Biometrika (2000), **87,** 4, *pp.* 954–959

Printed in Great Britain

A new family of power transformations to improve normality or symmetry

BY IN-KWON YEO

Department of Control & Instrumentation of Engineering, Kangwon National University, Chunchon, 200-701, Korea

inkwon@multimedia.kangwon.ac.kr

AND RICHARD A. JOHNSON

Department of Statistics, University of Wisconsin-Madison, Wisconsin 53706, U.S.A

rich@stat.wisc.edu

SUMMARY

We introduce a new power transformation family which is well defined on the whole real line and which is appropriate for reducing skewness and to approximate normality. It has properties similar to those of the Box–Cox transformation for positive variables. The large-sample properties of the transformation are investigated in the contect of a single random sample.

Some key words: Kullback–Leibler information; Maximum likelihood inference; Power transformation; Relative skewness.

1. INTRODUCTION

A major step towards an objective way of determining a transformation was made by Box & Cox (1964). The Box–Cox transformation $\psi^{\mathrm{BC}}(\lambda, x)$ is given by

$$\psi^{\mathrm{BC}}(\lambda, x) = \begin{cases} (x^{\lambda} - 1)/\lambda & (\lambda \neq 0), \\ \log(x) & (\lambda = 0), \end{cases}$$

for positive x. They considered selecting transformations for achieving approximate normality. The Box–Cox transformation is, however, only valid for positive x. Although a shift parameter can be introduced to handle situations where the response is negative but bounded below, the standard asymptotic results of maximum likelihood theory may not apply since the range of the distribution is determined by the unknown shift parameter; see Atkinson (1985, pp. 195–9). To circumvent these problems, some statisticians consider the signed power transformation, see for instance Bickel & Doksum (1981),

$$\psi^{\mathrm{SP}}(\lambda, x) = \{\mathrm{sgn}(x)|x|^{\lambda} - 1\}/\lambda \quad (\lambda > 0),$$

which covers the whole real line. Since ψ^{SP} is, however, designed to handle kurtosis rather than skewness, it has a serious drawback when it is applied to a skewed distribution. For instance, suppose X has the mixture density

$$f(x) = 0{\cdot}3\phi(x) + 0{\cdot}7\gamma(x), \tag{1·1}$$

where $\phi(.)$ is the standard normal density and $\gamma(.)$ is the gamma density

$$\gamma(x) = \frac{1}{6}(x+2)^3 \exp\{-(x+2)\} \quad (x > -2).$$

We are interested in transforming the random variable X so that the transformed distribution is approximately normal. Following Hernandez & Johnson (1980), we select ψ^{SP} to minimise the Kullback–Leibler information number

$$\int g_\lambda(u) \log \left\{ \frac{g_\lambda(u)}{\phi_{\mu,\sigma^2}(u)} \right\} du, \tag{1·2}$$

where $g_\lambda(.)$ is the probability density function of the transformed variable $\psi^{SP}(\lambda, X)$ and $\phi_{\mu,\sigma^2}(.)$ is the probability density function of a normal distribution with mean μ and variance σ^2. The minimum is over a suitable range of λ, μ and σ^2, and then the best choice of λ produces a g_λ that is closest, in the sense of Kullback–Leibler information, to a target normal density ϕ_{μ,σ^2}.

Even though ψ^{SP} is increasing in x, its Jacobian changes from a decreasing function of x to an increasing function as x changes sign. The cusp occurs at $\psi^{SP}(\lambda, 0) = -1/\lambda$, so the transformed density is bimodal and looks far from normal, as shown in Fig. 1. Most extended transformations, including the modulus transformation introduced by John & Draper (1980), lead to a bimodal distribution in our example where the support consists of the whole real line. A new family of transformations is therefore needed.

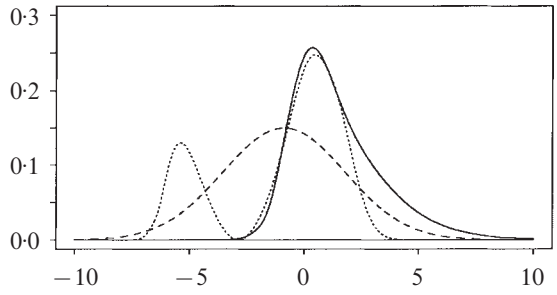

Fig. 1. Plots of the mixture density $f(.)$ (solid line), the transformed density $g_\lambda(.)$ based on ψ^{SP} (dotted line) and the target normal density $\phi_{\mu,\sigma^2}(.)$ (dashed line).

2. The new transformation

When searching for transformations that improve the symmetry of skewed data or distributions, it is helpful to recall the concept of relative skewness introduced by van Zwet (1964, p. 3). To motivate the definition, let X be a random variable having a continuous distribution function F with inverse F^{-1}, and let I_F be the smallest interval for which $\mathrm{pr}(X \in I_F) = 1$. Then the distribution F is said to be symmetric about x_0 if $F(x_0 + x) + F(x_0 - x) = 1$ and all real $x \in I_F$. Let Y be another random variable with a continuous distribution function G and inverse G^{-1}. Define $\psi(x) = G^{-1}\{F(x)\}$. Then G is the distribution function of $\psi(X)$ so the random variable $\psi(X)$ has the same distribution as the random variable Y. Van Zwet (1964) shows that $G^{-1}F$ is convex, respectively concave, if and only if G is the distribution function of a nondecreasing convex, respectively concave, transformation $\psi(X)$ of X. He then defines relative skewness, as follows.

Definition. *The distribution function G is more right-skewed, respectively more left-skewed, than the distribution F if $G^{-1}\{F(.)\}$ is a nondecreasing convex, respectively concave, function.*

Since a nondecreasing convex, respectively concave, transformation of a random variable effects a contraction of the lower, respectively upper, part of the support and an extension of the upper, respectively lower, part, it decreases the skewness to the left, respectively right. The Box–Cox transformation, for example, is concave in x for $\lambda < 1$ and convex in x for $\lambda > 1$. However, ψ^{SP} changes from convex to concave as x changes sign so it is not to be recommended when data that

can be positive or negative are skewed. John & Draper (1980) and Burbidge, Magee & Robb (1988) studied specific cases of other convex-to-concave transformations.

To motivate our choice of power transformations, we first consider a modified modulus transformation which has different transformation parameters on the positive and negative line. Let

$$\psi(\lambda_+, \lambda_-, x) = \begin{cases} \{(x+1)^{\lambda_+} - 1\}/\lambda_+ & (x \geqslant 0, \lambda_+ \neq 0), \\ \log(x+1) & (x \geqslant 0, \lambda_+ = 0), \\ -\{(-x+1)^{\lambda_-} - 1\}/\lambda_- & (x < 0, \lambda_- \neq 0), \\ -\log(-x+1) & (x < 0, \lambda_- = 0). \end{cases}$$

Next, we impose the condition that the second derivative $\partial^2\psi(\lambda_+, \lambda_-, x)/\partial x^2$ be continuous at $x = 0$. This forces the transformation to be smooth and implies that $\lambda_+ + \lambda_- = 2$. Consequently, we define the power transformation, $\psi(.,.): R \times R \to R$, where

$$\psi(\lambda, x) = \begin{cases} \{(x+1)^{\lambda} - 1\}/\lambda & (x \geqslant 0, \lambda \neq 0), \\ \log(x+1) & (x \geqslant 0, \lambda = 0), \\ -\{(-x+1)^{2-\lambda} - 1\}/(2-\lambda) & (x < 0, \lambda \neq 2), \\ -\log(-x+1) & (x < 0, \lambda = 2). \end{cases} \tag{2·1}$$

Then, by Lemma 1 below, $\psi(\lambda, x)$ is concave in x for $\lambda < 1$ and convex for $\lambda > 1$. Here, the constant 1 in parentheses makes the transformed value have the same sign as the original value, and allows us to prove Lemma 1 by working separately with the positive and negative domain. It also reduces ψ to the identity transformation for $\lambda = 1$.

Figure 2 shows the differences between the Box–Cox transformations, $(x^{\lambda} - 1)/\lambda$, and the new transformations. In fact, the new transformations on the positive line are equivalent to the generalised Box–Cox transformations, $\{(x+1)^{\lambda} - 1\}/\lambda$, for $x > -1$, where the shift constant 1 is included. We also see from Fig. 2 that, if the sign of x is changed, so that a right-skewed distribution becomes left-skewed, or the reverse, then the value of λ is replaced by $2 - \lambda$. We first establish properties of transformation (2·1).

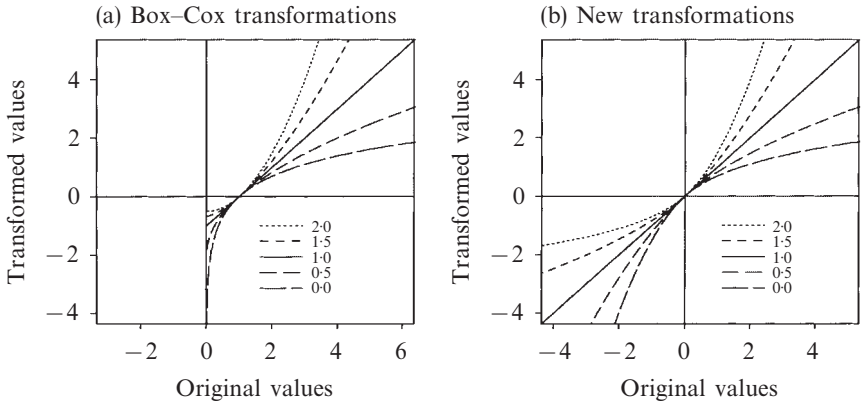

Fig. 2. A comparison of (a) the Box–Cox and (b) the new transformations, with $\lambda = 0{\cdot}0$, 0·5, 1·0, 1·5 and 2·0.

LEMMA 1. *The transformation function* $\psi(.,.)$ *defined in* (2·1) *satisfies the following*:

(i) $\psi(\lambda, x) \geqslant 0$ *for* $x \geqslant 0$, *and* $\psi(\lambda, x) < 0$ *for* $x < 0$;

(ii) $\psi(\lambda, x)$ *is convex in* x *for* $\lambda > 1$ *and concave in* x *for* $\lambda < 1$;

(iii) $\psi(\lambda, x)$ *is a continuous function of* (λ, x);

(iv) *if* $\psi^{(k)} = \partial^k \psi(\lambda, x)/\partial \lambda^k$ *then, for* $k \geqslant 1$,

$$\psi^{(k)} = \begin{cases} [(x+1)^{\lambda}\{\log(x+1)\}^k - k\psi^{(k-1)}]/\lambda & (\lambda \neq 0, x \geqslant 0), \\ \{\log(x+1)\}^{k+1}/(k+1) & (\lambda = 0, x \geqslant 0), \\ -[(-x+1)^{2-\lambda}\{-\log(-x+1)\}^k - k\psi^{(k-1)}]/(2-\lambda) & (\lambda \neq 2, x < 0), \\ \{-\log(-x+1)\}^{k+1}/(k+1) & (\lambda = 2, x < 0), \end{cases}$$

is continuous in (λ, x);

(v) $\psi(\lambda, x)$ *is increasing in both* λ *and* x;

(vi) $\psi(\lambda, x)$ *is convex in* λ *for* $x > 0$ *and concave in* λ *for* $x < 0$.

Note that $\psi^{(0)} \equiv \psi(\lambda, x)$.

The proofs are straightforward but tedious. Details are given in a University of Wisconsin technical report by the authors.

The mixture density $f(.)$ in (1·1) is skewed to the right so, according to van Zwet (1964), a good transformation should be concave in order to lead to near symmetry. Following Hernandez & Johnson (1980), we show in § 3 that it is reasonable to select the transformation, $\psi(\lambda, x)$, to minimise the Kullback–Leibler information number (1·2). By numerical integration, we obtain $\lambda = 0{\cdot}555$. As expected, this transformation is concave; see (ii) of Lemma 1. Figure 3 shows that the normal approximation is much improved since the transform has pulled down the right tail and pushed out the left tail.

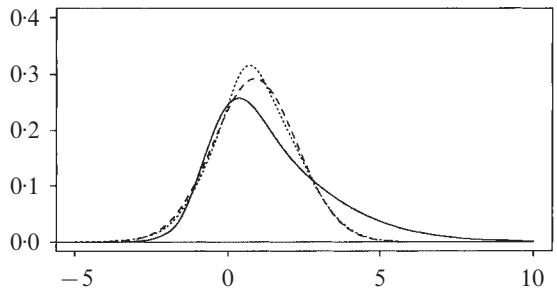

Fig. 3. Plots of the mixture density $f(.)$ (solid line), the transformed density $g_{\lambda}(.)$ based on ψ (dotted line) and the target normal density $\phi_{\mu,\sigma^2}(.)$ (dashed line).

3. Transforming to near normality

In this section we focus our attention on transforming a random sample from a parent distribution, with probability density function $f(.)$, to near normality. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables and denote the transformed variables by $\psi(\lambda, X_1), \ldots, \psi(\lambda, X_n)$. We assume that, for some λ, the transformed observations can be treated as normally distributed with some mean μ and variance σ^2. Under this assumption, the loglikelihood function is

$$l_n(\theta \mid x) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{\psi(\lambda, x_i) - \mu\}^2 + (\lambda - 1)\sum_{i=1}^{n}\operatorname{sgn}(x_i)\log(|x_i| + 1), \tag{3·1}$$

where $\theta = (\lambda, \mu, \sigma^2)'$ and $x = (x_1, \ldots, x_n)'$.

Holding λ fixed, we initially maximise $l_n(\lambda, ., .\,|\,x)$, yielding

$$\hat{\mu}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \psi(\lambda, x_i), \quad \hat{\sigma}^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \{\psi(\lambda, x_i) - \hat{\mu}(\lambda)\}^2. \tag{3·2}$$

The maximum likelihood estimate, $\hat{\lambda}$, of λ is obtained by maximising the profile loglikelihood function and then $\hat{\theta} = (\hat{\lambda}, \hat{\mu}(\hat{\lambda}), \hat{\sigma}^2(\hat{\lambda}))'$ maximises the loglikelihood function (3·1).

Under certain regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is a strongly consistent estimator of θ_0 which minimises the Kullback–Leibler information given by (1·2). Let

$$\nabla l_1(\theta_0\,|\,X) = \left(\frac{\partial}{\partial\theta_i} l_1(\theta\,|\,X)\bigg|_{\theta=\theta_0}\right)$$

be the gradient of the loglikelihood function of one observation for $\theta = (\theta_1, \theta_2, \theta_3)' = (\lambda, \mu, \sigma^2)'$, and let

$$\nabla^2 l_1(\theta_0\,|\,X) = \left(\frac{\partial^2}{\partial\theta_i\,\partial\theta_j} l_1(\theta\,|\,X)\bigg|_{\theta=\theta_0}\right)$$

be the Hessian of the loglikelihood function. Then $n^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically normal with mean 0 and covariance matrix $\Sigma(\theta_0) = V(\theta_0)W(\theta_0)V(\theta_0)'$, where

$$V(\theta_0) = E_f\{\nabla^2 l_1(\theta_0\,|\,X)\}^{-1}, \quad W(\theta_0) = E_f[\nabla l_1(\theta_0\,|\,X)\{\nabla l_1(\theta_0\,|\,X)\}'].$$

The details of the regularity conditions are given in the authors' technical report.

The assumption of homogeneity of the variance has been considered to be particularly important in many applications. However, the variance is often represented by a simple function of the mean. In practice, this relationship between mean and variance plays the major role in determining the transformation. Bartlett (1947) claims that a variance stabilising transformation 'often has the effect of improving the closeness of the distribution to normality'; his justification is based on 'correlation of variability with mean level on the original scale often implying excessive skewness which tends to be eliminated after the transformation'.

Let $\mathrm{var}(X) = \sigma^2(\mu)$ be a function of the mean. Without loss of generality, we assume that $\sigma^2(\mu)$ is an increasing function of μ. For the new transformation, we approximate the variance of the transformed variable, $\mathrm{var}\{\psi(\lambda, X)\} \doteq \sigma^2(\mu)h(\lambda, \mu)$, where

$$h(\lambda, \mu) = \begin{cases} (\mu+1)^{2(\lambda-1)} & (\mu \geqslant 0), \\ (-\mu+1)^{2(1-\lambda)} & (\mu < 0). \end{cases}$$

Since $h(\lambda, \mu)$ is decreasing in μ for $\lambda < 1$, we may choose a λ corresponding to variance stabilisation so that the variance becomes nearly constant on the transformed scale.

4. EXAMPLE

Darwin (1876) studied the effect of cross- and self-fertilisation on the growth of plants. Fifteen pairs of seedlings of the same age, one produced by cross-fertilisation and the other by self-fertilisation, were grown together so that the members of each pair were reared under nearly identical conditions. His aim was to demonstrate the greater vigour of the cross-fertilised plants. The differences between the final heights of plants in each pair after a fixed period of time were 6·1, −8·4, 1·0, 2·0, 0·7, 2·9, 3·5, 5·1, 1·8, 3·6, 7·0, 3·0, 9·3, 7·5 and −6·0.

The paired t-statistic is 2·142 ($p = 0{\cdot}025$), so the data support Darwin's claim at the 5% level of significance. However, both the Q–Q plot and the sample skewness statistic, $\sqrt{b_1} = 4{\cdot}713$, cast doubt on the normality assumption required for the paired t-test and indicate the data to be skewed to the left.

When the new transformation is applied, the parameter estimates for θ and the corresponding

estimated covariance matrix are

$$\begin{pmatrix} \hat{\lambda} \\ \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} 1{\cdot}305 \\ 4{\cdot}570 \\ 29{\cdot}786 \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} 0{\cdot}434 & & \\ 3{\cdot}270 & 54{\cdot}451 & \\ 26{\cdot}282 & 198{\cdot}227 & 3367{\cdot}608 \end{pmatrix}.$$

The likelihood ratio χ^2_1 statistic for $\lambda = 1$ is 3·873 ($p = 0{\cdot}0499$). The sample skewness statistic of the transformed values is $\sqrt{b_1} = 0{\cdot}093$ and the Shapiro–Wilk statistic is $W = 0{\cdot}975$ ($p = 0{\cdot}887$), showing that the normality of transformed data is much improved.

Let $t^{-1}(\gamma, n-1)$ be the γth quantile of the t distribution with $(n-1)$ degrees of freedom. Then the γth quantile of the mean difference can be approximated by

$$\hat{q}_\gamma = \psi^{-1}\{\hat{\lambda}, \hat{\mu}(\hat{\lambda}) + t^{-1}(\gamma, n-1)\hat{\sigma}(\hat{\lambda})/n^{1/2}\};$$

see Carroll & Ruppert (1991). In this example, the estimated 0·01th quantile is $\hat{q}_{0{\cdot}01} = 0{\cdot}790$. Since this is not negative, we strongly conclude that the cross-fertilised plants grow with the greater vigour, on average.

References

Atkinson, A. C. (1985). *Plots, Transformations and Regression*. Oxford: Oxford University Press.

Bartlett, M. S. (1947). The use of transformations. *Biometrics* **3**, 39–52.

Bickel, P. J. & Doksum, K. A. (1981). An analysis of transformations revisited. *J. Am. Statist. Assoc.* **76**, 296–311.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with Discussion). *J. R. Statist. Soc.* B **26**, 211–52.

Burbidge, J. B., Magee, L. & Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *J. Am. Statist. Assoc.* **83**, 123–7.

Carroll, R. J. & Ruppert, D. (1991). Prediction and tolerance intervals with transformation and/or weighting. *Technometrics* **33**, 197–210.

Darwin, C. (1876). *The Effect of Cross- and Self-fertilization in the Vegetable Kingdom*, 2nd ed. London: John Murray.

Hernandez, F. & Johnson, R. A. (1980). The large-sample behavior of transformations to normality. *J. Am. Statist. Assoc.* **75**, 855–61.

John, J. A. & Draper, N. R. (1980). An alternative family of transformations. *Appl. Statist.* **29**, 190–7.

van Zwet, W. R. (1964). *Convex Transformations of Random Variables*. Amsterdam: Mathematisch Centrum.

[*Received November* 1998. *Revised July* 2000]