Communications for Statistical Applications and Methods
2014, Vol. 21, No. 3, 213–224

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.3.213
ISSN 2287-7843

An Empirical Characteristic Function Approach to Selecting a Transformation to Normality

In-Kwon Yeo[1,a], Richard A. Johnson[b], XinWei Deng[c]

[a]Department of Statistics, Sookmyung Women's University, Korea
[b]Department of Statistics, University of Wisconsin-Madison, USA
[c]Department of Statistics, Virginia Tech, USA

Abstract

In this paper, we study the problem of transforming to normality. We propose to estimate the transformation parameter by minimizing a weighted squared distance between the empirical characteristic function of transformed data and the characteristic function of the normal distribution. Our approach also allows for other symmetric target characteristic functions. Asymptotics are established for a random sample selected from an unknown distribution. The proofs show that the weight function t^{-2} needs to be modified to have thinner tails. We also propose the method to compute the influence function for M-equation taking the form of U-statistics. The influence function calculations and a small Monte Carlo simulation show that our estimates are less sensitive to a few outliers than the maximum likelihood estimates.

Keywords: Box-Cox transformation, influence function, Yeo-Johnson transformation.

1. Introduction

Transformation of data is a useful tool that permits the use of an assumption such as normality, when the observed data seriously violate this condition. It is well-known that, under the normality assumption, the maximum likelihood estimator of the Box-Cox transformation parameter is very sensitive to outliers, see Andrews (1971). There are two ways to circumvent this problem. The first approach is to construct diagnostics which attempt to identify influential observations and then to remove these observations before estimating the transformation parameter. Cook and Wang (1983), Hinkley and Wang (1988), Tsai and Wu (1990) and Kim *et al.* (1996) studied case deletion diagnostics for the Box-Cox transformation. However, for multiple outliers, these diagnostic procedures are quite complicated and require an extensive computation. The second approach is to perform a robust estimation procedure that is not strongly affected by outliers. Carroll (1980) proposed a robust method for selecting a power transformation to achieve approximate normality in a linear model. Hinkley (1975) and Taylor (1985) suggested methods to estimate the transformation parameter in the Box-Cox transformation when the goal is to obtain approximate symmetry rather than normality. Yeo and Johnson (2001) and Yeo (2001) introduced an M-estimator obtained by minimizing the integrated square of the imaginary part of the empirical characteristic function of Yeo-Johnson transformed data.

In this paper, we propose a robust method to estimate a transformation parameter as well as the mean and the variance of the target distribution. The estimators are obtained by minimizing a squared

This research was supported by the Sookmyung Women's University Research Grants 2013.
[1] Corresponding author: Department of Statistics, Sookmyung Women's University, Seoul 140-742, Korea.
E-mail: inkwon@sookmyung.ac.kr

Published online 31 May 2014 / journal homepage: http://csam.or.kr

distance between the empirical characteristic function of the transformed data and the target characteristic function which is often that of a normal distribution. Specifically, we minimize the integral of the squared modulus of the difference of the two characteristic functions multiplied by a weight function. It is assumed that, for some interval about zero, the weight function equals t^{-2} which is used to compute the distance covariance by Szekely *et al.* (2007).

Many authors such as Koutrouvelis (1980), Koutrouvelis and Kellermeier (1981), Fan (1997), Klar and Meintanis (2005), and Jimenez-Gamero *et al.* (2009) have proposed goodness-of-fit test statistics based on measuring differences between the empirical characteristic function and the characteristic function in the null hypothesis. Conversely, we employ this statistic as a measurement to estimate the transformation parameter as well as the mean and the variance. Our estimation procedure can be viewed as solving estimating equations based on a U-statistic, see Lee (1990). In order to calculate the influence function, we take the approach of calculating this function in terms of the asymptotically equivalent statistic that is a sum of independent and identically distributed terms. The resulting expressions show that our procedure is more robust than the maximum likelihood procedure.

2. Estimation

Let $\psi(x, \lambda)$ be a family of transformations indexed by the transformation parameter λ, for instances, Box and Cox (1964), John and Draper (1980), and Yeo and Johnson (2000). Generally, a main goal of transforming data is to enhance the normality of data. According to the definition of the relative skewness by van Zwet (1964), a convex transformation reduces left-skewness or vice versa. The Box-Cox transformation and the Yeo-Johnson transformation are either convex or concave and can be applied to skewed data to improve symmetry. By contrast, the modulus transformation by John and Draper (1980) is convex-concave and can be applied to symmetric data to reduce kurtosis.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution function $F(\cdot)$. It is assumed that there exists a transformation parameter λ for which $\psi(X, \lambda)$ is normally distributed with mean μ and variance σ^2. This assumption can be extended to any symmetric distribution with the location and the scale parameter, μ and σ.

Let $\phi(t)$ be the characteristic function of the standard normal distribution, that is $\phi(t) = e^{-t^2/2}$, and let $\phi_n(\boldsymbol{\theta}, t)$ be the empirical characteristic function of standardized transformed variables $Z_j(\boldsymbol{\theta}) = \{\psi(X_j, \lambda) - \mu\}/\sigma$, $j = 1, \ldots, n$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T = (\lambda, \mu, \sigma^2)^T$. Then,

$$\phi_n(\boldsymbol{\theta}, t) = n^{-1} \sum_{j=1}^{n} \exp\left(itZ_j(\boldsymbol{\theta})\right) = \phi_{cn}(\boldsymbol{\theta}, t) + i\phi_{sn}(\boldsymbol{\theta}, t),$$

where

$$\phi_{cn}(\boldsymbol{\theta}, t) = n^{-1} \sum_{j=1}^{n} \cos\left(tZ_j(\boldsymbol{\theta})\right) \quad \text{and} \quad \phi_{sn}(\boldsymbol{\theta}, t) = n^{-1} \sum_{j=1}^{n} \sin\left(tZ_j(\boldsymbol{\theta})\right).$$

Here, the parameter space $\boldsymbol{\Theta}$ is assumed to be a compact set of the form

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} \mid a_i \le \theta_i \le b_i, \text{ where } 0 < a_3, \ |a_i|, |b_i| < \infty, \text{ for } i = 1, 2, 3\}. \tag{2.1}$$

The goodness-of-fit test statistics based on measuring differences between the empirical characteristic function and the characteristic function in the null hypothesis have been extensively investigated

in the literature. For example, Jimenez-Gamero *et al.* (2009) consider the statistic

$$T_{n,G}\left(\hat{\theta}\right) = n \int \left|\phi_n(t) - \phi\left(t;\hat{\theta}\right)\right|^2 dG(t),$$

where $\phi_n(t)$ is the empirical characteristic function, $\phi(t;\theta)$ is the characteristic function of target distribution, $\hat{\theta}$ is a consistent estimator and G is a distribution function. In this paper, we propose the method to estimate $\boldsymbol{\theta}$ by minimizing an integrated weighted version of the distance between the empirical characteristic function of $Z(\boldsymbol{\theta})$ and $e^{-t^2/2}$,

$$\varphi_n(\boldsymbol{\theta}) = \|\phi_n(\boldsymbol{\theta}) - \phi\|_w^2 = \int \left\{\phi_n(\boldsymbol{\theta},t) - e^{-\frac{t^2}{2}}\right\}\left\{\overline{\phi_n(\boldsymbol{\theta},t)} - e^{-\frac{t^2}{2}}\right\} w(t)dt,$$

where the overline denotes the complex conjugate and $w(t)$ is a non-negative real-valued weight function. Therefore,

$$\varphi_n(\boldsymbol{\theta}) \propto \int \phi_n(\boldsymbol{\theta},t)\overline{\phi_n(\boldsymbol{\theta},t)}w(t)\,dt - \int \left\{\phi_n(\boldsymbol{\theta},t) + \overline{\phi_n(\boldsymbol{\theta},t)}\right\} e^{-\frac{t^2}{2}} w(t)\,dt.$$

Here

$$\int \phi_n(\boldsymbol{\theta},t)\overline{\phi_n(\boldsymbol{\theta},t)}w(t)\,dt = \frac{1}{n} + \frac{2}{n^2}\sum_{j<k}\int \cos\left(t\left\{Z_j(\boldsymbol{\theta}) - Z_k(\boldsymbol{\theta})\right\}\right) w(t)\,dt,$$

$$\int \left\{\phi_n(\boldsymbol{\theta},t) + \overline{\phi_n(\boldsymbol{\theta},t)}\right\} e^{-\frac{t^2}{2}} w(t)\,dt = \frac{2}{n}\sum_{j=1}^{n}\int \cos\left(tZ_j(\boldsymbol{\theta})\right) e^{-\frac{t^2}{2}} w(t)\,dt.$$

Let $\phi(\boldsymbol{\theta},t) = E[\exp(itZ(\boldsymbol{\theta}))]$ denote the characteristic function of the standardized transformed variable $Z(\boldsymbol{\theta})$ and

$$\varphi(\boldsymbol{\theta}) = \|\phi(\boldsymbol{\theta}) - \phi\|_w^2 = \int \left\{\phi(\boldsymbol{\theta},t) - e^{-\frac{t^2}{2}}\right\}\left\{\overline{\phi(\boldsymbol{\theta},t)} - e^{-\frac{t^2}{2}}\right\} w(t)dt.$$

The distribution of $Z(\boldsymbol{\theta})$ is equivalent to the standard normal distribution if and only if $\varphi(\boldsymbol{\theta})$ is zero. Hence, a reasonable approach to estimation is to select the value $\hat{\boldsymbol{\theta}} = (\hat{\lambda},\hat{\mu},\hat{\sigma})^T$ which minimizes $\varphi_n(\boldsymbol{\theta})$.

In order to compute $\varphi_n(\boldsymbol{\theta})$, we have to specify the weight function. If the normal density with the mean 0 and the variance δ^2 is employed as the weight function, then $\int \cos(tz)w(t)\,dt$ is the characteristic function of the normal distribution and so $\int \cos(tz)w(t)\,dt = \exp\{-\delta^2 z^2/2\}$ and $\int \cos(tz)e^{-t^2/2}w(t)\,dt$ $= (1+\delta^2)^{-1/2}\exp\{-\delta^2 z^2/(2(1+\delta^2))\}$. Therefore, estimates are obtained by minimizing

$$\varphi_n(\boldsymbol{\theta}) \propto \frac{1}{n}\sum_{j<k}\exp\left\{-\frac{\delta^2}{2}\left(Z_j(\boldsymbol{\theta}) - Z_k(\boldsymbol{\theta})\right)^2\right\} - \frac{1}{\sqrt{1+\delta^2}}\sum_{j=1}^{n}\exp\left\{-\frac{\delta^2 Z_j(\boldsymbol{\theta})^2}{2(1+\delta^2)}\right\}.$$

Here, according to Epps and Pulley (1983), δ must be a small value . The behavior in neighborhood of zero is important for characteristic functions. They mentioned that $w(t)$ should assign high weight in some interval around the origin.

As in Szekely *et al.* (2007), an alternative weight function is $w(t) = t^{-2}$ on some interval containing zero and we define integrals as the principle values. The integral on 0 to ∞ is the limit as $\epsilon \to 0$ of

the integral over $(\epsilon, \epsilon^{-1})$. However, when we later consider properties of estimators, including the influence function, the upper and lower tails are too heavy and we find it necessary to impose a moment condition on $w(t)$. In this paper, we consider the truncated weight function such as $w(t) = t^{-2}$ if $t \in (-\delta, \delta)$ for a finite δ and 0 otherwise. The estimation procedure with this weight function involves some difficult numerical integrations and the proof of the asymptotic results is somewhat cumbersome; however, a simulation study shows that this weight function gives a better result. The asymptotic results are derived with this weight function.

3. Asymptotic Theory

Before stating our results, we introduce some notations. For any function $f(\boldsymbol{\theta})$, for $j, k = 1, 2, 3$,

$$\nabla f(\boldsymbol{\theta}_*) = \left(\left. \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \right) \quad \text{and} \quad \nabla^2 f(\boldsymbol{\theta}_*) = \left(\left. \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \right)$$

are the gradient and the Hessian of f evaluated at $\boldsymbol{\theta}_*$, respectively, for $j, k = 1, 2, 3$. We also write

$$\nabla_j f(\boldsymbol{\theta}_*) = \left. \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \quad \text{and} \quad \nabla^2_{jk} f(\boldsymbol{\theta}_*) = \left. \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*},$$

and $\psi_k(x, \lambda) = \partial^k \psi(x, \lambda)/\partial \lambda^k$ where $\psi_0(x, \lambda) = \psi(x, \lambda)$.

Theorem 1. *Let the parameter space $\boldsymbol{\Theta}$ be given by (2.1) and let $w(t) = t^{-2}$ if $t \in (-\delta, \delta)$ and 0 otherwise. Assume that, for $k = 0, 1, 2$, $\psi_k(x, \lambda)$ is continuous in (x, λ) and that there exists a function $h_k(x)$ that satisfies $|\psi_k(x, \lambda)| \le h_k(x)$ and $E[h_k(X)^2] < \infty$ on $\boldsymbol{\Theta}$. Suppose $\varphi(\boldsymbol{\theta})$ has a unique global minimum at $\boldsymbol{\theta}_0 = (\lambda_0, \mu_0, \sigma_0)^T$ where $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$. Then,*

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

$$\lim_{n\to\infty} \left\{ \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \varphi_n(\boldsymbol{\theta}) \right\} = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \varphi(\boldsymbol{\theta}) \quad \textit{with probability one.}$$

(2) $\hat{\boldsymbol{\theta}}$ is a strong consistent estimator of $\boldsymbol{\theta}_0$

(3) $n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0)$ is asymptotically distributed with $N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is specified in the proof

(4) $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$ is asymptotically distributed with $N(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^T)$, where $\mathbf{V}(\boldsymbol{\theta}_0) = (\nabla^2\varphi(\boldsymbol{\theta}_0))^{-1}$.

Proof:

(1) Note that

$$\varphi_n(\boldsymbol{\theta}) = \int \left\{ \left(\phi_{cn}(\boldsymbol{\theta}, t) - e^{-\frac{t^2}{2}} \right)^2 + \phi_{sn}(\boldsymbol{\theta}, t)^2 \right\} w(t)\, dt.$$

To verify the finiteness of $\varphi_n(\boldsymbol{\theta})$, we first bound the integrand by adding and subtracting 1. Since $|(\cos(tz) - 1)/t| \le |z|$ and $|\sin(tz)/t| \le |z|$,

$$\begin{aligned} \varphi_n(\boldsymbol{\theta}) &\le \int \left\{ 2\left(\phi_{cn}(\boldsymbol{\theta}, t) - 1\right)^2 + 2\left(1 - e^{-\frac{t^2}{2}}\right)^2 + \phi_{sn}(\boldsymbol{\theta}, t)^2 \right\} w(t) dt \\ &\le \frac{6\delta}{n} \sum_{j=1}^{n} Z_j(\boldsymbol{\theta})^2 + c, \end{aligned}$$

where $c = 2\int(1 - e^{-t^2/2})^2 w(t)dt$ is a finite number. Thus it is clear that $\varphi_n(\boldsymbol{\theta})$ is finite.

Rather than work with the V-statistics in brackets, we will take a U-statistic approach. Let $C(t, z, \boldsymbol{\theta}) = \cos(tz(\boldsymbol{\theta})) - e^{-t^2/2}$ and $S(t, z, \boldsymbol{\theta}) = \sin(tz(\boldsymbol{\theta}))$. We define

$$\eta(z_1, z_2; \boldsymbol{\theta}) = \int \{C(t, z_1, \boldsymbol{\theta})C(t, z_2, \boldsymbol{\theta}) + S(t, z_1, \boldsymbol{\theta})S(t, z_2, \boldsymbol{\theta})\}\ w(t)\, dt \tag{3.1}$$

and then have

$$\varphi_n(\boldsymbol{\theta}) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\eta\left(Z_j, Z_k; \boldsymbol{\theta}\right) = \frac{n-1}{n}U_n(\boldsymbol{\theta}) + \frac{1}{n^2}\sum_{j=1}^{n}\eta\left(Z_j, Z_j; \boldsymbol{\theta}\right), \tag{3.2}$$

where

$$U_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1}\sum_{j<k}\eta\left(Z_j, Z_k; \boldsymbol{\theta}\right).$$

Since $\eta(z_1, z_2; \boldsymbol{\theta})$ is bounded and continuous in $(z_1, z_2; \boldsymbol{\theta}) \in \Omega_M = S_M \times S_M \times \boldsymbol{\Theta}$ where $S_M = [-M, M]$, the uniform strong law of large numbers of U-statistics in Yeo and Johnson (2001) ensures that

$$U_n(\boldsymbol{\theta}) \xrightarrow{a.s.} E\left[\int \{C(t, Z_1, \boldsymbol{\theta})C(t, Z_2, \boldsymbol{\theta}) + S(t, Z_1, \boldsymbol{\theta})S(t, Z_2, \boldsymbol{\theta})\}\, w(t)dt\right] \equiv \eta(\boldsymbol{\theta})$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\eta(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Note that Z_1 and Z_2 are independent and identically distributed,

$$\eta(\boldsymbol{\theta}) = \int \left\{E\left[C(t, Z_1, \boldsymbol{\theta})\right]^2 + E\left[S(t, Z_1, \boldsymbol{\theta})\right]^2\right\}\, w(t)\, dt = \varphi(\boldsymbol{\theta}). \tag{3.3}$$

Furthermore, by the uniform strong law of large numbers in Rubin (1956),

$$\frac{1}{n}\sum_{j=1}^{n}\eta\left(Z_j, Z_j; \boldsymbol{\theta}\right) \xrightarrow{a.s.} E\left[\eta\left(Z_j, Z_j; \boldsymbol{\theta}\right)\right] \tag{3.4}$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and this limit function in (3.4) is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and finally we conclude that

$$\varphi_n(\boldsymbol{\theta}) \xrightarrow{a.s.} \varphi(\boldsymbol{\theta})$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the limit is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

(2) Since $\varphi_n(\boldsymbol{\theta}) \xrightarrow{a.s.} \varphi(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta}$ and $\varphi(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and, by assumption, $\boldsymbol{\theta}_0$ is unique minimizer of $\varphi(\boldsymbol{\theta})$, Lemma 2 in Yeo and Johnson (2001) allows us to conclude that $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$.

(3) Since $|\psi(x, \lambda)|$ and $|\psi_1(x, \lambda)|$ are bounded by integrable functions and $\boldsymbol{\Theta}$ is compact, each entry of $\nabla z_1(\boldsymbol{\theta})$ is bounded and each entry is dominated by an integrable function. We verify that $\nabla\eta(z_1, z_2; \boldsymbol{\theta})$ can be obtained by differentiating under the integral sign in (3.1). The result is

$$\begin{aligned}\nabla\eta(z_1, z_2; \boldsymbol{\theta}) = &\int \{T_1(t, z_1, z_2, \boldsymbol{\theta}) + T_2(t, z_1, z_2, \boldsymbol{\theta})\}\, t\, w(t)\, dt \\ &+ \int \{T_1(t, z_2, z_1, \boldsymbol{\theta}) + T_2(t, z_2, z_1, \boldsymbol{\theta})\}\, t\, w(t)\, dt,\end{aligned} \tag{3.5}$$

where

$$T_1(t, z_1, z_2, \boldsymbol{\theta}) = \sin(tz_1(\boldsymbol{\theta}))\left\{e^{-\frac{t^2}{2}} - \cos(tz_2(\boldsymbol{\theta}))\right\}\nabla z_1(\boldsymbol{\theta}),$$
$$T_2(t, z_1, z_2, \boldsymbol{\theta}) = \cos(tz_1(\boldsymbol{\theta}))\sin(tz_2(\boldsymbol{\theta}))\nabla z_1(\boldsymbol{\theta}).$$

Using $|\sin(tz)/t| \le |z|$ again, it is easy to show that entries of $\nabla\varphi_n(\boldsymbol{\theta})$ are finite. From (3.2), we see that

$$\nabla\varphi_n(\boldsymbol{\theta}) = \frac{n-1}{n}\nabla U_n(\boldsymbol{\theta}) + \frac{1}{n^2}\sum_{j=1}^{n}\nabla\eta\left(Z_j, Z_j; \boldsymbol{\theta}\right), \tag{3.6}$$

where

$$\nabla U_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1}\sum_{j<k}\nabla\eta\left(Z_j, Z_k; \boldsymbol{\theta}\right).$$

Again, by the uniform strong law of large numbers, the second term in (3.6) can be neglected. Note that $\nabla\eta(z_1, z_2; \boldsymbol{\theta})$ is a symmetric kernel and so $\nabla U_n(\boldsymbol{\theta})$ is also a U-statistics. Thus, the multivariate central limit theorem for random samples and U-statistics ensure the asymptotic normality of $\nabla\varphi_n(\boldsymbol{\theta}_0)$ with the mean vector $\nabla\varphi(\boldsymbol{\theta}_0) = \mathbf{0}$ and the covariance matrix $\mathbf{W}_n(\boldsymbol{\theta}_0)$, where the (j, k)-th element of $\mathbf{W}_n(\boldsymbol{\theta}_0)$ is

$$W_n^{(j,k)}(\boldsymbol{\theta}_0) = \frac{(n-1)^2}{n^2}\binom{n}{2}^{-1}\left\{2(n-2)E\left[\nabla_j\eta\left(Z_1, Z_2; \boldsymbol{\theta}_0\right)\nabla_k\eta\left(Z_1, Z_3; \boldsymbol{\theta}_0\right)\right]\right.$$
$$\left. + E\left[\nabla_j\eta\left(Z_1, Z_2; \boldsymbol{\theta}_0\right)\nabla_k\eta\left(Z_1, Z_2; \boldsymbol{\theta}_0\right)\right]\right\}.$$

Therefore, $n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0)$ is asymptotically normally distributed as $N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0))$, where the (j, k)-th element of $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ is

$$\Sigma^{(j,k)}(\boldsymbol{\theta}_0) = 4E\left[\nabla_j\eta(Z_1, Z_2; \boldsymbol{\theta}_0)\nabla_k\eta(Z_1, Z_3; \boldsymbol{\theta}_0)\right].$$

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

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

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

From (3.5) and (3.6), we have

$$\nabla^2\eta(z_1, z_2; \boldsymbol{\theta}) = \int\left\{S_1(t, z_1, z_2, \boldsymbol{\theta}) + tC_1(t, z_1, z_2, \boldsymbol{\theta})\right\}e^{-\frac{t^2}{2}}\, t\, w(t)\, dt$$
$$+ \int\left\{S_2(t, z_1, z_2, \boldsymbol{\theta}) - tC_2(t, z_1, z_2, \boldsymbol{\theta})\right\}t\, w(t)\, dt, \tag{3.7}$$

where

$$
\begin{aligned}
S_1(t, z_1, z_2, \boldsymbol{\theta}) &= \sin\left(tz_1(\boldsymbol{\theta})\right) \nabla^2 z_1(\boldsymbol{\theta}) + \sin\left(tz_2(\boldsymbol{\theta})\right) \nabla^2 z_2(\boldsymbol{\theta}), \\
C_1(t, z_1, z_2, \boldsymbol{\theta}) &= \cos\left(tz_1(\boldsymbol{\theta})\right) \nabla z_1(\boldsymbol{\theta}) \nabla z_1(\boldsymbol{\theta})^T + \cos(tz_2(\boldsymbol{\theta})) \nabla z_2(\boldsymbol{\theta}) \nabla z_2(\boldsymbol{\theta})^T, \\
S_2(t, z_1, z_2, \boldsymbol{\theta}) &= \sin\left(t(z_1(\boldsymbol{\theta}) - z_2(\boldsymbol{\theta}))\right) \nabla^2 (z_2(\boldsymbol{\theta}) - z_1(\boldsymbol{\theta})), \\
C_2(t, z_1, z_2, \boldsymbol{\theta}) &= \cos\left(t(z_1(\boldsymbol{\theta}) - z_2(\boldsymbol{\theta}))\right) \nabla (z_2(\boldsymbol{\theta}) - z_1(\boldsymbol{\theta})) \nabla \left(z_2(\boldsymbol{\theta}) - z_1(\boldsymbol{\theta})\right)^T
\end{aligned}
$$

and $\nabla^2 \varphi_n$ can be written as

$$
\nabla^2 \varphi_n(\boldsymbol{\theta}) = \frac{n-1}{n} \nabla^2 U_n(\boldsymbol{\theta}) + \frac{1}{n^2} \sum_{j=1}^{n} \nabla^2 \eta\left(Z_j, Z_j; \boldsymbol{\theta}\right), \tag{3.8}
$$

where

$$
\nabla^2 U_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{j<k} \nabla^2 \eta\left(Z_j, Z_k; \boldsymbol{\theta}\right).
$$

Note that, since $|\psi(x, \lambda)|$, $|\psi_1(x, \lambda)|$, and $|\psi_2(x, \lambda)|$ are bounded by integrable functions and $\boldsymbol{\Theta}$ is compact, each entry of $\nabla^2 z_1(\boldsymbol{\theta})$ and $\nabla z_1(\boldsymbol{\theta})(\nabla z_2(\boldsymbol{\theta}))^T$ is bounded. Hence, by the uniform strong law of large numbers, the second term in (3.8) can be neglected. Applying the uniform strong law of large numbers for U-statistic to $\nabla^2 U_n(\boldsymbol{\theta})$ we conclude that $\nabla^2 \varphi_n(\boldsymbol{\theta})$ converges almost surely to $\nabla^2 \varphi(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\nabla^2 \varphi(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. Hence, using the uniform convergence of $\nabla^2 \varphi_n$ and the continuity of $\nabla^2 \varphi$ with almost sure convergence of $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}_0$; therefore, it is easy to show that

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

Slutsky's theorem along with asymptotic normality of $n^{1/2} \varphi_n(\boldsymbol{\theta}_0)$ and (3.9) ensure $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically distributed with $N(\mathbf{0}, \mathbf{V}(\boldsymbol{\theta}_0) \boldsymbol{\Sigma}(\boldsymbol{\theta}_0) \mathbf{V}(\boldsymbol{\theta}_0)^T)$, where $\mathbf{V}(\boldsymbol{\theta}_0) = (\nabla^2 \varphi(\boldsymbol{\theta}_0))^{-1}$. □

Remark 1. Note that, for $a_1 \leq \lambda \leq b_1$, the Box-Cox transformation, the Yeo-Johnson transformation, and the modulus transformation satisfy the following inequalities:

$$
\begin{aligned}
|\psi(x, \lambda)| &\leq |\psi(x, a_1)| + |\psi(x, b_1)| = h(x), \\
|\psi_1(x, \lambda)| &\leq |\psi_1(x, a_1)| + |\psi_1(x, b_1)| = h_1(x), \\
|\psi_2(x, \lambda)| &\leq |\psi_2(x, a_1)| + |\psi_2(x, b_1)| = h_2(x).
\end{aligned}
$$

It is also shown that $\psi(x, \lambda)$, $\psi_1(x, \lambda)$, and $\psi_2(x, \lambda)$ are continuous in (x, λ).

4. Influence Function

The influence function is a basic tool to measure the robustness of estimators. Whereas the influence function of M-estimators with independent random variables has been extensively studied in literature, there is no literature that addresses the influence function for U-statistics. Hence we adopt the

following approach. Instead of working with the statistic $U_n(\boldsymbol{\theta}_0)$ in (3.2), we use the asymptotically equivalent statistic $n^{-1}\sum_{j=1}^{n}\eta(Z_j;\boldsymbol{\theta}_0)$, where

$$\begin{aligned}\eta(z_1;\boldsymbol{\theta}_0) &= E[\eta(z_1,Z_2;\boldsymbol{\theta}_0)]\\ &= \int C(t,z_1,\boldsymbol{\theta}_0)E\left[C(t,Z_2,\boldsymbol{\theta}_0)\right]w(t)\,dt + \int S(t,z_1,\boldsymbol{\theta}_0)E\left[S(t,Z_2,\boldsymbol{\theta}_0)\right]w(t)\,dt. \end{aligned} \tag{4.1}$$

Since $\eta(Z_j;\boldsymbol{\theta}_0)$'s are independent and identically distributed, we can use the usual M-equation theory of robustness based on estimating equations $\mathbf{0} = n^{-1}\sum_{j=1}^{n}\nabla\eta(Z_j;\boldsymbol{\theta}_0)$, where the derivative becomes

$$\begin{aligned}\nabla\eta(z_1;\boldsymbol{\theta}_0) &= E\left[\nabla\eta(z_1,Z_2;\boldsymbol{\theta}_0)\right]\\ &= \int\left\{E\left[T_1(z_1,Z_2,t,\boldsymbol{\theta}_0)\right] + E\left[T_2(z_1,Z_2,t,\boldsymbol{\theta}_0)\right]\right\}t\,w(t)\,dt\\ &\quad + \int\left\{E\left[T_1(Z_2,z_1,t,\boldsymbol{\theta}_0)\right] + E\left[T_2(Z_2,z_1,t,\boldsymbol{\theta}_0)\right]\right\}t\,w(t)\,dt. \end{aligned} \tag{4.2}$$

Assume that the conditions of Theorem 1 hold. Then, the influence function is given by

$$\mathrm{IF}(x_0,F) = \mathbf{M}^{-1}\nabla\eta(z;\boldsymbol{\theta}_0), \tag{4.3}$$

where

$$\begin{aligned}\mathbf{M} &= -E\left[\nabla^2\eta(Z_1;\boldsymbol{\theta}_0)\right]\\ &= 2\int E\left[\cos(tZ_1(\boldsymbol{\theta}))\right]E\left[\cos(tZ_1(\boldsymbol{\theta}))\nabla Z_1(\boldsymbol{\theta}_0)\nabla Z_1(\boldsymbol{\theta}_0)^T\right]t^2w(t)\,dt\\ &\quad 2\int E\left[\sin(tZ_1(\boldsymbol{\theta}))\right]E\left[\sin(tZ_1(\boldsymbol{\theta}))\nabla Z_1(\boldsymbol{\theta}_0)\nabla Z_1(\boldsymbol{\theta}_0)^T\right]t^2w(t)\,dt\\ &\quad - 2\int E\left[\cos(tZ_1(\boldsymbol{\theta}))\nabla Z_1(\boldsymbol{\theta}_0)\right]E\left[\cos(tZ_1(\boldsymbol{\theta}))\nabla Z_1(\boldsymbol{\theta}_0)^T\right]t^2w(t)\,dt\\ &\quad - 2\int E\left[\sin(tZ_1(\boldsymbol{\theta}))\nabla Z_1(\boldsymbol{\theta}_0)\right]E\left[\sin(tZ_1(\boldsymbol{\theta}))\nabla Z_1(\boldsymbol{\theta}_0)^T\right]t^2w(t)\,dt\\ &\quad + 2\int E\left[\cos(tZ_1(\boldsymbol{\theta}))\right]E\left[\sin(tZ_1(\boldsymbol{\theta}))\nabla^2 Z_1(\boldsymbol{\theta}_0)\right]tw(t)\,dt\\ &\quad - \int\left\{E\left[S_1(t,Z_1,Z_1,\boldsymbol{\theta}_0)\right] + tE\left[C_1(t,Z_1,Z_1,\boldsymbol{\theta}_0)\right]\right\}e^{-\frac{t^2}{2}}t\,w(t)\,dt, \end{aligned}$$

because Z_1 and Z_2 are independent and identically distributed. The influence function (4.3) is the standard result for M-estimator in an independent and identically distributed setting. Under our assumptions, all integrals and expectations in (4.1), (4.2), and (4.3) are finite.

For the Box-Cox transformation, the Yeo-Johnson transformation and the modulus transformation, $\lambda_0 = 1$ implies that the transformation is not necessary to improve normality. For $\lambda_0 = 1$, the influence function of maximum likelihood estimator under normality is proportional to $x^2\log(x)$, for $x > 0$, when the Box-Cox transformation is applied. Under our characteristic function approach, the influence function of the proposed estimator is proportional to $x\log(x)$. This calculation shows that the proposed estimator of λ is still sensitive, but less sensitive than the normal maximum likelihood estimate to an outlier. It can also be easily shown that the influence of an outlier to estimate μ is bounded. In the case where $\mu = 0$ and $\sigma^2 = 1$, Carroll (1980) proposed an estimator based on robustness arguments that also has influence function proportional to $x\log(x)$ when a just estimation of λ is considered.

5. Comparison with MLE

In this section, we present a small simulation that supports the influence function calculation by showing that our empirical characteristic function approach provides an estimator of λ that is more robust than the maximum likelihood estimator. Moreover, our proposed estimator appears to be slightly better than Carroll's robust estimator under the mixed alternatives considered below.

Our proposed method to estimate λ is compared with the maximum likelihood method and Carroll's method. When the Box-Cox transformation is employed, the maximum likelihood method estimates λ by maximizing the profile log-likelihood function

$$l(\lambda; \boldsymbol{x}) = -\frac{n}{2} \log\left(\hat{\sigma}(\lambda)^2\right) + (\lambda - 1) \sum_{i=1}^{n} \log(x_i),$$

where $\hat{\sigma}(\lambda)^2 = n^{-1} \sum_{j=1}^{n} (\psi(x_j, \lambda) - \hat{\mu}(\lambda))^2$ and $\hat{\mu}(\lambda) = n^{-1} \sum_{j=1}^{n} \psi(x_j, \lambda)$. Carroll follows Huber (1964) with the choice of loss function

$$\rho(x) = \begin{cases} \dfrac{x^2}{2}, & |x| \le k, \\ k\left(|x| - \dfrac{k}{2}\right), & |x| > k. \end{cases}$$

He expresses the likelihood in terms of the density with 'normal center-exponential tails' as follows;

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \sigma^{-n} \prod_{i=1}^{n} \exp\left\{-\rho\left(\frac{\psi(x_i, \lambda) - \mu}{\sigma}\right) + (\lambda - 1) \log(x_i)\right\}.$$

Thus the log-likelihood is

$$l(\boldsymbol{\theta}; \boldsymbol{x}) = -n \log(\sigma) + \sum_{i=1}^{n} \left\{-\rho\left(\frac{\psi(x_i, \lambda) - \mu}{\sigma}\right) + (\lambda - 1) \log(x_i)\right\}.$$

The estimation of λ consists of the following steps: For a fixed λ, take a starting value of σ and then estimate μ by solving

$$\sum_{i=1}^{n} \tau\left(\frac{\psi(x_i, \lambda) - \mu}{\sigma}\right) = 0,$$

where τ is the derivative of ρ. The iterative procedure estimates σ by solving

$$\frac{1}{n-1} \sum_{i=1}^{n} \tau^2\left(\frac{\psi(x_i, \lambda) - \mu}{\sigma}\right) = E_{\Phi}\left[\tau^2(Z)\right],$$

where the last expectation is taken with respect to a standard normal variable Z. To find the value of λ, maximize the likelihood function in

$$\hat{\lambda} = \arg\max_{\lambda} L(\mu(\lambda), \sigma(\lambda), \lambda).$$

As in Carroll (1980), we take $k = 2$ in the definition of $\rho(x)$.

Table 1: Simulation results for the estimation of λ with sample size $n = 100$. The weight function for our method is a truncated version of $1/t^2$. The mean, MSE and MAE are calculated from 1000 runs.

Distribution	Summary	Estimation Method			λ^*
		Proposed	MLE	Carroll	
(1)	Bias	0.0017	0.0009	0.0017	0
	SD	0.0014	0.0012	0.0012	
	MSE	0.0018	0.0015	0.0016	
	MAE	0.0343	0.0303	0.0305	
(2)	Bias	0.0006	0.0005	0.0007	0
	SD	0.0015	0.0016	0.0016	
	MSE	0.0021	0.0026	0.0026	
	MAE	0.0366	0.0402	0.0401	
(3)	Bias	0.0173	−0.0098	−0.0459	0.5
	SD	0.0079	0.0062	0.0064	
	MSE	0.0620	0.0385	0.0427	
	MAE	0.1982	0.1558	0.1612	
(4)	Bias	−0.1265	−0.2159	−0.2391	0.5
	SD	0.0082	0.0074	0.0074	
	MSE	0.0838	0.1008	0.1122	
	MAE	0.2301	0.2579	0.2753	
(5)	Bias	−0.0398	−0.0578	−0.1164	1
	SD	0.0131	0.0115	0.0052	
	MSE	0.1721	0.1348	0.1655	
	MAE	0.3444	0.2961	0.3231	
(6)	Bias	−0.2649	−0.4374	−0.4838	1
	SD	0.0152	0.0144	0.0147	
	MSE	0.3001	0.3999	0.4495	
	MAE	0.4410	0.5107	0.5468	

To evaluate the performance of the three methods of estimating λ, we generate pseudo random numbers from six different distributions. We write $N(\mu, 1)$ for a normal random variable having mean μ and variance 1. Also, for instance, the log-normal random variable whose logarithm is normal and has mean 2 and variance 2^2 is written as $\exp(2N(0, 1))$.

(1) log-normal distribution: $\exp(2N(0, 1))$.

(2) mixture of log-normal distributions: $0.98\exp(2N(0, 1)) + 0.02\exp(5N(0, 1))$.

(3) normal squared distribution: $N(5, 1)^2$.

(4) mixture of normal squared distributions: $0.98N(5, 1)^2 + 0.02N(8, 1)^2$.

(5) normal distribution: $N(5, 1)$.

(6) mixture of normal distributions: $0.98N(5, 1) + 0.02N(8, 1)$.

Note that for situations (2), (4) and (6), the generated data will contain about 2% outliers. For practical purposes, we truncated $w(t) = t^{-2}$ at $|t| \leq 50$. Initially, we also tried some weight functions that place less mass near 0.

For each situation, we generate samples of size $n = 100$ and obtain the estimate $\hat{\lambda}$ from each estimation method. This procedure is repeated $N = 1000$ times. For each estimation method, we summarize performance by calculating the mean and the standard deviation of $\hat{\lambda}$, the mean squared

error(MSE) of $\hat{\lambda}$, and the mean absolute error (MAE) of $\hat{\lambda}$. Specifically, we calculate

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\lambda}_i - \lambda^*\right)^2, \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\lambda}_i - \lambda^*\right|.$$

where $\hat{\lambda}_i$ is the estimate of λ from the i^{th} replication and λ^* denotes the value of true λ.

Table 1 reports the bias and the standard error of $\hat{\lambda}$, the MSE, and the MAE for the three methods. Carroll (1980) also gives a simulation where his and the Box-cox estimator have nearly the same behavior under a log-normal distribution. Based on these summary statistics, the proposed method shows worse performance when underlying distribution is not a mixture. However, in all three cases where the underlying distribution is a mixture, our proposed method based on the empirical characteristic function has a smaller bias, MSE and MAE than the Box-Cox estimation and Carroll's estimation. This implies that the proposed method is less sensitive to outliers.

References

Andrews, D. F. (1971). A note on the selection of data transformations, *Biometrika*, **58**, 249–254.

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.

Cook, R. A. and Wang, P. C. (1983). Transformations and influential cases in regression, *Technometrics*, **25**, 337–345.

Epps, T. W. and Pulley, L. B. (1983). A test for normality based on the empirical characteristic function, *Biometrika*, **70**, 723–726.

Fan, Y. (1997). Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function, *Journal of Multivariate Analysis*, **62**, 36–63.

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

Hinkley, D. V. and Wang, S. (1988). More about transformations and influential cases in regression. *Technometrics*, **30**, 435–440.

Huber, P. J. (1964). Robust estimation of a location parameter, *Annals of Statistics*, **53**, 73–101.

Jimenez-Gamero, M. D., Alba-Fernandez, V., Munoz-Garcia, J. and Chalco-Cano, Y. (2009). Goodness-of-fit tests based on empirical characteristic functions, *Computational Statistics and Data Analysis*, **53**, 3957–3971.

John, J. A. and Draper, N. R. (1980). An alternative family of transformations, *Applied Statistics*, **29**, 190–197.

Kim, C., Storer, B. E. and Jeong, M. (1996). A note on Box-Cox transformation diagnostics, *Technometrics*, **38**, 178–180.

Klar, B. and Meintanis, S. G. (2005) Tests for normal mixtures based on the empirical characteristic function, *Computational Statistics and Data Analysis*, **49**, 227–242.

Koutrouvelis, I. A. (1980). A goodness-of-fit test of simple hypothesis based on the empirical characteristic function, *Biometrika*, **67**, 238–240.

Koutrouvelis, I. A. and Kellermeier, J. (1981). A goodness-of-fit based on the empirical characteristic function when parameters must be estimated, *Journal of the Royal Statistical Society*, **B 43**, 173–176.

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.

Szekely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances, *Annals of Statistics*, **35**, 2769–2794.

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

Tsai, C. L. and Wu, X. (1990). Diagnostics in transformation and weighted regression, *Technometrics*, **32**, 315–322.

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

Yeo, I. K. (2001). Selecting a transformation to reduce skewness, *Journal of the Korean Statistical Society*, **30**, 563–571.

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.

Received November 29, 2013; Revised March 7, 2014; Accepted April 29, 2014