Public Health, Iowa City, IA, USA; <sup>b</sup>Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

## ABSTRACT

Normalization transformations have recently experienced a resurgence in popularity in the era of machine learning, particularly in data preprocessing. However, the classical methods that can be adapted to cross-validation are not always effective. We introduce Ordered Quantile (ORQ) normalization, a one-to-one transformation that is designed to consistently and effectively transform a vector of arbitrary distribution into a vector that follows a normal (Gaussian) distribution. In the absence of ties, ORQ normalization is guaranteed to produce normally distributed transformed data. Once trained, an ORQ transformation can be readily and effectively applied to new data. We compare the effectiveness of the ORQ technique with other popular normalization methods in a simulation study where the true data generating distributions are known. We find that ORQ normalization is the only method that works consistently and effectively, regardless of the underlying distribution. We also explore the use of repeated cross-validation to identify the best normalizing transformation when the true underlying distribution is unknown. We apply our technique and other normalization methods via the `bestNormalize` R package on a car pricing data set. We built `bestNormalize` to evaluate the normalization efficacy of many candidate transformations; the package is freely available via the Comprehensive R Archive Network.

## ARTICLE HISTORY

Received 15 October 2018

Accepted 4 June 2019

## KEY WORDS

High-dimensional data analysis; preprocessing; predictive modeling; machine learning; normalizing transformation

## 1. Introduction

The normal (or Gaussian) distribution has laid the groundwork for countless statistical methodological frameworks, not the least of which is classical linear regression. As such, since Box and Cox introduced their seminal normalization transformation in 1964 [5], statisticians and researchers alike have recognized the widespread application of mappings that ‘Gaussianize’ data. Transforming data so that methods which require normality could be applied circumvented the arguably more challenging problem of developing statistical techniques that could accommodate non-normal data.

---

**CONTACT** Ryan A. Peterson  [ryan-peterson@uiowa.edu](mailto:ryan-peterson@uiowa.edu)  Department of Biostatistics, University of Iowa College of Public Health, 145 N. Riverside Dr., Iowa City, IA 52245, USA; Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, 13001 E. 17th Pl., Aurora, CO 80045, USA

One popular use of these transformations is to mediate the (sometimes problematic) assumption of normality of the outcome, conditional on the covariates, in the setting of classical linear regression. Due to the ubiquity of right-skewed outcomes, a common approach to this problem is to simply to model the log of the outcome, but a more complex transformation might be warranted. Over the years, many alternative (and arguably more elegant) frameworks have been developed to relax the normality assumption: generalized linear modeling, quantile regression, etc. Yet still, the practice of ‘beating the data to look normal via some kind of normalizing transformation’ is still widely employed for its simplicity. While perhaps not the most elegant solution to the problem, often this technique works well as a quick and pragmatic solution.

In the era of high-dimensional data, another increasingly popular application of normalization occurs in applied regression settings with highly skewed or irregular distributions in some of the covariates. Such settings often yield high leverage points (and thus possibly highly influential points), even when one centers and scales the covariates. When examining interactions, these influential points can become especially problematic. If a covariate has even one high-leverage value, each interaction with that covariate will amplify the leverage of that point. Normalization of the covariates mitigates the leverage and potential influence of these covariates to an extent, which in some cases will allow for more robust model selection. As a result, popular model selection packages in R such as `caret` and `recipes` have built-in mechanisms to normalize the covariates automatically [11,13]. This method is called a ‘preprocessing’ technique [12], and it essentially forgoes the assumption of linearity between the outcome and the covariate, opting instead for the premise of a linear relationship between the outcome and the *transformed* value of the covariate (which in many cases may be more plausible). Interestingly, Box and Tidwell advised a similar approach in 1962 [6]. A large benefit of preprocessing in this manner is that any transformations done *prior* to investigating the relationship between covariates and the response will yield valid inferences – the process is unsupervised. Many methods exist that can determine optimal transformations of the covariates and the response jointly, by minimizing some objective loss function. However, these supervised methods (i.e. methods that are trained using both the response and the covariates) must account for the tendency to overfit when making inferences [9].

Normalization transformations do not always provide an easy panacea. Since its introduction, the Box-Cox transformation has had known weaknesses. Thus, many statisticians have offered modifications and alternatives that work in more general cases. The Yeo-Johnson transformation attempts to minimize the Kullback-Leibler divergence between the normal distribution and the distribution of the transformed data [19]. The Lambert WxF ‘Gaussianizing’ transformation is a similar approach that uses maximum likelihood to estimate a parametric function that best normalizes a skewed or a heavy-tailed distribution [7]. The Lambert technique is applicable under the assumption of the existence of the first two moments; see [16]. Other examples of improvements and extensions of the Box-Cox transformation can be found in [4,10,14].

Unfortunately, even many of these now popular parametric normalization methods are neither consistent nor accurate; each parametric method relies on certain assumptions about the shape of the generating distribution. Thus, no method is guaranteed to normalize a vector, and no method will always be the optimal. This is problematic, especially in situations without prior knowledge of the shape of the distribution (for instance,

in high-dimensional regression settings where normalization needs to be automatic). Ideally, a normalization procedure would be consistent and effective on any continuous data, regardless of its generating distribution.

Many nonparametric normalization transformations purport to work well in cases where parametric approaches fail. Numerous authors have contributed to this topic already; see [3] for a comprehensive literature review on this subject. In fact, the first instance of a rank-based nonparametric normalization technique that we could find predates the Box-Cox transformation by over a decade; see [2,18]. However, most nonparametric techniques are limited to the observed range of the data, and this presents a substantive problem for the application of normalization techniques in the era of cross-validation and bootstrapping. In order for a transformation to be useful in such contexts, the technique should be easily trained on arbitrary subsets of the data and seamlessly applied to new data. Ideally, the normalization would not only be defined outside the bounds of the training data, but effective in its treatment of new data.

Based on this impetus, we introduce the Ordered Quantile (ORQ) normalization technique. ORQ is fundamentally based on a rank mapping of the observed data to the normal distribution, which *guarantees* normally distributed transformed data when ties are not present. Like all other normalization transformations, the ORQ transformation is reversible (i.e. one-to-one), which allows for straightforward interpretation; any analysis performed on the normalized data can be interpreted using the original units. (We illustrate this property in a subsequent application.) ORQ normalization combines this rank-mapping with a shifted logit approximation that allows the transformation to work effectively on data outside the original domain.

In this paper, we will compare and contrast the performance of the ORQ normalization transformation to that of Box-Cox, Lambert WxF of type  $s$ , Lambert WxF of type  $h$ , and Yeo-Johnson. In Section 2, we formally introduce Ordered Quantile normalization and explain its underlying intuition. In Section 3, we present simulations that show how and when ORQ normalization will be useful compared to these other methods. In Section 4, we apply QRQ normalization in a ‘transform-both-sides’ regression context and compare the results to those produced by generalized additive models and ordinary least squares, primarily focusing on predictive efficacy.

## 2. The ordered quantile normalization technique

The ORQ normalization procedure is a semiparametric approach that uses the original values of a sample, the corresponding ranks, interpolation, and nonlinear extrapolation in order to estimate a normalizing transformation function that can readily be applied to new data. The first step is a simple rank-mapping from the empirical distribution function for the original sample to the normal distribution function; the observed data is thereby *forced* to follow a normal distribution. For data outside the domain of the original data, a nonlinear parametric model extrapolates the transformation. New data that fall within the domain of the original data are transformed via interpolation.

More formally, let  $\mathbf{x}$  refer to the original data, a vector of length  $n$ , and let  $\mathbf{z}$  refer to the ranks of  $\mathbf{x}$  (conformably ordered). Let  $x_i$  and  $z_i$  refer to individual values within the vector  $\mathbf{z}$ , indexed by  $i$ . Also, let  $x^*$  refer to a new observation that may or may not be represented among the original  $\mathbf{x}$ .

We define  $f(x_i) = \Phi^{-1}((z_i - 1/2)/n)$ ; a slightly modified version of the rank-based mapping found in [18]. Noting that  $((z_i - 1/2)/n)$  provides an estimated percentile of  $x_i$ , one can interpret  $f(x_i)$  as the inverse normal distribution function evaluated at the estimated percentile of  $x_i$ . We will henceforth denote the sample percentile as  $((z_i - 1/2)/n)$  as  $p_i$ , where  $p_i \in (0, 1)$ . We will use  $\pi_i$  to denote the corresponding population percentile.

Note that  $f(a)$  is only defined when  $a$  is an already observed value, i.e.  $a \in \mathbf{x}$ . Finally, let  $x_l = \max\{a \in \mathbf{x} \mid a < x^*\}$  and  $x_u = \min\{a \in \mathbf{x} \mid a > x^*\}$ ; in other words,  $x_l, x_u$  refer to the closest points to  $x^*$  that appeared in the original data  $\mathbf{x}$ .

The ORQ normalization transformation is defined as follows:

$$g(x^* \mid \mathbf{x}) = \begin{cases} f(x^*) & \text{if } x^* \in \{\mathbf{x}\} \\ \frac{f(x_u) - f(x_l)}{x_u - x_l} & \text{if } x^* \notin \{\mathbf{x}\} \text{ and } \min \mathbf{x} < x^* < \max \mathbf{x} \\ r(x^*; \mathbf{x}) & \text{if } x^* < \min \mathbf{x} \text{ or } x^* > \max \mathbf{x} \end{cases}$$

Here  $r(x^*; \mathbf{x})$  is an extrapolation function that will be subsequently defined.

The function  $r(x^*; \mathbf{x})$  is determined first by fitting a generalized (logit-link) linear model with parameters  $\beta_0, \beta_1$  of the following structure:

$$\text{logit } (\pi_i) = \beta_0 + \beta_1 x_i$$

To fit this model, we employ an objective function based on the form of the log-likelihood for a logistic regression model arising from the binomial distribution:

$$\sum_{i=1}^n [(p_i n) (\beta_0 + \beta_1 x_i) - n \log(1 + \exp(\beta_0 + \beta_1 x_i))].$$

At first appearance, the preceding is a strange model: here, the logit of the population percentile corresponding to the original  $x_i$  is characterized by a linear form with  $x_i$  as the lone covariate. However, though the model may not provide a good fit for the data at hand, we have found that it provides a robust mechanism for estimating quantiles for values outside the original domain of  $\mathbf{x}$ , as will be seen in subsequent simulations.

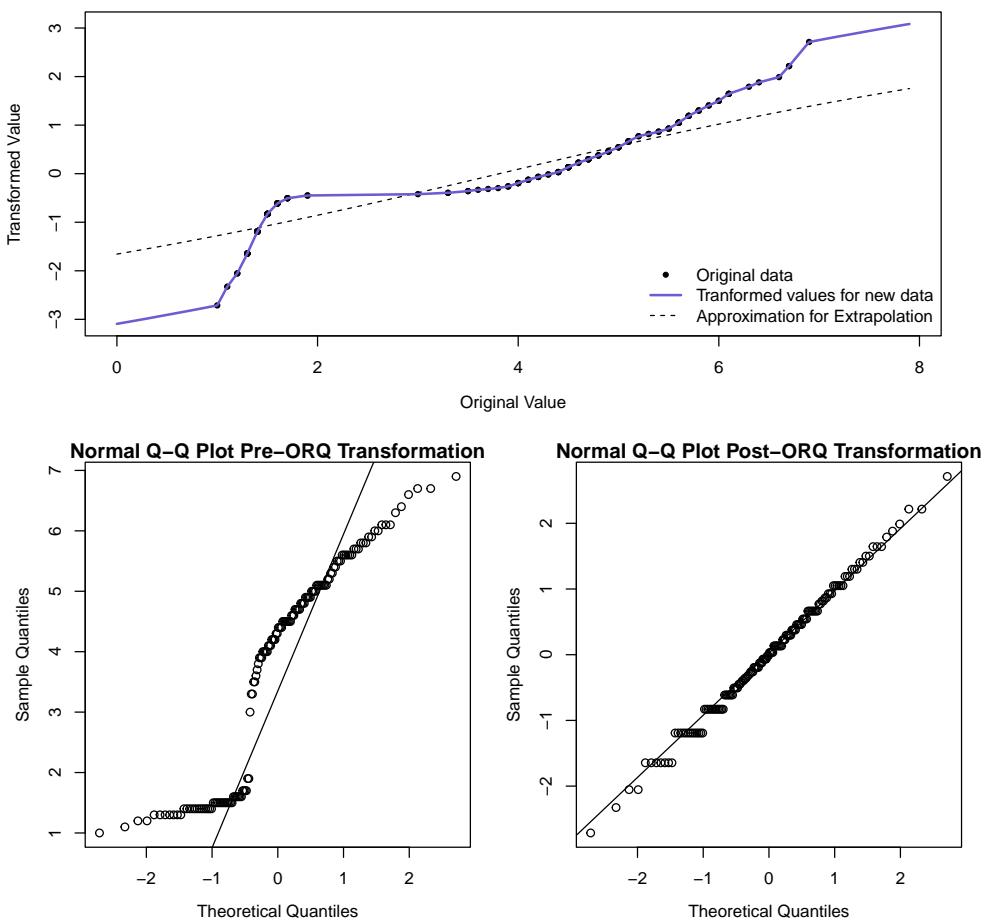
With this fitted model, we then use the estimate  $(\hat{\beta}_0, \hat{\beta}_1)$  to inform the ORQ extrapolations. If we let

$$l(a) = \Phi^{-1} \left( \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 a)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 a)} \right)$$

refer to the normal quantile of the model's prediction for a quantity  $a$ , we can then define  $r(x^*; \mathbf{x})$  as

$$r(x^*; \mathbf{x}) = \begin{cases} l(x^*) + \min_i[f(x_i)] - \min_i[l(x_i)] & \text{if } x^* < \min \mathbf{x} \\ l(x^*) + \max_i[f(x_i)] - \max_i[l(x_i)] & \text{if } x^* > \max \mathbf{x} \end{cases}$$

The preceding yields the ‘shifted’ logit approximation of the nonparametric transformation of the original data. The *shift* ensures that the transformation is smooth and one-to-one where the extrapolation function meets the original domain. The *logit* transformation increases the robustness of the transformation for new data points that must be



**Figure 1.** A visualization of ordered quantile normalization applied to petal length in Fisher's 'iris' data set.

extrapolated. As a side note, we also explored using a shifted linear extrapolation on the transformed values, but this was found to be highly inaccurate in cases where generating distributions had sufficiently heavy tails, such as the Cauchy distribution. The shifted logit approximation, applied to the percentiles, mitigates the effect of these extreme observations. By using every observation's rank in the estimation of the extrapolation function, the extrapolation becomes more robust to the variability present at the bounds of the domain.

Overall, the ORQ transformation can be viewed as semiparametric; it is nonparametric along the original domain of  $\mathbf{x}$ , but it is parametric outside the original domain of  $\mathbf{x}$ . However, the parametric component is constructed using the nonparametric component, in some sense 'borrowing strength' from it. Similar to other rank-based procedures, some information gets lost during the transformation process.

ORQ normalization is visualized in Figure 1 via Fisher's 'iris' data set. The top plot shows the transformation for the 'petal length' variable. Outside the bounds of the original domain of  $\mathbf{x}$  (in this case,  $x^* \notin [1, 6.9]$ ), we observe the extrapolation function  $r(x^*; \mathbf{x})$  as

a shift in the dotted line, which arises via the aforementioned logit model. Figure 1 also displays the normal Q-Q plots for the data both before and after the ORQ transformation.

The effect of uncertainty in the parametric component's approximation will typically be minimal since we do not expect to see many observations outside the observed range if the sample size is large enough (unless the tails of the distribution are very heavy). It should be noted that the ORQ technique will not guarantee normally transformed data in the presence of ties, but it could still yield the best normalizing transformation when compared to other alternatives. When a data vector has a large number of ties, normalization transformations (that are also one-to-one) cannot be effective since the transformed vector will also have a large number of ties.

The ORQ transformation can also be conceptualized as an approximation of a ‘true’ normalizing function. If  $X$  follows any distribution that has a transformation  $h(X)$  such that  $h(X) \sim N(0, 1)$ , then the ORQ transformation provides an approximation to  $h$  given a realized sample  $\mathbf{x}$ . This means that if a vector is right-skewed such that a log transformation truly normalizes it, then the ORQ transformation will approximate a log transformation. Similarly, if a vector is left-skewed such that an exponential transformation truly normalizes it, the ORQ transformation will approximate the exponential transformation. In fact, *if any normalizing transformation function exists, the ORQ transformation will approximate the function*. The approximation will improve as the amount of data used in the training of the transformation increases.

### 3. Simulations

In this section, we investigate the properties of various normalization transformation techniques when applied to various originating distributions. Specifically, we generate data from the following distributions:

$$x_1 \sim \Gamma(1, 1)$$

$$x_2 \sim 100 - \Gamma(1, 1)$$

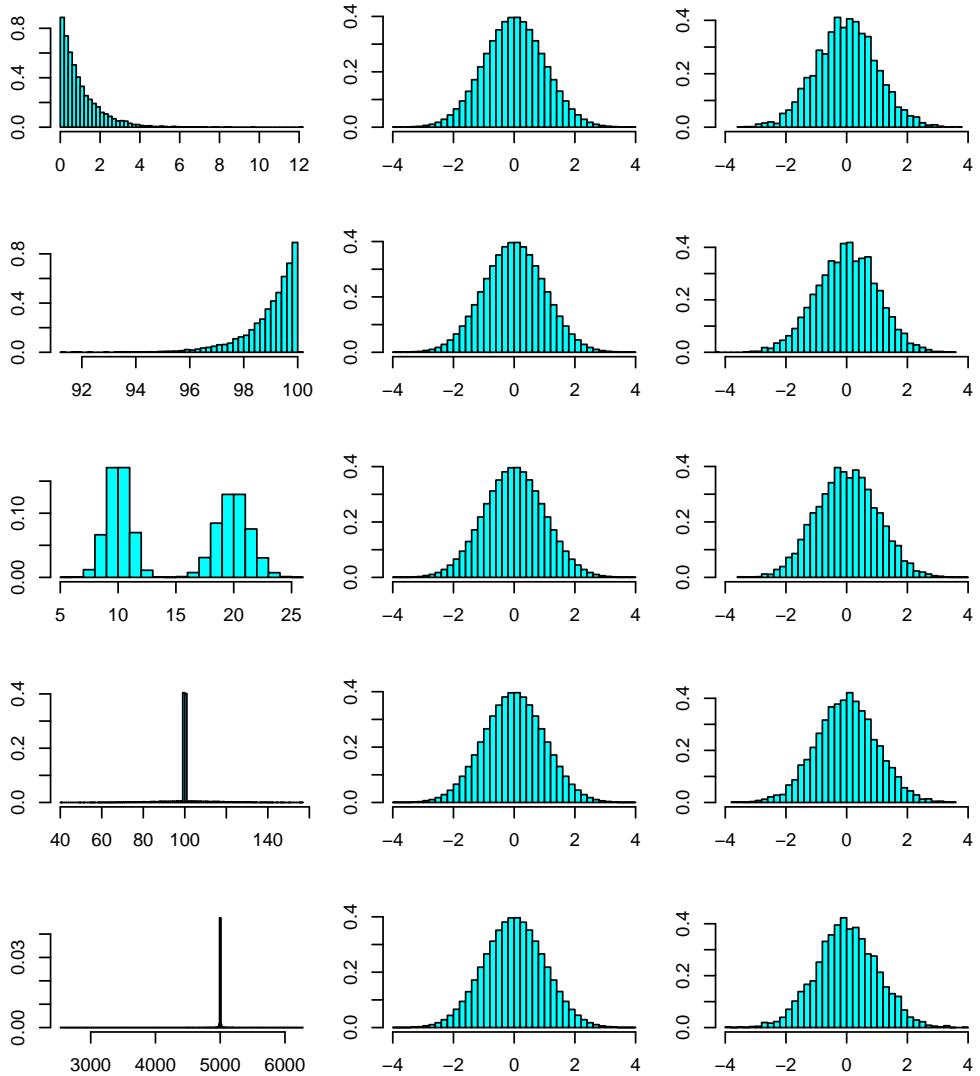
$$x_3 \sim N(10, 1) + \text{Bernoulli}(.5) * N(20, 1)$$

$$x_4 \sim p * N(100, .25) + (1 - p) * N(100, 16) \text{ where } p \sim \text{Bernoulli}(.8)$$

$$x_5 \sim \text{Cauchy}(\text{location} = 1000)$$

This set of generating distributions is quite broad (see Figure 2);  $x_1$  is right skewed,  $x_2$  is left-skewed,  $x_3$  is bimodal,  $x_4$  is a heavy-tailed mixture of normal variables, and  $x_5$  is a very heavy-tailed Cauchy distribution. The distributions are all shifted towards the positive domain, as the Box-Cox technique requires positive data. Although it is worth noting that positivity is an inherent limitation of the Box-Cox, shifting these distributions at least provides all of the techniques with a level playing field.

In order to determine the efficacy of the normalization transformations, we will use a variant of the Pearson's goodness-of-fit test statistic divided by the statistic's degrees of freedom [8]. This particular test statistic  $P$  follows a chi-squared distribution under the null hypothesis of normality. We divide  $P$  by its degrees of freedom to ensure that the result is interpretable; values close to one are ideal in that they show the least evidence of lack of



**Figure 2.** Candidate generating distributions (left), in-sample ORQ transformed values (center), and out-of-sample ORQ transformed values (right), with  $n = 10,000$  in both training and testing samples.

normality. Moreover, as the data of interest become increasingly non-normal, P/df grows accordingly.

Table 1 shows the average in-sample P/df statistic for normalization transformations for  $n = 1,000$  observations of each candidate distribution and across  $S = 10,000$  simulations. Curiously, ORQ normalization appears to be doing *too* well in these transformations – rendering the results somewhat suspicious. This is essentially because the method is ‘cheating’ – the transformed data are set to their exact theoretical normal quantiles. The nature of our transformation forces the Pearson’s normalization test statistic to be even less than one would expect under the null, as we have also stripped away any random deviations that would be inherent under the null hypothesis of normality.

**Table 1.** In-sample transformation efficacy measured by P/df on the original samples ( $n = 1,000$ ) after transformation. Values close to one indicate normally transformed data.

	Right-skewed	Left-skewed	Bimodal	Normal Mixture	Cauchy
Box-Cox	1.05	23.08	51.01	339.90	367.45
Lambert S	1.24	1.24	51.57	340.26	307.50
Lambert H	22.93	22.92	53.24	9.92	1.13
Yeo-Johnson	4.78	21.15	51.09	339.79	498.53
ORQ	0.01	0.01	0.01	0.01	0.01

A better platform for comparison would be to investigate how well the normalization works on newly observed data, which is exhibited in the subsequent tables. In Table 1, we also see that barring ORQ, no transformation works well for every type of originating distribution, even for the in-sample data.

Since comparing ORQ to other normalization transformations on in-sample metrics is unfair, Table 2 illustrates the transformation efficacy on newly generated data points for multiple sizes of training and test sets (the size of the test set is set to one tenth the size of the training set). Table 2 presents the average P/df statistic taken across 10,000 simulated test and training sets. For  $n = 10,000$ , we observe very similar results as Table 1 except for ORQ, which still performs very well but not suspiciously so. In fact, ORQ performs well on every candidate distribution, including the bimodal and the normal mixture distributions, which every other transformation function failed to normalize. Figure 2 displays a generated sample of size 10,000 from each distribution, their in-sample ORQ normalized values, and ORQ transformations of a newly generated sample of size 10,000.

As the sample size of the training set (and the test set) decreases, the ORQ transformation performs slightly worse as its approximation to the true normalizing transformation becomes less precise. This pattern indicates that as the sample size increases, the ORQ transformation will be more and more accurate and effective in producing a successful normalizing transformation for each of these distributions. The same pattern was found for each transformation that was effective for a particular distribution; as the sample size increases, the Lambert S transformation generally worked better for skewed distributions, the Lambert H transformation worked increasingly well for the Cauchy distribution, and the Box-Cox transformation worked increasingly well for the right-skewed generating distribution. Curiously, in situations where a transformation was not effective, higher sample sizes seem to reduce its efficacy. However, this phenomenon is due to the increased power of the Pearson test, and is not necessarily attributable to a worsening transformation in these settings; as the number of observations in the test set increases, departures from normality are going to increasingly push the P/df statistic away from 1, regardless of the estimation accuracy of the transformation itself.

We know that in-sample estimates are biased in favor of the ORQ normalization, but how are we to correct for this when the true distribution is unknown? We can use repeated cross-validation (CV) in this circumstance. Specifically, we split the data into  $k$  folds, then for each fold  $j$  we iteratively fit each normalizing transformation using the other  $k-1$  folds as a training set. We then calculate the Pearson's P/df for the transformed data in fold  $j$ , and repeat for all other  $j$  in  $k$ , taking the mean value as the best estimate of the transformation efficacy. Finally, we repeat this process  $r$  times to lessen the impact of the randomness in the partitioning of the folds, taking the mean across repeats as the final estimate of

**Table 2.** Out-of-sample transformation efficacy measured by P/df on newly generated samples for various sample sizes. The size of the test data is one tenth that of the training data. Values close to 1 indicate normally transformed data.

	Right-skewed	Left-skewed	Bimodal	Normal Mixture	Cauchy
<i>n</i> = 50					
Box-Cox	1.59	2.20	3.30	4.76	3.11
Lambert S	1.57	1.58	3.22	4.54	2.68
Lambert H	2.10	2.08	3.33	1.83	1.68
Yeo-Johnson	1.66	2.15	3.30	4.79	3.66
ORQ	1.71	1.74	1.72	1.80	1.88
<i>n</i> = 100					
Box-Cox	1.16	2.15	3.62	7.11	3.76
Lambert S	1.16	1.16	3.60	7.10	3.21
Lambert H	2.03	2.01	3.67	1.46	1.23
Yeo-Johnson	1.22	2.09	3.63	7.12	4.39
ORQ	1.28	1.29	1.29	1.30	1.37
<i>n</i> = 1000					
Box-Cox	1.04	6.14	13.76	58.12	34.66
Lambert S	1.06	1.05	13.94	58.23	30.62
Lambert H	5.21	5.19	14.20	2.63	1.08
Yeo-Johnson	1.45	5.77	13.81	58.10	45.57
ORQ	1.13	1.12	1.13	1.13	1.13
<i>n</i> = 10,000					
Box-Cox	1.07	23.13	51.07	346.00	433.92
Lambert S	1.24	1.24	51.60	345.88	402.67
Lambert H	22.67	22.71	53.28	9.78	1.08
Yeo-Johnson	4.82	21.20	51.15	346.00	462.30
ORQ	1.11	1.11	1.10	1.11	1.10

out-of-sample P/df. If the number of folds is selected such that the test/training ratio is close to what it is for a given problem, then the estimate of Pearson's P/df should be close to the true value. However, the accuracy of the ORQ transformation depends to an extent on the sample size used to train it, so cross-validation will underestimate the true efficacy of ORQ normalization compared to when it has been trained using the full sample (i.e. the CV estimated P/df statistic will tend to be higher than its value would be if ORQ had been trained using the whole sample).

In order to show that this CV method is effective whether or not the true generating distribution is known, in Table 3 we present results from a hybrid of the prior simulations and the application in the next section. The sample size for this simulation is set to 6,283, which is the same as the sample size in the application. We use five-fold cross-validation with

**Table 3.** Estimated out-of-sample normalization efficacy using five-fold cross-validation with five repeats (*n* = 6,283). Car price is included to show this method can be effective even when the true generating distribution is unknown; see Section 4 for more details.

	Right-skewed	Left-skewed	Bimodal	Normal Mixture	Cauchy	Car Price
Box-Cox	1.21	25.58	58.35	423.65	381.27	4.00
Lambert S	1.48	1.32	59.80	424.68	269.29	4.02
Lambert H	25.75	24.30	60.61	10.77	1.08	5.80
Yeo-Johnson	5.74	24.89	58.42	423.79	481.33	4.00
ORQ	1.30	1.21	1.28	1.19	1.15	1.26

five repeats to estimate the normalization efficacy of our candidate distributions (generating distributions we know), as well as ‘car price’ (a generating distribution that we do not know). By comparing the results from Tables 3 and 2, we find that repeated cross-validation yields the same out-of-sample efficacy results as we would see if we gathered a genuine new sample. This means that although we do not know the true generating model for car price, we can be confident that these efficacy statistics are representative of out-of-sample performance.

## 4. Application

The autotrader data set was scraped from the Autotrader website for inclusion with the `bestNormalize` package [15]. The variables include car mileage, price, and age (as well as model and make). We apply the `bestNormalize` package functionality, along with the ORQ transformation, to normalize mileage, age, and price. Finally, we build a pricing model using these transformed data.

### 4.1. The `bestNormalize` package

We produced the `bestNormalize` R package for several reasons. First, it facilitates the use of ORQ normalization as well as other transformations. Second, we have found that applied statistical problems frequently involve the need to normalize variables, but it can be difficult to assess a wide range of normalization techniques. Moreover, it is often unclear how to test the methods against one another; many practitioners will opt for the first method they uncover that seems to work. The `bestNormalize` framework allows users to easily compare the extra-sample normalization efficacy of many different candidate functions. With one line of code, users can investigate the Box-Cox, the Yeo-Johnson, the Lambert WxF, and ORQ transformations, as well as other more parsimonious transformations such as log, exponential, square-root, and hyperbolic arc sine.

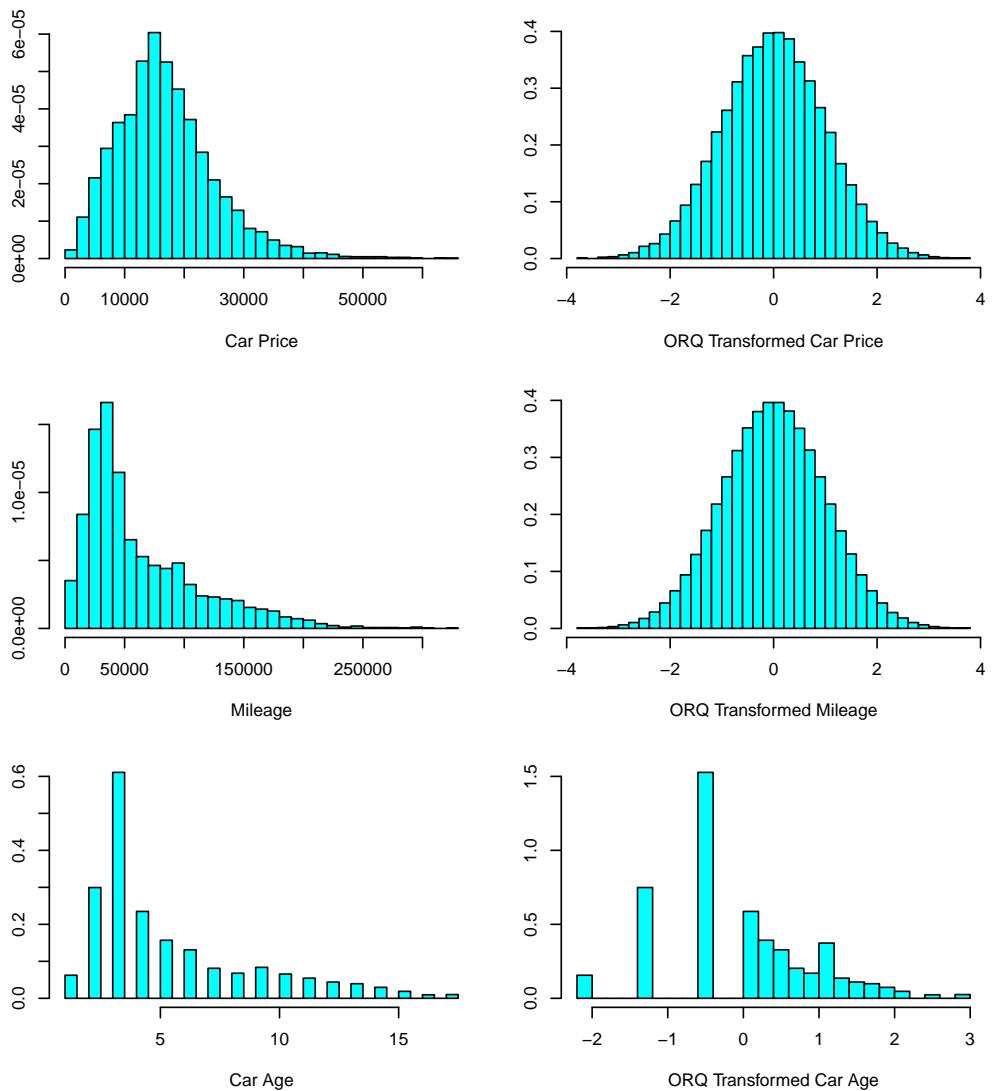
The `bestNormalize` R package is completely open source and is freely available on GitHub and CRAN. A detailed tutorial is available in the form of a package vignette on the CRAN website.

### 4.2. Estimating transformation functions for car price, mileage, and age

As evident in Figure 3, car price, mileage, and age are all highly right skewed. Furthermore, all three variables have ties in their distributions, which further complicates normalization. Using `bestNormalize`, we can test multiple possible transformations for normalization efficacy. In Table 4, we use five-fold cross-validation with five repeats to estimate the normalization efficacy for a set of transformations.

We see that the estimated normality statistics for the ORQ transformation are close to one for both price and mileage, despite the presence of ties in the data set. ORQ is performing considerably better than all of the other candidate transformations. Interestingly, a square-root transformation performed almost as well in normalizing car price as did the Box-Cox, Yeo-Johnson, or Lambert S transformations.

However, age was not very well normalized by any candidate transformation. Age has many more ties than price and mileage, and this makes it very difficult to find a normalizing



**Figure 3.** Distributions of pre- and post- ORQ transformations of car price, mileage, and age.

transformation (see Figure 3). Even so, the ORQ transformation has the lowest estimated P/df statistic.

#### 4.3. Fitting a model on the transformed data

With the normally distributed transformed data, we can fit a linear model using the transformed values of each variable. This process is referred to as ‘transform-both-sides’ regression. We will then be able to use the reverse transformations to assess the relationships among the variables in terms of their original units. There are some purported benefits of this type of modeling approach; see [9]. Specifically, in many applications, a

**Table 4.** Estimated extra-sample normalization statistics for autotrader data (five folds, five repeats).

	Price	Mileage	Age
Arc-Sinh	6.64	4.56	170.63
Box-Cox	4.04	4.11	170.72
Lambert H	6.16	20.93	172.08
Lambert S	4.11	4.12	169.39
Log	6.64	4.55	170.25
No Transform	6.59	22.09	171.56
ORQ	1.27	1.15	167.83
Square-root	4.21	7.14	170.60
Yeo-Johnson	4.04	4.11	171.87

simple marginal normalization for both the outcome and the covariates will lead to approximate normality of the residuals. Assuming that the residuals are also uncorrelated and homoscedastic, the actual coverage of confidence and prediction intervals should be close to the nominal level.

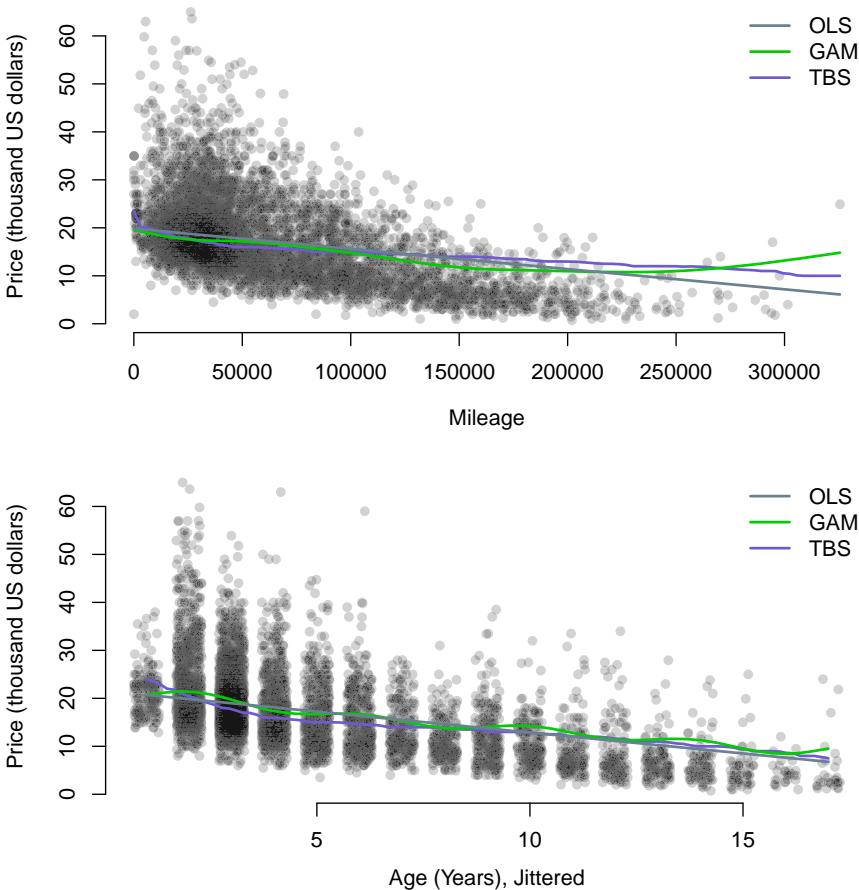
We will investigate the results from the transform-both-sides (TBS) regression compared to alternative modeling strategies in the next section.

It should be noted that with the ORQ transformation, there is no universally appropriate interpretation of the coefficients on the covariates, since the transformation function itself will be characteristically different for various (generating) distributions. Furthermore, when the ORQ transformation is employed, the scale of the original units is not preserved; only the ordering of the observations is preserved. However, the reverse ORQ transformation can easily be applied, allowing us to characterize covariate effects on the scale of the original units. While such an approach does not yield an interpretation of the fitted model's specific coefficient estimates, it does provide a sense of the shape of the relationship between covariates and the outcome. In a sense, the setting is akin to that of fitting a generalized additive model (GAM), where it is difficult to interpret each coefficient, and a picture is worth a thousand words.

Thus, in order to adequately interpret the model, we produce plots of the predictions across all of the observed values of each covariate. For the sake of illustration, we also plot the predictions resulting from fitting a generalized additive model and a simple ordinary least squares model (Figure 4). We see that both mileage and age decrease the predicted car price (which is unsurprising). The magnitude of the effect is highest for low values of mileage and age, and smallest on high values. Interestingly, the effect of mileage on price in the GAM is not monotonic; the price is expected to increase past 200,000 miles. This is odd, but understandable in the context of influential points (e.g. there was a 2012 GMC Sierra with 325,000 miles listed for sale at \$24,900).

#### 4.4. Comparing models

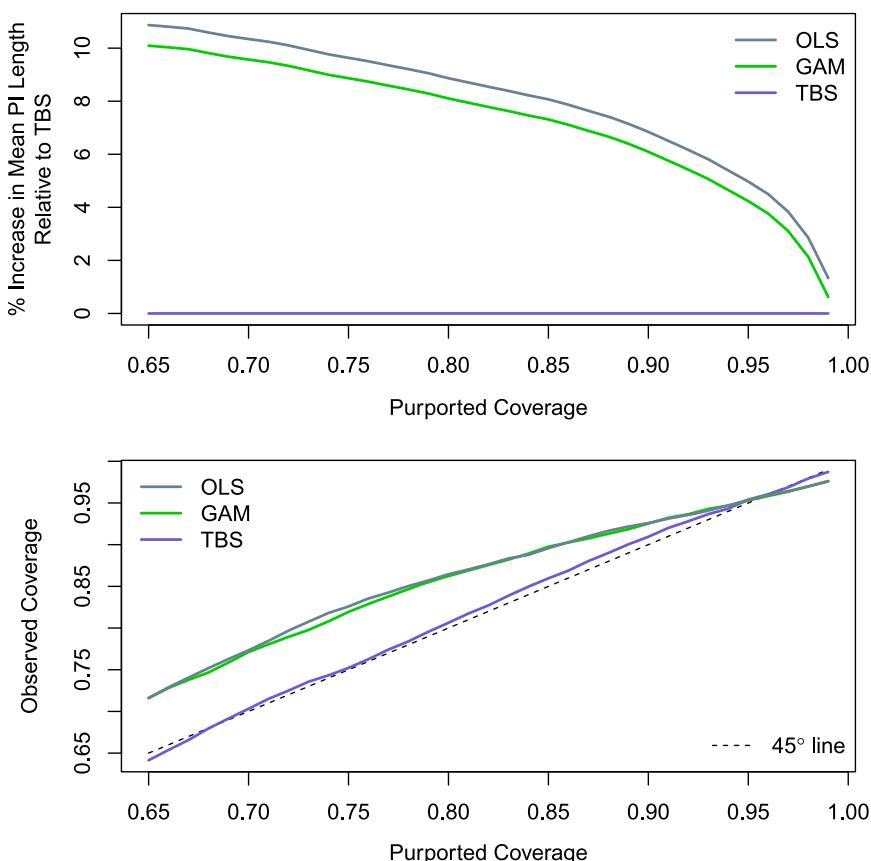
In order to compare the TBS regression with the GAM and the OLS model, we investigate the leave-one-out predictions and their corresponding prediction intervals. Specifically, for each car listing in the data set, all of the ORQ transformations were re-trained without that observation, as were the TBS, GAM, and OLS models. Then, each model was used to predict the omitted car price and to derive a corresponding prediction interval. We considered



**Figure 4.** Effects of mileage (top) and age (bottom) on car price for generalized additive models (GAM), ordinary least squares (OLS) models, and transform-both-sides (TBS) regression, where the ORQ transformation is used on car price, mileage, and age.

prediction intervals with multiple levels of coverage (see Figure 5 and Table 5). The prediction intervals for the TBS model had observed coverage that was much closer to nominal compared to the other two methods, which were too conservative. This suggests that the residuals for the TBS model were closer to normally distributed than the residuals for the other two methods; indeed, this is what we observe to be empirically evident. In the full TBS model, the P/df statistic for the residuals is 10.5, compared to 21.5 and 22.2 for the GAM and OLS fits, respectively. We also find that the mean prediction interval (PI) length was lowest for the TBS model for all levels of coverage. In other words, the TBS model yields more precise prediction than the other two methods.

Note that the reduction in the length of the prediction intervals for TBS models is not similar for different levels of the covariates. In the TBS model, newer cars (and those with less mileage) have wider prediction intervals than for older, higher-mileage cars. This reflects the right-skewed nature of the originating distribution of price; we should expect pricier cars to have greater variability.



**Figure 5.** Prediction interval length (top) and coverage (bottom) in generalized additive models (GAM) and ordinary least squares (OLS) models compared to transform-both-sides (TBS) regression, where the ORQ transformation is used on the outcome and the covariates. Intervals were calculated and evaluated with leave-one-out cross-validation.

**Table 5.** Prediction interval (PI) performance for transform-both-sides (TBS) regression, compared to generalized additive models (GAM) and ordinary least squares (OLS). Intervals were calculated and evaluated with leave-one-out cross-validation.

Nominal Level (%)	Empirical Coverage			Mean PI Length		
	(%)			(\$1000)		
	TBS	GAM	OLS	TBS	GAM	OLS
75	75.2	81.9	82.6	14.3	15.5	15.6
80	80.6	86.2	86.5	16.0	17.3	17.4
85	86.0	89.8	89.6	18.1	19.4	19.6
90	90.9	92.5	92.6	20.9	22.2	22.4
95	95.4	95.3	95.4	25.4	26.5	26.7

## 5. Discussion

We have found the ORQ transformation to be a remarkably effective normalization technique. For any candidate generating distribution, the ORQ transformation will approximate the best normalizing transformation (if one exists). We applied ORQ normalization in the context of ‘transform-both-sides’ regression in order to predict the price of a car based on its mileage and age. Through this approach, we were able to reduce the width of the prediction intervals while retaining close to nominal coverage, obtaining superior results compared to those arising from generalized additive and ordinary least squares fits.

We envision several avenues of future extensions to this work. As we noted in the introduction, it may be worthwhile to use such a normalization procedure as a means of automatically preprocessing features in a model selection context in order to reduce the leverage of potentially influential points among candidate predictors. In this context, one can pair the ORQ transformation with a robust procedure for assessing departures from normality, such as the methods proposed in [17]; variables could be screened as to whether or not a transformation is merited in the first place, and only those judged to be sufficiently non-normal would undergo a transformation. Additionally, while we investigated several possible generating distributions, further exploration of the efficacy of the ORQ transformation could be performed using the p-outliers model proposed originally in [1].

Finally, we produced and utilized the `bestNormalize` R package, which can perform ORQ normalization and compare its out-of-sample performance to a suite of other candidate normalization techniques.

## Acknowledgements

We wish to thank the referees for their valuable feedback, which served to improve the original version of this manuscript.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Ryan A. Peterson  <http://orcid.org/0000-0002-4650-5798>

Joseph E. Cavanaugh  <http://orcid.org/0000-0002-0514-7664>

## References

- [1] N. Balakrishnan, *Permanents, order statistics, outliers, and robustness*, Rev. Mat. Complut. 20 (2007), pp. 7–107.
- [2] M.S. Bartlett, *The use of transformations*, Biometrics 3 (1947), pp. 39–52. Available at <http://www.jstor.org/stable/3001536>.
- [3] T.M. Beasley, S. Erickson, and D.B. Allison, *Rank-based inverse normal transformations are increasingly used, but are they merited?*, Behav. Genet. 39 (2009), pp. 580–595. Available at <https://doi.org/10.1007/s10519-009-9281-0>.
- [4] P.J. Bickel and K.A. Doksum, *An analysis of transformations revisited*, J. Am. Stat. Assoc. 76 (1981), pp. 296–311. Available at <https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1981.10477649>.

- [5] G.E.P. Box and D.R. Cox, *An analysis of transformations*, J. R. Stat. Soc. Ser. B 26 (1964), pp. 211–252. Available at <http://www.jstor.org/stable/2984418>.
- [6] G.E. Box and P.W. Tidwell, *Transformation of the independent variables*, Technometrics 4 (1962), pp. 531–550. Available at <http://www.tandfonline.com/doi/abs/10.1080/00401706.1962.10490038>.
- [7] G.M. Goerg, *Lambert w random variables-a new family of generalized skewed distributions with applications to risk estimation*, Ann. Appl. Stat. 5 (2011), pp. 2197–2230. Available at <https://doi.org/10.1214/11-AOAS457>.
- [8] J. Gross and U. Ligges, *nortest: Tests for normality* (2015). Available at <https://CRAN.R-project.org/package=nortest>, R package version 1.0-4.
- [9] F. Harrell, *Regression Modeling Strategies*, Springer, New York, 2015.
- [10] J.A. John and N.R. Draper, *An alternative family of transformations*, J. R. Stat. Soc. Ser. C 29 (1980), pp. 190–197. Available at <http://www.jstor.org/stable/2986305>.
- [11] M. Kuhn, *caret: Classification and regression training* (2017). Available at <https://CRAN.R-project.org/package=caret>, R package version 6.0-78.
- [12] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, New York, 2013.
- [13] M. Kuhn and H. Wickham, *recipes: Preprocessing tools to create design matrices* (2018). Available at <https://CRAN.R-project.org/package=recipes>, R package version 0.1.2.
- [14] B.F.J. Manly, *Exponential data transformations*, J. R. Stat. Soc. Ser. D 25 (1976), pp. 37–42. Available at <http://www.jstor.org/stable/2988129>.
- [15] R.A. Peterson, *bestNormalize: A suite of normalizing transformations* (2019). Available at <https://github.com/petersonR/bestNormalize>, R package version 3.6.0.
- [16] M. Stehlík and P. Hermann, *Letter to the editor*, Ann. Appl. Stat. 9 (2015), p. 2051. Available at <https://doi.org/10.1214/15-AOAS864>.
- [17] M. Stehlík, L. Střelec, and M. Thulin, *On robust testing for normality in chemometrics*, Chemometr. Intell. Lab. Syst. 130 (2014), pp. 98–108.
- [18] B. Van der Waerden, *Order tests for the two-sample problem and their power*, in *Indagationes Mathematicae (Proceedings)*, Vol. 55, Elsevier, Amsterdam, 1952, pp. 453–458.
- [19] I. Yeo and R.A. Johnson, *A new family of power transformations to improve normality or symmetry*, Biometrika 87 (2000), pp. 954–959. Available at <http://dx.doi.org/10.1093/biomet/87.4.954>.