Journal of the Korean Statistical Society (2001), **30**: *4, pp 563–571*

Selecting a Transformation to Reduce Skewness

In-Kwon Yeo[1]

ABSTRACT

In this paper, we study selecting a transformation so that the transformed variable is nearly symmetrically distributed. The large sample properties of an M-estimator of transformation parameter that is obtained by minimizing the integrated square of the imaginary part of the empirical characteristic function are investigated when a random sample is selected from some unspecified distribution. According to influence function calculations and Monte Carlo simulations, these estimates are less sensitive, than the normal model maximum likelihood estimates, to a few outliers.

Keywords: Empirical characteristic function; Power transformations; Influence function

1. INTRODUCTION

Many statistical techniques are based on assumptions about the form of the population distribution. The validity of the results obtained depends, sometimes critically, on the assumed conditions being satisfied, at least approximately. When these assumptions are seriously violated, a transformation of the data may permit the valid use of these techniques. In this paper, the goal is to achieve a distributional symmetry via the 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). Van Zwet defines the distribution function G to be more right-skewed (more left-skewed) than the distribution function F if $G^{-1}(F(x))$ is a non-decreasing convex (concave) function. According to van Zwet's definition, the right-skewness is increased if the transformation is convex, while the right-skewness is decreased if the transformation is concave. Since a non-decreasing convex (concave) transformation of a random variable effects a

[1]Department of Control and Instrumentation Engineering, Kangwon National University, Chunchon 200-701, Korea

contraction of the lower (upper) part of the support and an extension of the upper (lower) part, it decreases the skewness to the left (right). It is also shown that a convex-concave transformation preserves the ordering of the standardized odd central moments, one of the best known measures of skewness.

Box and Cox (1964) considered a family of power transformations satisfying this property. The Box-Cox transformation is, however, only valid for positive x. Since it is assumed that the transformed variables are approximately symmetrically distributed around an arbitrary center, the Box-Cox transformations may not be applied, at least not with adding a data dependent translation parameter. In this paper, the development is based on the extended family of power transformations

$$\psi(\lambda, x) = \begin{cases} \left\{(x+1)^{\lambda} - 1\right\}/\lambda & \text{for } x \geq 0, \lambda \neq 0, \\ \log(x+1) & \text{for } x \geq 0, \lambda = 0, \\ -\left\{(-x+1)^{2-\lambda} - 1\right\}/(2-\lambda) & \text{for } x < 0, \lambda \neq 2, \\ -\log(-x+1) & \text{for } x < 0, \lambda = 2, \end{cases} \tag{1.1}$$

which is introduced in Yeo and Johnson (2000). This family has properties similar to the Box-Cox transformation. In particular, $\psi(\lambda, x)$ is concave in x for $\lambda < 1$ and convex for $\lambda > 1$. Note that $\psi(\lambda, x)$ reduces to the identity function when $\lambda = 1$.

It is well known that the maximum likelihood estimator of transformation parameter is very sensitive to outliers, see Andrews (1971). Carroll (1980) has proposed a robust method for selecting a power transformation to achieve approximate normality in a linear model. Hinkley (1975) and Taylor (1985) have suggested methods for estimating λ in the Box-Cox transformation when the goal is to obtain approximate symmetry rather than normality. In the following sections, a robust method for estimating the transformation parameter λ is suggested by using the empirical characteristic function and the asymptotics and influence of the estimator is investigated.

2. ESTIMATION METHOD AND ASYMPTOTIC RESULTS

To motivate the choice of the deviation measure, we recall a well-known fact that a random variable X is symmetrically distributed around a location parameter μ if and only if the the characteristic function with factor $\exp(-i\mu t)$ is real. Assume that, for some λ, the distribution of the transformed variable $\psi(\lambda, X)$ is symmetric about a location parameter μ. Let $\phi_n(\lambda, t)$ be the empirical characteristic function of transformed variables $\psi(\lambda, X_1), \ldots, \psi(\lambda, X_n)$, that is,

$$\phi_n(\lambda, t) = n^{-1}\sum_{j=1}^{n}\exp(it\psi(\lambda, X_j)) = \phi_{cn}(\lambda, t) + i\phi_{sn}(\lambda, t),$$

where $\phi_{cn}(\lambda, t) = n^{-1}\sum_{j=1}^{n}\cos(t\psi(\lambda, X_j))$ and $\phi_{sn}(\lambda, t) = n^{-1}\sum_{j=1}^{n}\sin(t\psi(\lambda, X_j))$. Yeo and Johnson (2001) propose to transform X according to $\psi(\lambda, X)$ and then to select λ and μ to make the integrated square of the imaginary part of the empirical characteristic function of $\psi(\lambda, X_1), \ldots, \psi(\lambda, X_n)$ with factor $\exp(-it\mu)$ minimized,

$$\begin{aligned}\varphi_n(\boldsymbol{\theta}) &= \int [\mathrm{Im}\{\exp(-it\mu)\phi_n(\lambda, t)\}]^2 \, dG(t)\\ &= \int \left[n^{-1}\sum_{j=1}^{n}\sin\{t(\psi(\lambda, X_j) - \mu)\}\right]^2 dG(t),\end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)' = (\lambda, \mu)'$ and $G(\cdot)$ is a symmetric distribution function. Let $G(\cdot)$ have a characteristic function $\nu(\cdot)$. Then, we can write

$$\varphi_n(\boldsymbol{\theta}) = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{\nu[\psi(\lambda, X_i) - \psi(\lambda, X_j)] - \nu[\psi(\lambda, X_i) + \psi(\lambda, X_j) - 2\mu]\}, \quad (2.1)$$

since $\sin(x)\sin(y) = \{\cos(x-y) - \cos(x+y)\}/2$.

Let $\phi(\lambda, t)$ be the characteristic function of $\psi(\lambda, X)$. The distribution of $\psi(\lambda, X)$ is symmetric about μ if and only if $\mathrm{Im}\{\exp(-it\mu)\phi(\lambda, t)\}$ is zero. Hence, it can be a proper estimating method to select the value $\hat{\boldsymbol{\theta}}$ which minimizes $\varphi_n(\boldsymbol{\theta})$. Usually, in a given instance, it may not be possible to select a λ so that $\psi(\lambda, X)$ has a symmetric distribution. Nevertheless, we make that assumption. Recall the Box and Cox (1964) assumption of normality and see Hernandez and Johnson (1980) examples.

Let $\boldsymbol{\theta}_0 = (\lambda_0, \mu_0)'$ be the minimizer of $\varphi(\boldsymbol{\theta}) = \int [\mathrm{Im}\{\exp(-it\mu)\phi(\lambda, t)\}]^2 \, dG(t)$. Assume that there is a compact set defined as

$$\boldsymbol{\Theta} = \{(\lambda, \mu) \mid a \le \lambda \le b \text{ and } c \le \mu \le d\}$$

over which the following conditions are satisfied:

(i) $\boldsymbol{\theta}_0$ is unique on $\boldsymbol{\Theta}$,

(ii) $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$,

(iii) $E[I_{(X<0)}(-X)^{2(2-a)}\log(-X+1)^2] < \infty$ and $E[I_{(X\geq 0)}X^{2b}\log(X+1)^2] < \infty$,

(iv) $\nabla\varphi(\boldsymbol{\theta}_0) = \left(\left.\frac{\partial\varphi(\boldsymbol{\theta})}{\partial\theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) = \mathbf{0}$,

(v) $\nabla^2\varphi(\boldsymbol{\theta}_0) = \left(\left.\frac{\partial^2\varphi(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right)$ is non-singular.

Theorem 2.1. *Let $\nu(\cdot)$ be the characteristic function of the distribution function $G(\cdot)$ with $\int t^2 dG(t) < \infty$, where $G(t^-) + G(-t) = 1$ for all $t \in \boldsymbol{R}$ and $G(t) - G(-t) > 0$ for any $t > 0$. Denote $\tau(\boldsymbol{\theta}, x_1, x_2) = \nu[\psi(\lambda, x_1) - \psi(\lambda, x_2)] - \nu[\psi(\lambda, x_1) + \psi(\lambda, x_2) - 2\mu]$.*

(1) $\varphi_n(\boldsymbol{\theta}) \xrightarrow{a.s} \varphi(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \Theta$. Further $\varphi(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$.

(2) Assume condition (i). Then, $\hat{\boldsymbol{\theta}} = \arg\min\varphi_n(\boldsymbol{\theta})$ converges almost surely to $\boldsymbol{\theta}_0$.

(3) Assume conditions (i)-(iv). Then, $n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0)$ is asymptotically distributed with $N(\mathbf{0}, \Sigma(\boldsymbol{\theta}_0))$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = E_{X_2}[E_{X_1}[\nabla\tau(\boldsymbol{\theta}_0, X_1, X_2)\,(\nabla\tau(\boldsymbol{\theta}_0, X_1, X_2))']]$

(4) Assume conditions (i)-(v). Then, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically distributed with $N(0, \mathbf{V}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)')$, where $\mathbf{V}(\boldsymbol{\theta}_0) = \left(\nabla^2\varphi(\boldsymbol{\theta}_0)\right)^{-1}$.

Proofs.
(1) See Yeo and Johnson (2001).
(2) See Yeo and Johnson (2001).
(3) Decomposing $n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0)$, we see, from (2.1), that

$$n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0) = \frac{1}{2n^{3/2}}\sum_{j=1}^{n}\nabla\tau(\boldsymbol{\theta}_0, X_j, X_j) + \frac{n-1}{2n}n^{1/2}U_{1n}(\boldsymbol{\theta}_0), \tag{2.2}$$

where $U_{1n}(\boldsymbol{\theta}_0) = \binom{n}{2}^{-1}\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}\nabla\tau(\boldsymbol{\theta}_0, X_j, X_k)$. The strong law of large numbers ensures that the first term on the right hand side of (2.2) converges almost surely to $\mathbf{0}$. The multivariate central limit theorem for U-statistics (see Lee (1990), page 76) gives that $n^{1/2}U_{1n}(\boldsymbol{\theta}_0)$ is asymptotically distributed with $N(\mu(\boldsymbol{\theta}_0), 4\boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$, where $\mu(\boldsymbol{\theta}_0) = 2\nabla\varphi(\boldsymbol{\theta}_0) = \mathbf{0}$ and

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}_0) = E_{X_2}[E_{X_1}[(\nabla\tau(\boldsymbol{\theta}_0, X_1, X_2))(\nabla\tau(\boldsymbol{\theta}_0, X_1, X_2))']].$$

Slutsky's theorem allows to conclude that

$$n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0) \text{ is asymptotically distributed with } N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)). \tag{2.3}$$

(4) Expanding $n^{1/2}\nabla\varphi_n(\hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}_0$, we obtain that

$$n^{1/2}\nabla\varphi_n(\hat{\boldsymbol{\theta}}) = n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0) + \nabla^2\varphi_n(\tilde{\boldsymbol{\theta}})n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\tilde{\boldsymbol{\theta}} = \alpha_n\hat{\boldsymbol{\theta}} + (1-\alpha_n)\boldsymbol{\theta}_0$ for $\alpha_n \in [0,1]$ and $n \geq 1$. Since $n^{1/2}\nabla\varphi(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ at the minimum when $\hat{\boldsymbol{\theta}}$ lies in the interior of Θ, $n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0) - (-\nabla^2\varphi_n(\tilde{\boldsymbol{\theta}}))n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in probability to $\mathbf{0}$. From (2.2), $\nabla^2\varphi_n$ can be written as

$$\nabla^2\varphi_n(\boldsymbol{\theta}) = (2n^2)^{-1}\sum_{j=1}^{n}\nabla^2\tau(\boldsymbol{\theta}, X_j, X_j) + \frac{n-1}{2n}U_{2n}(\boldsymbol{\theta}), \tag{2.4}$$

where $U_{2n}(\boldsymbol{\theta}) = \binom{n}{2}^{-1}\sum_{j=1}^{n-1}\sum_{k=j+1}^{n}\nabla^2\tau(\boldsymbol{\theta}, X_j, X_k)$. Applying the uniform convergence by Rubin (1956) and the uniform strong law of large numbers for U-statistic by Yeo and Johnson (2001), we conclude that $\nabla^2\varphi_n(\boldsymbol{\theta})$ converges almost surely to $\nabla^2\varphi(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \Theta$. Further, the limit function $\nabla^2\varphi(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. Hence, using uniform convergence of $\nabla^2\varphi_n$ and the continuity of $\nabla^2\varphi$ with almost sure convergence of $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}_0$, it is easy to show that

$$\nabla^2\varphi_n(\tilde{\boldsymbol{\theta}}) \text{ converges almost surely to } \nabla^2\varphi(\boldsymbol{\theta}_0). \tag{2.5}$$

By Slutsky's theorem along with (2.3) and (2.5), we conclude that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically distributed with $N(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)')$, where $\mathbf{V}(\boldsymbol{\theta}_0) = (\nabla^2\varphi(\boldsymbol{\theta}_0))^{-1}$. □

3. EXAMPLES FOR THE WEIGHT DISTRIBUTION G

To gain some understanding regarding the choice of a weight function, we first specify absolutely continuous weight distributions by their density. We consider the weight density functions;

$$\begin{aligned} g(t) &= (2\pi\sigma^2)^{-1/2}\exp(-t^2/(2\sigma^2)) && -\infty < t < \infty, \\ g(t) &= (2\Gamma(\alpha)\sigma^\alpha)^{-1}|t|^{\alpha-1}\exp(-|t|/\sigma) && -\infty < t < \infty, \\ g(t) &= 1/(2\sigma) && -\sigma < t < \sigma. \end{aligned}$$

These weight distributions satisfy the conditions in Theorem 2.1. Note that each of the three weight distribution families is indexed by a scale parameter $\sigma > 0$, so we write

$$\varphi_n(\boldsymbol{\theta}) = \varphi_n(\boldsymbol{\theta}, \sigma) = \int_{-\infty}^{\infty}\phi_{sn}^2(\boldsymbol{\theta}, t)\, dG(t, \sigma),$$

where $\phi_{sn}(\boldsymbol{\theta}, t) = n^{-1}\sum_{j=1}^{n} \sin(t(\psi(\lambda, x_j) - \mu))$. Since the weight distributions above have finite high order moments, the contribution of some neighborhood of the origin, $-\delta \leq t \leq \delta$, is almost equal to the whole integral when σ is small. For $-\delta \leq t \leq \delta$ such that δ is sufficiently small, $\phi_{sn}^2(\boldsymbol{\theta}, t)$ is successfully approximated by simpler function, that is,

$$\phi_{sn}^2(\boldsymbol{\theta}, t) = t^2\Big(n^{-1}\sum_{j=1}^{n}\psi(\lambda, x_j) - \mu\Big)^2 + O(t^4).$$

Hence, if we take σ small sufficiently, then, for small values of δ,

$$\varphi_n(\boldsymbol{\theta}, \sigma) \sim \Big(n^{-1}\sum_{j=1}^{n}\psi(\lambda, x_i) - \mu\Big)^2 \int_{-\delta}^{\delta} t^2 \, dG(t, \sigma).$$

That is, for a wide choice of weight distributions, the particular choice negligibly affects the estimation of λ when small σ is used.

4. INFLUENCE OF OUTLIERS

The influence function serves to describe the effect of an outlier on estimation. Let T be a functional. Then, the proposed estimator can be written as a function $T(F_n)$ where F_n stands for the empirical distribution function. The influence function evaluated at a point x_0 is defined as

$$IF(x_0, F) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F + \varepsilon\delta(x_0)] - T(F)}{\varepsilon},$$

where $\delta(x_0)$ is the probability measure which puts mass one at the point x_0. In the case where $\mu = 0$, $\sigma^2 = 1$, the influence function of the estimation method proposed by Carroll (1980) is proportional to $\rho(\psi(\lambda_0, x_0))\psi^{(1)}(\lambda_0, x_0)$ where, for some $k > 0$, $\rho(a) = a$ if $a \leq k$ and $= \text{sign}(a)k$ otherwise and

$$\begin{aligned}
\psi^{(1)}(\lambda, x) &= \frac{\partial}{\partial\lambda}\psi(\lambda, x) \\
&= \begin{cases}
\{(x+1)^{\lambda}\log(x+1) - \psi(\lambda, x)\}/\lambda & \text{for } \lambda \neq 0,\ x \geq 0, \\
\log(x+1)^2/2 & \text{for } \lambda = 0,\ x \geq 0, \\
\{(-x+1)^{2-\lambda}\log(-x+1) + \psi(\lambda, x)\}/(2-\lambda) & \text{for } \lambda \neq 2,\ x < 0, \\
\log(-x+1)^2/2 & \text{for } \lambda = 2,\ x < 0.
\end{cases}
\end{aligned}$$

When no transformation is necessary, i.e. $\lambda_0 = 1$, for a large deviation x from 0, the influence function of the normal theory maximum likelihood estimator and

Carroll's robust estimator are proportional to $\text{sign}(x)x^2 \log(|x|+1)$ and $x\log(|x|+1)$, respectively.

The proposed estimate $\hat{\lambda}$ can be written as the solution to

$$\begin{aligned}\frac{\partial}{\partial\lambda}\varphi_n(\boldsymbol{\theta}) &= \int t\left\{\int \cos(t\psi(\lambda,x))\psi^{(1)}(\lambda,x)dF_n(x)\right\} \\ &\quad\times\left\{\int \sin(t\psi(\lambda,y))dF_n(y)\right\}dG(t) = 0.\end{aligned}$$

For $\lambda_0 = 1$, the influence function of our estimation is proportional to

$$x\log(|x|+1)\int t\ \sin(ty)\cos(tx)\ dG(t).$$

Since $\int |t\sin(ty)\cos(tx)|dG(t) \le \int |t|dG(t)$, applying a symmetric weight distribution $G(\cdot)$ having finite moments, e.g. standard normal, we see that the influence function is proportional to $x\log(|x|+1)$. It is shown that the proposed estimate of λ is also sensitive, but less sensitive than the normal maximum likelihood estimate, to an outlier. Also, it can be easily shown that the influence of an outlier on estimating μ is bounded.

5. SIMULATION

To compare the performance of the estimation and influence of outliers with the normal maximum likelihood estimation, a simulation was performed. A series of 10000 replications, of samples of size 30, were generated for $\lambda_0 = 0.0$, 0.5, and 1.5 where $\psi(\lambda_0, X)$ is normally distributed with its mean 0 and its standard deviation 3, $N(0,3^2)$. The normal weight distribution with $\sigma = 0.01$ was employed. Table 5.1 gives the simulated bias, standard deviation and mean squared error of the two estimates for λ_0. MSEC denotes the proposed M-estimate. Maximum likelihood methods, see Yeo and Johnson(2000), perform well for all cases of the normal distribution. When the single positive outlier $\psi(\lambda_0, x_{31}) = 30$ to each data set was added, $N(0,3^2)+$, we see that a large increase in bias is traded for a small decrease in standard deviation, resulting in increase in MSE. Further, the bias of MLEs is highly increased so that the MSE of MLEs tends to become larger than that of MSECs. This implies that the MSEC may be less sensitive, than the normal model MLE, to the outliers. An interesting result is that, as λ_0 gets away from 1, both estimations tend to provide a more accurate estimate regardless of the existence of outliers.

Table 5.1: The Monte Carlo Bias (BIAS), standard deviation (STD) and mean squared error (MSE) of estimates for MLE and MSEC. Based on 10000 replications, $\lambda_0 = 0.0$, 0.5 and 1.5, and sample size $n = 30$.

		$\lambda_0 = 0.0$		$\lambda_0 = 0.5$		$\lambda_0 = 1.5$	
		MLE	MSEC	MLE	MSEC	MLE	MSEC
$N(0,3^2)$	BIAS	0.020	-.005	0.010	-.008	-.012	0.002
	STD	0.086	0.132	0.123	0.169	0.124	0.169
	MSE	0.008	0.017	0.015	0.029	0.016	0.029
$N(0,3^2)+$	Bias	-.125	-.109	-.339	-.191	-.391	-.223
	STD	0.018	0.087	0.059	0.112	0.095	0.120
	MSE	0.016	0.019	0.119	0.049	0.162	0.064

Acknowledgement

The author gratefully acknowledge the constructive comments of anonymous referees.

References

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika* **58**, 249-54.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, **B 26**, 211-252.

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *Journal of the Royal Statistical Society*, **B 42**, 71-78.

Hernandez, F. and Johnson, R. A. (1980). The large-sample behavior of transformations to normality. *Journal of the American Statistical Association*, **75**, 855-861.

Hinkley, D. V. (1975). On power transformations to symmetry, *Biometrika*, **62**, 101-111.

Lee, A. J. (1990). *U-statistics : Theory and Practice*, Marcel Dekker, New York.

Rubin, H. (1956). Uniform convergence of random functions with applications to statistics. *Annals of Mathematical Statistics*, **27**, 200-203.

Taylor, J. M. G. (1985). Power Transformations to Symmetry. *Biometrika*, **72**, 145-152.

van Zwet, W. R. (1964). *Convex transformations of random variables*, Mathematisch Centrum, Amsterdam.

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

Yeo, I. K. and Johnson, R. A. (2001). A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. *Statistics and Probability Letters*, **51**, 63-69.