A NEW FAMILY OF POWER TRANSFORMATIONS TO IMPROVE NORMALITY OR SYMMETRY
by In-Kwon Yea and Richard A. Johnson
A new family of power transformations

to improve normality or symmetry

BY IN-KWON YEO

Research Institute of Applied Statistics, Sungkyunkwan University, Seoul 110-745, Korea

AND RICHARD A. JOHNSON

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

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. In particular, there is a convexity (or concavity) property in the parameter. The large sample properties of the transformation are investigated in the context of a single random sample.

Our new transformation is also applied with the less restrictive objective of improving the approximation to symmetry. We propose an M-estimator of the transformation parameter that is obtained by minimizing the integrated square of the imaginary part of the empirical characteristic function. As part of our derivation of the asymptotic properties, we develop a uniform strong law of large numbers for U-statistics which contains some parameter. According to influence function calculations, our estimates are less sensitive, than the normal model maximum likelihood estimates, to a few outliers.

Some key words: Empirical characteristic function, Kullback-Leibler information, maximum likelihood inference, power transformation, uniform strong law of large numbers

1. INTRODUCTION

Many parametric techniques in statistics are based on the assumption that the underlying distribution of the data is normal. However, there are several situations where the normal assumptions are seriously violated. To address cases where the original distribution is highly skewed, we develop a procedure for transforming to near normality or to near symmetry. Inference then concerns the parameters in the transformed model.

A major step towards an objective way of determining a transformation was made by Box and Cox (1964). They considered selecting transformations for achieving approximate normality. The Box-Cox transformation \( \psi^{BC}(\lambda, x) \) is given by

\[
\psi^{BC}(\lambda, x) = \begin{cases} 
(x^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, \\
\log(x) & \text{if } \lambda = 0,
\end{cases}
\] (1.1)

for positive \( x \). Box and Cox (1964) also discussed estimating \( \lambda \) by the maximum likelihood approach and by a Bayesian method. The Box-Cox transformation is, however, only valid
for positive $x$. Although a shift parameter can be introduced to handle the situations where the response is negative but bounded below, the standard asymptotic results of maximum likelihood theory may not apply since the range of distribution is determined by the unknown shift parameter (see Atkinson (1985)). To circumvent these problems, some statisticians consider the signed power transformation (see for instance Bickel and Doksum (1981))

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

(1.2)

which covers the whole real line. Since $\psi^{SP}$ 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 distribution,

$$
f(x) = 0.3\phi(x) + 0.7\gamma(x),
$$

(1.3)

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

$$
\gamma(x) = \frac{(x + 2)^3 \exp(- (x + 2))}{\Gamma(4)}, \quad x > -2.
$$

We are interested in transforming a random variable $X$ so that the transformed distribution is approximately normal. Following Hernandez and Johnson (1980), we select $\psi^{SP}$ to minimize the Kullback-Leibler information number,

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

(1.4)
where \( g_\lambda(\cdot) \) is the p.d.f. of the transformed variable and \( \phi_{\mu,\sigma^2}(\cdot) \) is the p.d.f. of a normal distribution with mean \( \mu \) and variance \( \sigma^2 \). The minimum is over a suitable range of \( \mu, \sigma^2 \) and \( \lambda \).

Even though \( \psi^{SP} \) is increasing in \( x \), its Jacobian changes from a decreasing function of \( x \) to an increasing function as \( x \) changes sign. Since the transformed density function has the Jacobian as one component, it has a cusp when \( x = 0 \). The cusp occurs at \( \psi^{SP}(\lambda, 0) = -1/\lambda \), so the transformed density is bimodal and looks far from normal distribution (see Figure 1).

In the context of transforming a symmetric distribution to near normality, John and Draper (1980) introduced the modulus transformation,

\[
\psi^{JD}(\lambda, x) = \begin{cases} 
\text{sign}(x) \left\{ (|x| + 1)^\lambda - 1 \right\} / \lambda & \text{if } \lambda \neq 0, \\
\text{sign}(x) \log(|x| + 1) & \text{if } \lambda = 0.
\end{cases}
\]  

(1.5)

This transformation also leads to a bimodal distribution in our example. A new family of transformation is needed.

In this article, we propose a new transformation family. We first apply a new transformation for improving the approximation to normality and investigate the large-sample properties of estimators. Next, we consider a more general objective. We try to determine a transformation so that the transformed variable is nearly symmetrically distributed, but not necessarily normal.

2. 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). To motivate the definition, let \( X \) be a random variable having a continuous distribution function \( F \) with inverse \( F^{-1} \) and 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 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 (concave) if and only if \( G \) is the distribution of a non-decreasing convex (concave) transformation \( \psi(X) \) of \( X \). He then defines relative skewness;

the distribution function \( G \) is more right-skewed (more left-skewed) than the distribution function \( F \) if \( G^{-1}(F(\cdot)) \) is a non-decreasing convex (concave) function.

Since a non-decreasing convex (concave) transformation of a random variable effects a 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). The Box-Cox transformation, for example, is concave in \( x \) for \( \lambda < 1 \) and convex in \( x \) for \( \lambda > 1 \). However, \( \psi^{SP} \) changes from convex to concave as \( x \) changes sign so it is not recommendable when two sided (positive and negative) data are skewed. John and Draper (1980) and Burbidge, Magee and 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
Figure 2: A comparison of the Box-Cox and the new transformations

Let

\[
\psi(\lambda_+, \lambda_-, x) = \begin{cases} 
  \frac{(x + 1)^{\lambda_+} - 1}{\lambda_+} & \text{for } x \geq 0, \lambda_+ \neq 0, \\
  \log(x + 1) & \text{for } x \geq 0, \lambda_+ = 0, \\
  -\frac{((-x + 1)^{\lambda_-} - 1)}{\lambda_-} & \text{for } x < 0, \lambda_- \neq 0, \\
  -\log(-x + 1) & \text{for } x < 0, \lambda_- = 0.
\end{cases}
\]

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, \(\psi(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}\), where

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

Then, by Lemma 1 below, \(\psi(\lambda, x)\) is concave in \(x\) for \(\lambda < 1\) and convex 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 generalized Box-Cox transformations, \(((x+1)^\lambda - 1)/\lambda, x > -1\), where the shift constant 1 is included. We first establish properties of the transformation (2.1).

**Lemma 1** The transformation function \(\psi(\cdot, \cdot)\) defined in (2.1) satisfies

\[\]

4
Figure 3: Plots of mixture density $f(\cdot)$ (solid lines), transformed density $g_\lambda(\cdot)$ by $\psi$ (dotted lines) and target normal density $\phi_{\mu,\sigma^2}(\cdot)$ (broken lines).

(a) $\psi(\lambda, x) \geq 0$ for $x \geq 0$ and $\psi(\lambda, x) < 0$ for $x < 0$.

(b) $\psi(\lambda, x)$ is convex in $x$ for $\lambda > 1$ and concave in $x$ for $\lambda < 1$.

(c) $\psi(\lambda, x)$ is a continuous function of $(\lambda, x)$.

(d) Let $\psi^{(k)} = \partial^k \psi(\lambda, x)/\partial \lambda^k$. Then, for $k \geq 1$,

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

is continuous in $(\lambda, x)$. Here $\psi^{(0)} \equiv \psi(\lambda, x)$.

(e) $\psi(\lambda, x)$ is increasing in both $\lambda$ and $x$.

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

5
The proofs are straightforward but tedious. Details are given by Yeo (1997).

The mixture density \( f(\cdot) \) in (1.3) looks skewed to the right so, according to van Zwet (1964), a good transformation should be concave in order to lead to a near symmetry. Following Hernandez and Johnson (1980), we show in Theorem 2 below that it is reasonable to select the transformation, \( \psi(\lambda, x) \), to minimize the Kullback-Leibler information number (1.4). By numerical integration, we obtain \( \lambda = 0.555 \). As expected, this transformation is concave according to (b) of Lemma 1. Figure 3 shows that the normal approximation is much improved since the transform has pulled down in the right tail and pushed out the left hand tail.

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 \( f(\cdot) \), to near normality.

3.1 Estimation method

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 \( \lambda \), the transformed observations can be treated as normally distributed with some mean \( \mu \) and variance \( \sigma^2 \). Under this assumption, the log-likelihood function is

\[
\ell_n(\theta|x) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\psi(\lambda, x_i) - \mu)^2 \\
+ (\lambda - 1) \sum_{i=1}^{n} (I(x_i > 0) \log(x_i + 1) - I(x_i < 0) \log(-x_i + 1)),
\]  

(3.1)

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

Holding \( \lambda \) fixed, we initially maximize \( \ell_n(\cdot, \cdot, \lambda|x) \) yielding

\[
\hat{\mu}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \psi(\lambda, x_i) \quad \text{and} \quad \hat{\sigma}^2(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\psi(\lambda, x_i) - \hat{\mu}(\lambda))^2.
\]  

(3.2)

Substituting (3.2) into (3.1), we obtain the profile log-likelihood

\[
\ell_{\text{max}}(\lambda|x) = -\frac{n}{2} \log(2\pi) + 1) - \frac{n}{2} \log(\hat{\sigma}^2(\lambda)) \\
+ (\lambda - 1) \sum_{i=1}^{n} (I(x_i > 0) \log(x_i + 1) - I(x_i < 0) \log(-x_i + 1)).
\]  

(3.3)

The maximum likelihood estimate, \( \hat{\lambda} \), of \( \lambda \) is obtained by maximizing (3.3) and then \( \hat{\theta} = (\hat{\mu}(\hat{\lambda}), \hat{\sigma}^2(\hat{\lambda}), \hat{\lambda})' \) maximizes the log-likelihood function (3.1).

3.2 Asymptotic results

Since the small sample distribution of \( \hat{\theta} \) is difficult to derive, we now study the large sample behavior of \( \hat{\theta} \). Similar to the Box-Cox transformation considered in Hernandez and Johnson (1980), the maximum likelihood estimators (MLE’s) are consistent and asymptotically normal.
THEOREM 1 Let $X_1, \ldots, X_n$ be independent and identically distributed with probability density function $f(\cdot)$. Suppose the parameter space $\Theta$, the true $f(\cdot)$, and the log-likelihood function $l_1(\theta|x)$ satisfy the following conditions

(i) The parameter space $\Theta$ is a compact set defined as

$$\Theta = \left\{ \theta = (\mu, \sigma^2, \lambda)^t \mid |\mu| \leq r, c \leq \sigma \leq d, a \leq \lambda \leq b, \right.$$ 

with $-\infty < a < b < \infty$, $0 < r < \infty$ and $0 < c < d < \infty \right\},$$

(ii) $E_f[l_1(\theta|X)]$ has a unique global maximum at $\theta_0 \in \Theta$,

(iii) $E_f[I_{(X<0)}(-X)^{2(a-2)}] < \infty$, and $E_f[I_{(X>0)}X^{2b}] < \infty$.

Then

(A) The MLE $\hat{\theta}$ is a strongly consistent estimator of $\theta_0$.

Furthermore, if

(iv) $\theta_0$ is an interior point of $\Theta$.

(v) $E_f[I_{(X<0)}(-X)^{2(a-2)}(\log(-X + 1))^2] < \infty$ and $E_f[I_{(X>0)}X^{2b}(\log(X + 1))^2] < \infty$,

(vi) $E_f[\nabla l_1(\theta_0|X)] = 0$, where the column vector $\nabla l_1(\theta_0|X) = \left( \frac{\partial l_1(\theta|X)}{\partial \theta_i} \right)_{\theta=\theta_0}$ is the gradient of the log-likelihood function for $\theta = (\theta_1, \theta_2, \theta_3)^t = (\mu, \sigma^2, \lambda)^t$,

(vii) $W(\theta_0) = E_f[\nabla^2 l_1(\theta_0|X) (\nabla l_1(\theta_0|X))^t] = -E_f[\nabla^2 l_1(\theta_0|X)]$ is non-singular, where $\nabla^2 l_1(\theta_0|X) = \left( \frac{\partial^2 l_1(\theta|X)}{\partial \theta_i \partial \theta_j} \right)_{\theta=\theta_0}$ is the Hessian of the log-likelihood function,

Then

(B) $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{D}} N_3(0, W^{-1}(\theta_0)).$

The proof is given in Appendix.

REMARK 1 When $a = 2$, $E_f[I_{(X<0)}(-X)^{2(a-2)}]$ in (iii) and $E_f[I_{(X<0)}(-X)^{2(a-2)}(\log(-X + 1))^2]$ in (v) are replaced by $E_f[I_{(X<0)}(\log(-X + 1))^2]$ and $E_f[I_{(X<0)}(\log(-X + 1))^4]$, respectively. Similarly, when $b = 0$, $E_f[I_{(X>0)}X^{2b}]$ and $E_f[I_{(X>0)}X^{2b}(\log(X + 1))^2]$ are replaced by $E_f[I_{(X>0)}(\log(X + 1))^2]$ and $E_f[I_{(X>0)}(\log(X + 1))^4]$, respectively.
**Theorem 2** Let \(X_1, \ldots, X_n, \ldots\) be independent and identically distributed with density function \(f(\cdot)\). Under the assumptions of Theorem 1,

\[
\hat{\theta} = (\hat{\mu}(\hat{\lambda}), \hat{\sigma}^2(\hat{\lambda}), \hat{\lambda})' \xrightarrow{a.s.} \theta_* = (\mu_*(\lambda_*), \sigma_*^2(\lambda_*), \lambda_*)',
\]

where \(\theta_*\) is the value of \(\theta\) that minimizes the Kullback-Leibler information given by (1.4).

Furthermore, \(\theta_* = \theta_0\), where \(\theta_0\) is the value of \(\theta\) in assumption (ii) of Theorem 1.

**Proof.** According to (A) of Theorem 1, with probability one,

\[
\lim_{n \to \infty} \left( \max_{\theta \in \Theta} \frac{1}{n} l_n(\theta | X) \right) = E_f[l_1(\theta | X)]
\]

\[
= \max_{\theta \in \Theta} E_f[\log(f_{\mu,\sigma^2}(\psi(\lambda, X))) + (\lambda - 1)(I_{X \geq 0}(X + 1) - I_{X < 0}(\log(-X + 1))].
\]

Next,

\[
\min_{\theta \in \Theta} KL(g_{\lambda}; f_{\mu,\sigma^2}) = E_f[\log(f(X))]
\]

\[
- \max_{\theta \in \Theta} E_f[\log(f_{\mu,\sigma^2}(\psi(\lambda, X))) + (\lambda - 1)(I_{X \geq 0}(X + 1) - I_{X < 0}(\log(-X + 1))].
\]

Thus, the value of \(\theta\) in \(\Theta\) that maximizes \(E_f[l_1(\theta | X)]\) is the same value that minimizes \(KL(g_{\lambda}; f_{\mu,\sigma^2})\). Therefore, \(\hat{\theta} \xrightarrow{a.s.} \theta_*\) and \(\theta_* = \theta_0\). \(\square\)

4. **Transforming to near symmetry**

It is well known that the MLE of the Box-Cox 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 (1965) have suggested methods for estimating \(\lambda\) in the Box-Cox transformation when the goal is to obtain approximate symmetry rather than normality. We employ the empirical characteristic function for estimating the transformation parameter \(\lambda\). Since we assume that the transformed variables are approximately symmetrically distributed around an arbitrary center, this approach may not be applied when using the Box-Cox transformations; at least not without adding a data dependent translation parameter.

4.1 **Estimation method**

Let \(X_1, X_2, \ldots\), be independent and identically distributed random variables each with distribution function \(F(\cdot)\) and characteristic function \(\phi(t)\). Let \(F_n(x)\) be the empirical distribution function based on \(X_1, X_2, \ldots, X_n\), and let \(\phi_n(t) = \phi_n(t, X_1, \ldots, X_n)\) be the empirical characteristic function (ECF)

\[
\phi_n(t) = \int \exp(\imath tx) \, dF_n(x) = n^{-1} \sum_{j=1}^n \exp(\imath tX_j) = \phi_{cn}(t) + \imath \phi_{sn}(t),
\]
where
\[
\phi_{cn}(t) = n^{-1} \sum_{j=1}^{n} \cos(tX_j) \quad \text{and} \quad \phi_{sn}(t) = n^{-1} \sum_{j=1}^{n} \sin(tX_j).
\] (4.1)

Since the characteristic function of a random variable \(X\) is real if and only if \(X\) is symmetrically distributed around zero, we choose to look for a transform for which the imaginary part is, in some sense, small. In particular, we consider minimizing the integrated square of the imaginary part of the ECF,
\[
\varphi_n = \int \phi_{sn}^2(t) \, dG(t),
\]
where \(G(\cdot)\) is some specified weight function and \(\phi_{sn}\) is defined in (4.1). Feuerverger and Mureika (1977) proposed the statistic \(\varphi_n\) in the context of testing for symmetry. For the general cases where the center of symmetry is not specified, a generalized statistic,
\[
\varphi_n = \int [\text{Im}\{\exp(-it\mu)\phi_n(t)\}]^2 \, dG(t),
\] (4.2)
is considered.

Our approach is to transform \(X\) according to \(\psi(\lambda, X)\) and then to select \(\lambda\) and \(\mu\) to make (4.2) small. That is, we minimize the integrated square of the imaginary part of the ECF of \(\psi(\lambda, X_1), \ldots, \psi(\lambda, X_n)\) with factor \(\exp(-it\mu)\). Usually, in a given instance, it may not be possible to select a \(\lambda\) 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.

**Assumption** For some \(\lambda\), the distribution \(\psi(\lambda, X_1)\) is symmetric about a center \(\mu\).

To access the quality of the transformation \(\psi(\lambda, X)\), we let \(\phi_n(\lambda, t)\) be the empirical characteristic function of transformed variables \(\psi(\lambda, X_1), \ldots, \psi(\lambda, X_n)\),
\[
\phi_n(\lambda, t) = n^{-1} \sum_{j=1}^{n} \exp(it\psi(\lambda, X_j)) = \phi_{cn}(\lambda, t) + i\phi_{sn}(\lambda, t),
\]
where \(\phi_{cn}(\lambda, t) = n^{-1} \sum_{j=1}^{n} \cos(t\psi(\lambda, X_j))\) and \(\phi_{sn}(\lambda, t) = n^{-1} \sum_{j=1}^{n} \sin(t\psi(\lambda, X_j))\). Define the deviation measure
\[
\varphi_n(\theta) = \int [\text{Im}\{\exp(-it\mu)\phi_n(\lambda, t)\}]^2 \, dG(t)
= \int [\cos(-t\mu)\phi_{sn}(\lambda, t) + \sin(-t\mu)\phi_{cn}(\lambda, t)]^2 \, dG(t)
= \int [n^{-1} \sum_{j=1}^{n} \sin(t(\psi(\lambda, X_j) - \mu))]^2 \, dG(t),
\]
where \(\theta = (\lambda, \mu)'\) and \(G(\cdot)\) is some symmetric distribution function, so \(G(t^-) + G(-t^-) = 1\) for all \(t \in \mathbb{R}\), and which satisfies \(G(t) - G(-t) > 0\) for any \(t > 0\). Further, let \(G(\cdot)\) have a characteristic function \(\nu(\cdot)\). Then, \(\sin(x)\sin(y) = \{\cos(x - y) - \cos(x + y)\} / 2\) so we can write
\[
\varphi_n(\theta) = (2n^2)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\nu[\psi(\lambda, X_i) - \psi(\lambda, X_j)] - \nu[\psi(\lambda, X_i) + \psi(\lambda, X_j) - 2\mu]\}. \] (4.3)
If the distribution of $\psi(\lambda, X)$ is symmetric about $\mu$, then the imaginary part of the characteristic function of $\psi(\lambda, X)$ with $\exp(-it\mu)$ will be zero. We propose selecting the value $\theta_n$ which minimizes $\varphi_n(\theta)$.

**Remark 2** To simplify the calculations, we may consider the generalized alternative power transformation which has shift parameter $\mu$,

$$\psi(\lambda, \mu, x) = \begin{cases} \frac{(x - \mu + 1)^\lambda - 1}{\lambda} & \text{if } x \geq \mu, \lambda \neq 0, \\ \log(x - \mu + 1) & \text{if } x \geq \mu, \lambda = 0, \\ \frac{(-x + \mu + 1)^{2-\lambda} - 1}{(2-\lambda)} & \text{if } x < \mu, \lambda \neq 2, \\ -\log(-x + \mu + 1) & \text{if } x < \mu, \lambda = 2. \end{cases} \quad (4.4)$$

Then $\mu$ may be estimated by the sample median or other smoothed measure of median.

### 4.2 Asymptotic results

As part of our development, we first establish the almost sure uniform convergence of $U$-statistics. This result extends Rubin (1956).

**Theorem 3** Let $X_1, \ldots, X_n, \ldots$ be a sequence of independent and identically distributed random variables with a common distribution function $F$. Let $\eta(x; \theta)$ be a symmetric kernel, which depends on unknown parameters $\theta$, and which is a measurable function in $x = (x_1, \ldots, x_k)^T \in \mathbb{R}^k$ for each $\theta \in \Theta$. Let $\Theta$ be a compact topological space. Consider

$$U_{kn}(x, \theta) = \binom{n}{k}^{-1} \sum_{(n,k)} \eta(x_{i(1)}, \ldots, x_{i(k)}; \theta), \quad (4.5)$$

where the summation $\sum_{(n,k)}$ is taken over all subsets $1 \leq i(1) < \cdots < i(k) \leq n$ of $\{1, 2, \ldots, n\}$.

Set $\eta_j(x_1, \ldots, x_j; \theta) = E_F[\eta(x_1, \ldots, x_j, X_{j+1}, \ldots, X_k; \theta)]$ for $j = 1, \ldots, k$ and assume that

(i) there is an integrable and symmetric kernel $g$ such that, for all $\theta \in \Theta$ and $x \in \mathbb{R}^k$,

$$|\eta(x; \theta)| \leq g(x) \quad (4.6)$$

(ii) there is a sequence $S^n_M$ of measurable sets such that

$$P(\mathbb{R}^k - \bigcup_{M=1}^{\infty} S^n_M) = 0, \quad (4.7)$$

(iii) for each $M$ and for all $j = 1, \ldots, k$, $\eta_j(x_1, \ldots, x_j; \theta)$ is equicontinuous in $\theta$ for $(x_1, \ldots, x_j) \in S^n_M$, where $S^n_M = S^n_M \times S^{k-j}_M$. 

10
Then
\[ U_{kn}(X, \theta) \xrightarrow{a.s.} EF[\eta(X; \theta)] \] (4.8)

uniformly in \( \theta \in \Theta \) and \( EF[\eta(X; \theta)] \) is continuous.

The proof is given in Appendix.

**Lemma 2.** Suppose that the parameter space \( \Theta \) is a compact set. Then
\[ \varphi_n(\theta) \xrightarrow{a.s.} \varphi(\theta) = \int \phi_{\theta}^2(\theta, t) dG(t) \] (4.9)

uniformly in \( \theta = (\lambda, \mu)' \in \Theta \), where \( \phi_{\theta}(\theta, t) = EF[\sin(t(\psi(\lambda, X) - \mu))] \). Further \( \varphi(\theta) \) is continuous in \( \theta \).

**Proof:** Denote \( \tau(\theta, x, y) = \nu[\psi(\lambda, x) - \psi(\lambda, y)] - \nu[\psi(\lambda, x) + \psi(\lambda, y) - 2\mu] \). From (4.3), we obtain that
\[ \varphi_n(\theta) = (2n^2)^{-1} \sum_{j=1}^n \tau(\theta, X_j, X_j) + \frac{n-1}{2n} U_n(\lambda), \] (4.10)

where \( U_n(\theta) = \left( \frac{n}{2} \right)^{-1} \sum_{j=1}^{n-1} \sum_{k=j+1}^n \tau(\theta, X_j, X_k) \). Letting \( S_M = [-M, M] \), \( P(R - \bigcup_{M=1}^\infty S_M) = 0 \). Because \( \tau(\theta, x, x) \) is bounded and continuous in \((\theta, x) \in \Theta \times S_M\), it is equicontinuous in \( \theta \) for \( x \in S_M \). By Rubin (1956),
\[ n^{-1} \sum_{j=1}^n \tau(\theta, X_j, X_j) \xrightarrow{a.s.} EF[\tau(\theta, X_1, X_1)] \] uniformly in \( \theta \in \Theta \).

To apply **Theorem 3** to \( U_n(\theta) \), we define
\[ \eta(x_1, x_2; \theta) = \tau(\theta, x_1, x_2) = 2 \int \sin(t(\psi(\lambda, x_1) - \mu)) \sin(t(\psi(\lambda, x_2) - \mu)) dG(t) \]
and \( \eta_1(x_1; \theta) = EF[\eta(x_1, X; \theta)] \). Then, \( \eta(x_1, x_2; \theta) \) is a symmetric kernel that is bounded for all \( \theta \in \Theta \). Setting \( S_M^2 = [-M, M] \times [-M, M] \), we have that \( P(R^2 - \bigcup_{M=1}^\infty S_M^2) = 0 \). Since \( \eta(x_1, x_2; \theta) \) and \( \eta_1(x_1; \theta) \) are continuous on \((\theta, x_1, x_2) \in \Theta \times S_M^2\), and \((\theta, x_1) \in \Theta \times [-M, M]\), respectively, both \( \eta(\cdot, \cdot; \theta) \) and \( \eta_1(\cdot; \theta) \) are equicontinuous in \( \theta \). By **Theorem 3**
\[ U_n(\theta) \xrightarrow{a.s.} 2 \int EF^2[\sin(t(\psi(\lambda, X) - \mu))] dG(t) = 2\varphi(\theta) \]
uniformly in \( \theta \) and \( \varphi(\theta) \) is continuous. Therefore, \( \varphi_n(\theta) \xrightarrow{a.s.} \varphi(\theta) \) uniformly in \( \theta \in \Theta \). □

**Theorem 4.** Let \( \nu(\cdot) \) be the characteristic function of the distribution function \( G(\cdot) \), where \( G(t^-) + G(-t) = 1 \) for all \( t \in \mathbb{R} \) and \( G(t) - G(-t) > 0 \) for any \( t > 0 \). Suppose
(i) the parameter space $\Theta$ is a compact set defined as
\[ \Theta = \{ (\lambda, \mu) \mid a \leq \lambda \leq b \text{ and } c \leq \mu \leq d \} \]
with $-\infty < a < b < \infty$ and $-\infty < c < d < \infty$,

(ii) let $\varphi(\theta)$, defined by (4.9), have unique global minimum at $\theta_0 = (\lambda_0, \mu_0)'$. Then
\[ (A) \quad \hat{\theta}_n = \arg \min \varphi_n(\theta) \text{ is a strongly consistent estimator of } \theta_0. \]

Furthermore, if

(iii) $\theta_0$ is an interior point of $\Theta$,

(iv) $E_F[I_{X<0}(-X)^{2(2-a)}(\log(-X+1))^2]$, $E_F[I_{X \geq 0}X^{2b}(\log(X+1))^2]$ and $\int t^2 dG(t)$ are finite,

(v) $\nabla \varphi(\theta_0) = 0$,

(vi) $\nabla^2 \varphi(\theta_0)$ is non-singular,

then
\[ (B) \quad \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{L} N_2(0, V(\theta_0) \Sigma(\theta_0) V(\theta_0)'), \]
where $V(\theta_0) = (\nabla^2 \varphi(\theta_0))^{-1}$ and $\Sigma(\theta_0) = E_{X_2} [E_{X_1} [\nabla \tau(\theta_0, X_1, X_2)'(\nabla \tau(\theta_0, X_1, X_2))]].$

The proof is given in Appendix.

Remark 3 When $a = 2$, $E_F[I_{X<0}(-X)^{2(2-a)}(\log(-X+1))^2]$ in (iv) is replaced by $E_F[I_{X<0}(\log(-X+1))^4]$. When $b = 0$, $E_F[I_{X \geq 0}X^{2b}(\log(X+1))^2]$ is replaced by $E_F[I_{X \geq 0}(\log(X+1))^4]$.

Remark 4 Note that $E_F[\sin(t(\psi(\lambda_0, X) - \mu_0))] = 0$ for all $t$ when $\psi(\lambda_0, X)$ is exactly symmetric about $\mu_0$. Since $\varphi(\theta) = \int E_F^2[\sin(t(\psi(\lambda, X) - \mu))]dG(t)$, if $\psi(\lambda_0, X)$ is symmetric around $\mu_0$ and $G(\cdot)$ is not degenerate at $t = 0$, then
\[ \varphi(\theta_0) = \int E_F^2[\sin(t(\psi(\lambda_0, X) - \mu))]dG(t) = 0, \]
\[ \nabla \varphi(\theta_0) = 2 \int tE_F[\sin(t(\psi(\lambda, X) - \mu))] \]
\[ \quad \quad \quad E_F[\cos(t(\psi(\lambda, X) - \mu))] \nabla \varphi(\lambda_0, X) \mu_0]dG(t) = 0, \]
\[ \nabla^2 \varphi(\theta_0) = 2 \int t^2 E_F[\cos(t(\psi(\lambda, X) - \mu))] \nabla \varphi(\lambda_0, X) \mu_0] \]
\[ \quad \quad \quad E_F[\nabla^2 \psi(\lambda_0, X) \mu_0]dG(t). \]
\[ E_{X_2}[E_{X_1}[\nabla \tau(\theta_0, X_1, X_2)(\nabla \tau(\theta_0, X_1, X_2))]] \]

\[ = 4E_{X_2}[\left\{ \int t \sin(t(\psi(\lambda_0, X_2) - \mu_0))E_{X_1}[\cos(t(\psi(\lambda_0, X_1) - \mu_0))[\nabla(\psi(\lambda_0, X_1) - \mu_0)]dG(t) \right\} \nabla(\psi(\lambda_0, X_1) - \mu_0)]dG(t)]'. \]

4.3 Choices 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:

\[ g(t) = (2\pi\sigma^2)^{-1/2} \exp(-t^2/(2\sigma^2)) \quad -\infty < t < \infty, \]
\[ g(t) = (2\Gamma(\alpha)\sigma^\alpha)^{-1}|t|^\alpha-1 \exp(-|t|/\sigma) \quad -\infty < t < \infty, \]
\[ g(t) = 1/(2\sigma) \quad -\sigma < t < \sigma. \]

These weight distributions satisfy the conditions in Theorem 4. Note that each of the three weight distribution families is indexed by a scale parameter \( \sigma > 0 \), so we write

\[ \varphi_n(\theta) = \varphi_n(\theta, \sigma) = \int_{-\infty}^{\infty} \varphi_{sn}(\theta, t) \ dG(t, \sigma), \]

where \( \varphi_{sn}(\theta, t) = n^{-1} \sum_{j=1}^{n} \sin(t(\psi(\lambda, x_j) - \mu)) \). 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 \( \sigma \) is small. For \(-\delta \leq t \leq \delta\) such that \( \delta \) is sufficiently small, \( \varphi_{sn}^2(\theta, t) \) is successfully approximated by simpler function, that is,

\[ \varphi_{sn}(\theta, t) = t^2 \left( n^{-1} \sum_{j=1}^{n} \psi(\lambda, x_j) - \mu \right)^2 + O(t^4). \]

Hence, if we take \( \sigma \) small sufficiently, then, for small values of \( \delta \),

\[ \varphi_n(\theta, \sigma) \sim \left( n^{-1} \sum_{j=1}^{n} \psi(\lambda, x_i) - \mu \right)^2 \int_{-\delta}^{\delta} t^2 \ dG(t, \sigma). \]

That is, for a wide choice of weight distributions, the particular choice does not affect the estimation of \( \lambda \). In these cases, our approach is equivalent to minimizing \( \left( n^{-1} \sum_{j=1}^{n} \psi(\lambda, x_j) - \mu \right)^2 \), if \( \sigma \) is sufficiently small. However, under the limiting formulation some moments conditions,

\[ E_F[\psi(\lambda, X)] < \infty, \]

are required so we retain our original formulation and suggest that small \( \sigma \) be used.

4.4 Influence of outliers

The influence function serves to describe the effect of an outlier on estimation. Suppose that the estimator can be written as a function \( T(F_0) \). Then, the influence function evaluated at a point \( x_0 \) is defined as

\[ IF(x_0, F) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F + \varepsilon\delta(x_0)] - T(F)}{\varepsilon}, \]
Table 1: The Monte Carlo bias, standard deviation (S.D.) 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 $n = 30$.

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_0 = 0.0$</th>
<th></th>
<th>$\lambda_0 = 0.5$</th>
<th></th>
<th>$\lambda_0 = 1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLE</td>
<td>MSEC</td>
<td>MLE</td>
<td>MSEC</td>
<td>MLE</td>
</tr>
<tr>
<td>$N(0, 3^2)$</td>
<td>bias</td>
<td>0.020</td>
<td>-0.005</td>
<td>0.010</td>
<td>-0.008</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>0.086</td>
<td>0.132</td>
<td>0.123</td>
<td>0.169</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.008</td>
<td>0.017</td>
<td>0.015</td>
<td>0.029</td>
</tr>
<tr>
<td>$N(0, 3^2)+$</td>
<td>bias</td>
<td>-0.125</td>
<td>-0.109</td>
<td>-0.339</td>
<td>-0.191</td>
</tr>
<tr>
<td></td>
<td>S.D.</td>
<td>0.018</td>
<td>0.087</td>
<td>0.059</td>
<td>0.112</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.016</td>
<td>0.019</td>
<td>0.119</td>
<td>0.049</td>
</tr>
</tbody>
</table>

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 $\psi^{(1)}$ is defined in (d) of LEMMA 1. 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 respectively proportional to $x^2 \log(|x| + 1)$ and $|x| \log(|x| + 1)$.

Our estimate $\hat{\lambda}_n$ can be written as the solution to

$$
\int t \left[ \int \cos(t\psi(\hat{\lambda}_n, x))\psi^{(1)}(\hat{\lambda}_n, x)F_n(x) \right] \left[ \int \sin(t\psi(\hat{\lambda}_n, y))dF_n(y) \right] dG(t) = 0.
$$

This equation forces us to consider the influence of a point $(x, y)$. For $\lambda_0 = 1$, we obtain that the influence function of our estimation is proportional to $|x| \log(|x| + 1) \int t \sin(ty) \cos(tx) dG(t)$

Applying a weight distribution $G(\cdot)$ having finite moments, e.g. standard normal, we see that our estimate of $\lambda$ is also sensitive, but less sensitive than the normal maximum likelihood estimate, to a outlier. Also, it can be easily showed that the influence of an outlier on estimating $\mu$ is bounded.

To investigate the influence of a single outlier, we used a simulation approach. A series of 10000 replications, of samples of size 30, were generated for $\lambda_0 = 0.0$, 0.5 and 1.5 according to $X^{(\lambda_0)} \sim N(0, 3^2)$. The normal weight distribution with $\sigma = 0.01$ was employed. Table 1 gives the simulated bias, standard deviation and mean squared error of the two estimates for $\lambda_0$. Maximum likelihood methods perform well for all cases. In the row $N(0, 3^2)+$, we added the positive outlier $x_{31}^{(\lambda_0)} = 30$ to each data set of $N(0, 3^2)$. It is shown that the MLE’s are strongly affected by the outlier while the MSEC estimates are less sensitive to the outlier.
APPENDIX: proofs

The following lemma is used as auxiliary result to prove the strong consistency of \( \hat{\theta} \) establish in Theorem 1 and Theorem 4. Since it is a standard result, we omit the proof.

**Lemma A.1** Let \( \{T_n(\cdot)\} \) be a sequence of random functions defined on a probability space and which depend on \( \theta \in \Theta \), a compact set. Suppose that

(i) There exist a continuous function \( T(\theta) \) defined on \( \Theta \) such that \( T_n(\theta) \overset{a.s.}{\rightarrow} T(\theta) \) uniformly in \( \theta \in \Theta \),

(ii) \( T(\theta) \) has a unique minimum (or maximum) at \( \theta_0 \in \Theta \).

Then \( \hat{\theta}_n = \arg\min T_n(\theta) \) (or \( \hat{\theta}_n = \arg\max T_n(\theta) \)) is a strongly consistent estimator of \( \theta_0 \).

**Proof of Theorem 1.** (A) Since \( l_1(\theta|x) \) is dominated by an integrable function and is equicontinuous in \( \theta \) for \( x \in S_M \) where \( S_M = [-M, M] \) and \( P(R - \cup_{M=1}^{\infty} S_M) = 0 \), the convergence of \( \hat{\theta}_n = \arg\min \{ l_1(\theta|x) \} \) in \( \Theta \) is unique by assumption (ii), LEMMA A.1 leads to conclusion that \( \hat{\theta} \overset{a.s.}{\rightarrow} \theta_0 \).

(B) Expanding \( n^{-1/2}\nabla l_n(\theta|X) \) about \( \theta_0 \), we obtain that

\[
n^{-1/2}\nabla l_n(\theta|X) = n^{-1/2}\nabla l_n(\theta_0|X) + n^{-1}\nabla^2 l_n(\theta_0|X)\sqrt{n}\left(\theta - \theta_0\right), \tag{A.1}\]

where \( \theta_0 = c_n\hat{\theta} + (1 - c_n)\theta_0 \), \( c_n \in (0,1) \). Since \( n^{-1/2}\nabla l_n(\theta_0|X) = 0 \) at the maximum, the left hand side of (A.1) converges to zero in probability. By the central limit theorem, \( n^{-1/2}\nabla l_n(\theta_0|X) \) is asymptotically normal with mean \( 0 \) and variance \( W(\theta_0) \). Another application of Rubin and the consistency of \( \hat{\theta} \) ensure that \( -n^{-1}\nabla^2 l_n(\theta_0|X) \overset{P}{\rightarrow} -E_I[\nabla^2 l_1(\theta_0|X)] = W(\theta_0) \). By Slutsky’s theorem, we conclude that

\[
\sqrt{n}\left(\theta - \theta_0\right) \overset{L}{\rightarrow} N_3(0, W^{-1}(\theta_0)). \]

**Proof of Theorem 3.** Without loss of generality, we can take the \( S_M^k \) to be monotonically increasing. As in Lee (1990, page 26), we define symmetric kernels of degrees, 1, 2, ..., \( k \),

\[
\begin{align*}
h_1(x_1; \theta) &= \eta_1(x_1; \theta) - E_F[\eta(X; \theta)] \\
h_2(x_1, x_2; \theta) &= \eta_2(x_1, x_2; \theta) - h_1(x_1; \theta) - h_1(x_2; \theta) - E_F[\eta(X; \theta)] \\
&\vdots \\
h_k(x; \theta) &= \eta_k(x; \theta) - \sum_{j=1}^{k-1} \sum_{(i,j)} h_j(x_{i(1)}, \ldots, x_{i(j)}; \theta) - E_F[\eta(X; \theta)]. \tag{A.2}\end{align*}
\]
Then, by the \( H \)-decomposition of Hoeffding, we obtain

\[
U_{kn}(\mathbf{x}; \theta) = \left( \frac{n}{k} \right)^{-1} \sum_{(n,k)} \eta(x_{i(1)}, \ldots, x_{i(k)}; \theta) = E_F[\eta(X; \theta)] + \sum_{j=1}^{k} \left( \binom{k}{j} \right) H_{jn}(\theta),
\]

where \( H_{jn}(\theta) = \left( \binom{n}{j} \right)^{-1} \sum_{(n,j)} h_j(x_{i(1)}, \ldots, x_{i(j)}; \theta) \) is a U-statistic of degree \( j \) based on kernel \( h_j \).

To verify the uniform convergence of \( U_{kn}(X; \theta) \), it suffices to show that for all \( j = 1, \ldots, k \),

\[
H_{jn}(\theta) \xrightarrow{a.s.} 0 \quad \text{uniformly in } \theta.
\]

Let \( h_j^* = h_j + E_F[\eta(X; \theta)] \) so \( h_j^* \) is equicontinuous by (iii). Then, the assumptions (i) - (iii) imply those of Rubin (1956), so

\[
n^{-1} \sum_{i=1}^{n} h_j^* (X_i; \theta) = n^{-1} \sum_{i=1}^{n} \eta_1 (X_i; \theta) \xrightarrow{a.s.} E_F[\eta_1(X_1; \theta)]
\]

(A.3)

uniformly in \( \theta \) and \( E_F[\eta_1(X_1; \theta)] \) is continuous. But, by the property of U-statistic, \( E_F[\eta_1(X_1; \theta)] = E_F[\eta(X; \theta)] \).

Setting \( h_j^* = h_j + E_F[\eta(X; \theta)] \), we have \( h_j^* (\theta) = h_j^* (x_1, \ldots, x_j; \theta) - \sum_{l=1}^{j-1} \sum_{(l,d)} h_l(x_{i(1)}, \ldots, x_{i(l)}; \theta) \). Assumption (iii) and the continuity of \( E_F[\eta(X; \theta)] \) imply the equicontinuity of \( h_j^* (\theta), \ j = 2, \ldots, k \). Let \( \varepsilon > 0 \) be given. Then, the equicontinuity of \( h_j^* (\theta) \) and compactness of \( \Theta \) allow us to select \( \Theta_1, \ldots, \Theta_p \) and open subsets \( \Theta_1, \ldots, \Theta_p \) of \( \Theta \) such that \( \Theta \subseteq \bigcup_{l=1}^{p} \Theta_l, \ \theta_l \in \Theta_l \) and \( |h_j^* (\theta) - h_j^* (\theta_l)| < \varepsilon/4 \) for \( \theta \in \Theta_l \) and \((x_1, \ldots, x_j) \in S^l_M \).

Write \( h_{ij}^* (\theta) = h_j^* (X_{i(1)}, \ldots, X_{i(j)}; \theta) \). For \( j \geq 2 \), \( H_{jn}(\theta) = \left( \binom{n}{j} \right)^{-1} \sum_{(n,j)} h_j(x_{i(1)}, \ldots, x_{i(j)}; \theta) \) is a U-statistic, and applying (i) iteratively to the \( h_j (\cdot; \theta) \) in (A.2), we have, for \( j = 2, \ldots, k \), \( E[h_j(X_1, \ldots, X_j; \theta)] < \infty \). According to the strong law of large numbers for U-statistic (see Lee (1990), page 122), \( \left( \binom{n}{j} \right)^{-1} \sum_{(n,j)} h_{ij}^* (\theta_l) \xrightarrow{a.s.} E_F[\eta(X; \theta_l)] \).

Hence, for every \( \varepsilon > 0 \), we may choose an \( N = N_\varepsilon \) such that for every \( \delta > 0 \),

\[
P \left( \left| \left( \binom{n}{j} \right)^{-1} \sum_{(n,j)} h_{ij}^* (\theta_l) - E_F[\eta(X; \theta_l)] \right| \geq \varepsilon/4 \text{ for } n \geq N \right) < \frac{\delta}{2^p}.
\]

(A.4)

for \( l = 1, \ldots, p \).

Next, consider \( g(\mathbf{x}) \) in (i). Let \( g_j(x_1, \ldots, x_j) = E_F[g(x_1, \ldots, x_j, X_{j+1}, \ldots, X_k)] \) for \( j = 1, \ldots, k \). Recursively define \( g_j^*(x_1, \ldots, x_j) \) starting with \( g_1^*(x_1) = g_1(x) \) and \( g_2^*(x_1, \ldots, x_j) = g_j(x_1, \ldots, x_j) + \sum_{l=1}^{j-1} \sum_{(l,d)} g_l^*(x_{i(1)}, \ldots, x_{i(l)}) \). Then, \( g_j^* \) is symmetric in its arguments for \( j \geq 2 \).

Define

\[
Z_{ij} = \begin{cases} 
0 & \text{if } (X_{i(1)}, \ldots, X_{i(j)}) \in S^j_M, \\
0 & \text{otherwise.}
\end{cases}
\]
Since \( \left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} Z_{ij} \) is a U-statistic, whose kernel \( g^*_j \) is integrable, another application of the strong law of large numbers for U-statistics allows us to conclude that, for \( M \) sufficiently large,

\[
\left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} Z_{ij} \overset{a.s.}{\rightarrow} E_F[Z_{ij}] = \int \cdots \int_{\mathbb{R}^j} g^*_j(x_1, \ldots, x_j) \, dF(x_1, \ldots, x_j) < \varepsilon/5.
\]

Thus, for every \( \varepsilon > 0 \), we may select an \( N_2 \) such that for every \( \delta > 0 \),

\[
P\left( \left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} Z_{ij} > \varepsilon/4, \text{ for some } n \geq N_2 \right) < \delta/2. \tag{A.5}
\]

Note that if \( \theta \in \Theta_l \), \( \left| h^*_j(\theta) - h^*_j(\theta_l) \right| < \varepsilon/4 + 2Z_{ij} \) because

\[
\left| h^*_j(\theta) - h^*_j(\theta_l) \right| = \begin{cases} < \varepsilon/4 & \text{if } (X_{i(1)}, \ldots, X_{i(j)}) \in S^j, \\ \leq 2g^*_j(X_{i(1)}, \ldots, X_{i(j)}) < \varepsilon/4 + 2Z_{ij} & \text{otherwise.} \end{cases}
\]

Let \( N = \max(N_1, N_2) \). Then, for every \( \varepsilon > 0 \) and \( \delta > 0 \),

\[
P\left( \left| \left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} h^*_j(\theta) - E_F[\eta(X; \theta_l)] \right| \geq \varepsilon, \text{ for some } n \geq N \text{ and } \theta, \theta_l \in \Theta_l \text{ some } l \right)
\]

\[
\leq P\left( \left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} Z_{ij} \geq \varepsilon/4, \text{ for some } n \geq N \right) + P\left( \left| \left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} h^*_j(\theta_l) - E_F[\eta(X; \theta_l)] \right| \geq \varepsilon/4, \text{ for some } n \geq N \text{ and } \theta_l \in \Theta_l \text{ some } l \right)
\]

\[
\leq \delta/2 + \sum_{l=1}^{p} \delta/2p = \delta,
\]

by (A.4) and (A.5). This implies that, for each \( j = 2, \ldots, k \),

\[
\left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} h^*_j(X_{i(1)}, \ldots, X_{i(j)}; \theta) \overset{a.s.}{\rightarrow} E_F[\eta(X; \theta)] \tag{A.6}
\]

uniformly in \( \theta \in \Theta \). Consequently, by (A.3) and (A.6),

\[
H_{jn}(\theta) = \left( \frac{n}{j} \right)^{-1} \sum_{(n,j)} h^*_j(X_{i(1)}, \ldots, X_{i(j)}; \theta) \overset{a.s.}{\rightarrow} 0
\]

uniformly in \( \theta \in \Theta \) for all \( j = 1, \ldots, k \). Here, \( E_F[\eta(X; \theta)] \) is continuous in \( \theta \) by (A.3). \( \square \)

Proof of Theorem 4. \( \textbf{(A)} \) According to LEMMA 2, \( \varphi_n(\theta) \overset{a.s.}{\rightarrow} \varphi(\theta) \) uniformly in \( \theta \) and \( \varphi(\theta) \) is continuous. By assumption (ii), \( \theta_0 \) minimizes \( \varphi(\theta) \) and is unique. Consequently, LEMMA 1 allows us to conclude that \( \theta_n \overset{a.s.}{\rightarrow} \theta_0 \).

\( \textbf{(B)} \) To establish the asymptotic normality of \( \hat{\theta}_n \), we expand the product \( \sqrt{n} \) times the first derivative \( \nabla \varphi_n(\hat{\theta}_n) \) about \( \theta_0 \).

\[
\sqrt{n} \nabla \varphi_n(\hat{\theta}_n) = \sqrt{n} \nabla \varphi_n(\theta_0) + \nabla^2 \varphi_n(\hat{\theta}_n) \sqrt{n}(\theta_n - \theta_0), \tag{A.7}
\]
where $\tilde{\theta}_n = \alpha_n \hat{\theta}_n + (1 - \alpha_n) \theta_0$ for $\alpha_n \in [0, 1]$ and $n \geq 1$. Since $\sqrt{n} \nabla \varphi(\hat{\theta}_n) \to 0$ at the minimum when $\theta_n$ lies in the interior of $\Theta$, 

$$
\sqrt{n} \nabla \varphi_n(\theta_0) - (-\nabla^2 \varphi_n(\theta_0)) \sqrt{n}(\theta_n - \theta_0) \xrightarrow{p} 0.
$$

(A.8)

Decomposing the first term of left hand side in (A.8), we see, from (4.10), that

$$
\sqrt{n} \nabla \varphi_n(\theta_0) = \frac{1}{2n^{3/2}} \sum_{j=1}^{n} \nabla \tau(\theta_0, X_j, X_j) + \frac{n-1}{2n} \sqrt{n} U_{1n}(\theta_0),
$$

(A.9)

where $U_{1n}(\theta_0) = \left(\frac{n}{2}\right)^{-1} \sum_{j=1}^{n} \sum_{k=1}^{n} \nabla \tau(\theta_0, X_j, X_k)$. The strong law of large numbers ensures that

$$
n^{-1} \sum_{j=1}^{n} \nabla \tau(\theta_0, X_j, X_j) \xrightarrow{a.s.} E\tau[\nabla \tau(\theta_0, X_1, X_1)] < \infty,
$$

so the first term on the right hand side of (A.9) converges to 0 almost surely. Since $U_{1n}(\theta_0)$ is a Hoeffding $U$-statistic, the multivariate central limit theorem for $U$-statistics (see Lee (1990), page 76) gives

$$
\sqrt{n} U_{1n}(\theta_0) \xrightarrow{L} N_2(\mu(\theta_0), 4 \Sigma(\theta_0)),
$$

(A.10)

where $\mu(\theta_0) = E\tau[\nabla \tau(\theta_0, X_1, X_2)] = 2 \nabla \varphi(\theta_0) = 0$ and

$$
\Sigma(\theta_0) = E_{X_2}[(\nabla \tau(\theta_0, X_1, X_2))(\nabla \tau(\theta_0, X_1, X_2))^\top].
$$

Slutsky's theorem allows us to conclude from (A.9) and (A.10) that

$$
\sqrt{n} \nabla \varphi_n(\theta_0) \xrightarrow{L} N_2(0, \Sigma(\theta_0)).
$$

(A.11)

Finally, we show that the remainder term $\nabla^2 \varphi_n(\tilde{\theta}_n)$ converges to $\nabla^2 \varphi(\theta_0)$ in probability. From (A.9), $\nabla^2 \varphi_n$ can be written as

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

(A.12)

where $U_{2n}(\theta) = \left(\frac{n}{2}\right)^{-1} \sum_{j=1}^{n} \sum_{k=1}^{n} \nabla^2 \tau(\theta, X_j, X_k)$. Then, application of Rubin's theorem gives

$$
n^{-1} \sum_{j=1}^{n} \nabla^2 \tau(\theta, X_j, X_j) \xrightarrow{a.s.} E\tau[\nabla^2 \tau(\theta, X_1, X_1)] < \infty
$$

(A.13)

uniformly in $\theta \in \Theta$. Hence the first term on the right hand side of equation (A.12) converges to zero almost surely. To apply Theorem 3 to $U_{2n}(\theta)$, we define

$$
\eta(x_1, x_2; \theta) = \nabla^2 \tau(\theta, x_1, x_2) = 2 \nabla^2 \int \sin(t(\psi(\lambda, x_1) - \mu)) \sin(t(\psi(\lambda, x_2) - \mu)) \, dG(t)
$$

and $\eta_1(x_1; \theta) = E_{x_2}[\eta(x_1, X; \theta)]$. Then, $\eta(x_1, x_2; \theta)$ is a symmetric kernel and is dominated by an integrable function for all $\theta \in \Theta$, according to (iv). Setting $S_M^2 = [-M, M] \times$
$[-M, M]$, we have $P\left( R^2 - \bigcup_{M=1}^{\infty} S_M^2 \right) = 0$. Since $\eta(x_1, x_2; \theta)$ and $\eta(x_1; \theta)$ are continuous on $(\theta, x_1, x_2) \in \Theta \times S^2_f$, a compact set, and $(\theta, x_1) \in \Theta \times [-M, M]$, a compact set, respectively, both $\eta(-.; \theta)$ and $\eta(.; \theta)$ are equicontinuous in $\theta$ (see Kosmala (1995)). Consequently, $U_{2n}(\theta) \xrightarrow{a.s.} 2\nabla^2 \varphi(\theta)$ uniformly in $\theta$. Further, the limit function $\nabla^2 \varphi(\theta)$ is continuous in $\theta$. Hence, using uniform convergence of $\nabla^2 \varphi_n$ and the continuity of $\nabla^2 \varphi$ with $\theta_n \xrightarrow{a.s.} \theta_0$, it is easy to show that

$$\nabla^2 \varphi_n(\theta_n) \xrightarrow{a.s.} \nabla^2 \varphi(\theta_0).$$

From (A.8), we use (A.11) and (A.14) along with Slutsky’s theorem to conclude that

$$\sqrt{n}(\theta_n - \theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0)),$$

where $\Sigma(\theta_0) = (\nabla^2 \varphi(\theta_0))^{-1}$. □

References


Unpublished Reference Paper


The results in this appendix are related to the properties of the function $\psi : (\mathbb{R}^+ \times \mathbb{R}) \to (\mathbb{R}^+ \times \mathbb{R})$ defined as

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

where $\log(\cdot)$ is the natural logarithm. We first proceed to establish some inequalities which are used to prove Lemma B.2.

**Lemma B.1** For $y > 0$, the following inequalities hold.

1. $y \left(\log(y) - 1\right) + 1 \geq 0$.
2. $y \left[\left(\log(y) - 1\right)^2 + 1\right] - 2 \begin{cases} > 0 & \text{if } y > 1, \\
= 0 & \text{if } y = 1, \\
< 0 & \text{if } y < 1.
\end{cases}$

**Proof:** Let $f'$ and $f''$ be the first and second derivatives of a function $f$.

(1) Let $f_1(y) = y(\log(y) - 1) + 1$. Then $f_1(y) \to 1$ as $y \to 0$, $f_1(y) \to \infty$ as $y \to \infty$, $f_1'(y) = \log(y)$ and $f_1''(y) = 1/y > 0$ for $y > 0$. Hence $f_1(y)$ has the unique minimum at $y = 1$ and $f_1(1) = 0$.

(2) Let $f_2(y) = y \left[\left(\log(y) - 1\right)^2 + 1\right] - 2$. Then $f_2'(y) = \log^2(y) \geq 0$ and $f_2(1) = 0$ so that (2) holds. □

**Lemma B.2** The transformation function $\psi(\cdot, \cdot)$ defined in (B.1) satisfies

1. $\psi(\lambda, x) \geq 0$ for $x \geq 0$ and $\psi(\lambda, x) < 0$ for $x < 0$.
2. $\psi(\lambda, x)$ is convex in $x$ for $\lambda > 1$ and concave in $x$ for $\lambda < 1$.
3. $\psi(\lambda, x)$ is a continuous function of $(\lambda, x)$.
4. Let $\psi^{(k)} = \frac{\partial^k}{\partial \lambda^k} \psi(\lambda, x)$. Then for $k \geq 1$,

\[
\psi^{(k)} = \begin{cases} 
\frac{\{(x + 1)^\lambda \log^k(x + 1) - k\psi^{(k-1)}\}}{\lambda}, & \text{if } \lambda \neq 0, x \geq 0, \\
\log^{k+1}(x + 1)/(k + 1), & \text{if } \lambda = 0, x \geq 0, \\
\frac{-\{(x + 1)^{2-\lambda}(-\log(-x + 1))^k - k\psi^{(k-1)}\}}{(2-\lambda)} & \text{if } \lambda \neq 2, x < 0, \\
\frac{-\log(-x + 1)^{k+1}}{(k + 1)} & \text{if } \lambda = 2, x < 0,
\end{cases}
\]
is continuous in \((\lambda, x)\). Here \(\psi^{(0)} \equiv \psi(\lambda, x)\).

\((5)\) \(\psi(\lambda, x)\) is increasing in both \(x\) and \(\lambda\).

\((6)\) \(\psi(\lambda, x)\) is convex in \(\lambda\) for \(x > 0\) and concave in \(\lambda\) for \(x < 0\).

**Proof:**

(1) For \(x \geq 0\), we have
\[
\begin{cases}
(x + 1)^{\lambda} - 1 \geq 0 & \text{if } \lambda > 0, \\
(x + 1)^{\lambda} - 1 \leq 0 & \text{if } \lambda < 0.
\end{cases}
\]
When \(\lambda = 0\), \(\log(x + 1) \geq 0\) for \(x \geq 0\). Hence \(\psi(\lambda, x) \geq 0\) for all \(\lambda\) whenever \(x \geq 0\).

Similarly, for \(x < 0\), we have
\[
\begin{cases}
-((-x + 1)^{2-\lambda} - 1) > 0 & \text{if } \lambda > 2, \\
-((-x + 1)^{2-\lambda} - 1) < 0 & \text{if } \lambda < 2.
\end{cases}
\]
When \(\lambda = 2\), \(-\log(-x + 1) < 0\) for \(x < 0\). Hence \(\psi(\lambda, x) < 0\) for all \(\lambda\) whenever \(x < 0\).

(2) For \(\lambda > 1\),
\[
\frac{\partial^2 \psi(\lambda, x)}{\partial x^2} = \begin{cases}
(\lambda - 1)(x + 1)^{\lambda-2} > 0 & \text{if } x \geq 0, \\
(\lambda - 1)(-x + 1)^{-\lambda} > 0 & \text{if } x < 0 \text{ and } \lambda \neq 2, \\
1/(-x + 1)^2 > 0 & \text{if } x < 0 \text{ and } \lambda = 2.
\end{cases}
\]

Consequently, \(\frac{\partial^2 \psi(\lambda, x)}{\partial x^2} > 0\) for all \(x\) whenever \(\lambda > 1\). Similarly, for \(\lambda < 1\),
\[
\frac{\partial^2 \psi(\lambda, x)}{\partial x^2} = \begin{cases}
(\lambda - 1)(x + 1)^{\lambda-2} < 0 & \text{if } x \geq 0 \text{ and } \lambda \neq 0, \\
-1/(x + 1)^2 < 0 & \text{if } x \geq 0 \text{ and } \lambda = 0, \\
(\lambda - 1)(-x + 1)^{-\lambda} < 0 & \text{if } x < 0.
\end{cases}
\]
Here, we have \(\frac{\partial^2 \psi(\lambda, x)}{\partial x^2} < 0\) for all \(x\) whenever \(\lambda < 1\). Therefore, \(\psi\) is convex in \(x\) for \(\lambda \geq 1\) and concave in \(x\) for \(\lambda < 1\).

(3) It is clear that \(\psi(\lambda, x)\) is continuous in \(\lambda\) and \(x\) except \(\lambda = 0\) for \(x \geq 0\) and \(\lambda = 2\) for \(x < 0\). However, from the power series (B.2), \(\psi(0, x) = \log(x + 1)\) for \(x \geq 0\) and \(\psi(2, x) = -\log(-x + 1)\) for \(x < 0\). Hence \(\psi(\lambda, x)\) is continuous at \(\lambda = 0\) for \(x \geq 0\) and \(\lambda = 2\) for \(x < 0\).

(4) To establish \((4)\), we proceed by induction. Let \(k = 1\). Then for \(\lambda \neq 0\) and \(x \geq 0\),
\[
\frac{\partial \psi(\lambda, x)}{\partial \lambda} = \frac{1}{\lambda} \left\{ (x + 1)^{\lambda} \log(x + 1) - \psi(\lambda, x) \right\}
\]
so that the recurrence relation \((4)\) holds for \(k = 1\). Next assume that \((4)\) hold for \(k = n\) where \(n > 1\). Then
\[
\frac{\partial^{n+1} \psi(\lambda, x)}{\partial \lambda^{n+1}} = \frac{\partial}{\partial \lambda} \left\{ \frac{1}{\lambda} \left[ (x + 1)^{\lambda} \log^n(x + 1) - n \frac{\partial^n \psi(\lambda, x)}{\partial \lambda^n} \right] \right\}
\]
\[
\frac{1}{\lambda} \left\{ (x+1)^\lambda \log^{n+1}(x+1) - n \frac{\partial^n}{\partial \lambda^n} \psi(\lambda, x) \right. \\
- \left. \frac{1}{\lambda} \left[ (x+1)^\lambda \log^n(x+1) - n \frac{\partial^{n-1}}{\partial \lambda^{n-1}} \psi(\lambda, x) \right] \right\} \\
= \frac{1}{\lambda} \left\{ (x+1)^\lambda \log^{n+1}(x+1) - (n+1) \frac{\partial^n}{\partial \lambda^n} \psi(\lambda, x) \right\}
\]

Hence the recurrence relation holds for all \( k \geq 1 \) and \( \lambda \neq 0 \) when \( x \geq 0 \). Similarly, for \( \lambda \neq 2 \) and \( x < 0 \),

\[
\frac{\partial \psi(\lambda, x)}{\partial \lambda} = \frac{-1}{2-\lambda} \left\{ (-x+1)^{2-\lambda} \left( -\log(x+1) \right) - \psi(\lambda, x) \right\}
\]

Thus (4) holds for \( k = 1 \). Assume that (4) hold for \( k = n \) where \( n > 1 \). Then

\[
\frac{\partial^{n+1} \psi(\lambda, x)}{\partial \lambda^{n+1}} = \frac{\partial}{\partial \lambda} \left\{ \frac{-1}{2-\lambda} \left[ (-x+1)^{2-\lambda} \left( -\log(x+1) \right)^n - n \frac{\partial^{n-1}}{\partial \lambda^{n-1}} \psi(\lambda, x) \right] \right\} \\
= \frac{-1}{2-\lambda} \left\{ (-x+1)^{2-\lambda} \left( -\log(x+1) \right)^{n+1} - n \frac{\partial^n}{\partial \lambda^n} \psi(\lambda, x) \right\} \\
- \frac{-1}{2-\lambda} \left\{ (-x+1)^{2-\lambda} \left( -\log(x+1) \right)^n - n \frac{\partial^{n-1}}{\partial \lambda^{n-1}} \psi(\lambda, x) \right\} \\
= \frac{-1}{2-\lambda} \left\{ (-x+1)^{2-\lambda} \left( -\log(x+1) \right)^{n+1} - (n+1) \frac{\partial^n}{\partial \lambda^n} \psi(\lambda, x) \right\}
\]

Here, we have the recurrence relation for all \( k \geq 1 \) and \( \lambda \neq 2 \) when \( x < 0 \).

Writing \( \psi(\lambda, x) \) as power series, and taking the limit as \( \lambda \to 0 \) when \( x \geq 0 \) and \( \lambda \to 2 \) when \( x < 0 \), we conclude that, for all \( \lambda \),

\[
\psi(\lambda, x) = \begin{cases} 
\sum_{k=1}^{\infty} \lambda^{k-1} \log^k(x+1)/k!, & \text{for } x \geq 0 \\
-\sum_{k=1}^{\infty} (2-\lambda)^{k-1} \log^k(x+1)/k!, & \text{for } x < 0
\end{cases}
\]

(B.2)

The values at \( \lambda = 0 \) and \( \lambda = 2 \) coincide with (2.1). Taking limits of \( \left( \psi(\lambda, x) - \psi(0, x) \right)/\lambda \) for \( x \geq 0 \) and \( \left( \psi(\lambda, x) - \psi(2, x) \right)/(2-\lambda) \) for \( x < 0 \), we find

\[
\left. \frac{\partial \psi(\lambda, x)}{\partial \lambda} \right|_{\lambda=0} = \frac{1}{2} \log^2(x+1), & \text{for } x \geq 0, \\
\left. \frac{\partial \psi(\lambda, x)}{\partial \lambda} \right|_{\lambda=2} = \frac{1}{2} \left( -\log(x+1) \right)^2, & \text{for } x < 0.
\]

From the coefficients in the power series (B.2),

\[
\left. \frac{\partial^k \psi(\lambda, x)}{\partial \lambda^k} \right|_{\lambda=0} = \frac{1}{k+1} \log^{k+1}(x+1), & \text{for } x \geq 0, \\
\left. \frac{\partial^k \psi(\lambda, x)}{\partial \lambda^k} \right|_{\lambda=2} = \frac{1}{k+1} \left( -\log(x+1) \right)^{k+1}, & \text{for } x < 0.
\]

23
When \( \lambda \neq 0 \) and 2,

\[
\frac{\partial \psi(\lambda, x)}{\partial x} = \begin{cases} 
(x + 1)^{\lambda-1} > 0 & \text{for all } x \geq 0, \\
(-x + 1)^{1-\lambda} > 0 & \text{for all } x < 0,
\end{cases}
\]

so that \( \psi \) is increasing in \( x \). When \( \lambda = 0 \), \( \frac{\partial \psi(\lambda, x)}{\partial x} = 1/(x + 1) \geq 0 \) for \( x \geq 0 \). When \( \lambda = 2 \), \( \frac{\partial \psi(\lambda, x)}{\partial x} = 1/(1 - x) > 0 \) for \( x < 0 \). Therefore, \( \psi \) is increasing in \( x \).

For \( x \geq 0 \) and \( \lambda \neq 0 \),

\[
\frac{\partial \psi(\lambda, x)}{\partial \lambda} = \frac{1}{\lambda} \left\{ (x + 1)^{\lambda} \log(x + 1) - \frac{(x + 1)^{\lambda} - 1}{\lambda} \right\} = \frac{1}{\lambda^2} \left\{ (x + 1)^{\lambda} \left[ \log(x + 1)^{\lambda} - 1 \right] + 1 \right\}
\]

(B.3)

Letting \( y = (x + 1)^{\lambda} \) and employing Lemma B.1 part (1), we observe that \( \frac{\partial \psi(\lambda, x)}{\partial \lambda} \geq 0 \) for all \( x \geq 0 \) and \( \lambda \neq 0 \). For \( x \geq 0 \), \( \frac{\partial \psi(\lambda, x)}{\partial \lambda} = \left( \log(x + 1) \right)^2 / 2 \geq 0 \) when \( \lambda = 0 \). Similarly, for \( x < 0 \),

\[
\frac{\partial \psi(\lambda, x)}{\partial \lambda} = \frac{1}{2 - \lambda} \left\{ (-x + 1)^{2-\lambda} \left[ \log(-x + 1)^{2-\lambda} - 1 \right] - 1 \right\}
\]

(B.4)

Letting \( y = (-x + 1)^{2-\lambda} \) and utilizing Lemma B.1 part (1) again, \( \frac{\partial \psi(\lambda, x)}{\partial \lambda} > 0 \) for all \( x < 0 \) and \( \lambda \neq 2 \). For \( x < 0 \), \( \frac{\partial \psi(\lambda, x)}{\partial \lambda} = \left( \log(-x + 1) \right)^2 / 2 > 0 \) when \( \lambda = 2 \). This implies that \( \psi \) is increasing in \( \lambda \). Therefore, \( \psi \) is increasing in both variables.

(6) For fixed \( x \geq 0 \),

\[
\frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} = \frac{1}{\lambda} \left\{ (x + 1)^{\lambda} \log^2(x + 1) - 2 \left[ \frac{1}{\lambda} \left( (x + 1)^{\lambda} \log(x + 1) - \frac{(x + 1)^{\lambda} - 1}{\lambda} \right) \right] \right\} = \frac{1}{\lambda^3} \left\{ (x + 1)^{\lambda} \left[ \log^2(x + 1)^{\lambda} - 2(x + 1)^{\lambda} \log(x + 1)^{\lambda} + 2(x + 1)^{\lambda} - 2 \right] \right\} = \frac{1}{\lambda^3} \left\{ (x + 1)^{\lambda} \left[ \log(x + 1)^{\lambda} - 1 \right]^2 + 1 \right\}
\]

(B.5)

Letting \( y = (x + 1)^{\lambda} \) and employing Lemma B.1, part (2), we conclude that, for \( x \geq 0 \), \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} \geq 0 \) if \( \lambda > 0 \) and \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} \geq 0 \) if \( \lambda < 0 \). When \( \lambda = 0 \), \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} = \log^3(x + 1)/3 \geq 0 \) for \( x \geq 0 \). Hence \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} \geq 0 \) for all \( \lambda \) whenever \( x \geq 0 \). Similarly, for \( x < 0 \),

\[
\frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} = \frac{-1}{2 - \lambda} \left\{ (-x + 1)^{2-\lambda} \log^2(-x + 1) \right\} - \frac{2}{2 - \lambda} \left\{ (-x + 1)^{2-\lambda} \left[ \log(-x + 1)^{2-\lambda} - 1 \right] \right\}
\]

(24)
\[
\begin{align*}
&= \frac{-1}{(2-\lambda)^3} \left\{ (-x+1)^{2-\lambda} \left( \log(-x+1)^{2-\lambda} \right)^2 \\
&\quad -2(-x+1)^{2-\lambda} \log(-x+1)^{2-\lambda} + 2(-x+1)^{2-\lambda} - 2 \right\} \\
&= \frac{-1}{(2-\lambda)^3} \left\{ (-x+1)^{2-\lambda} \left[ \left( \log(-x+1)^{2-\lambda} - 1 \right)^2 + 1 \right] - 2 \right\} \\
\end{align*}
\] (B.6)

Setting \( y = (-x+1)^{2-\lambda} \) and utilizing LEMMA B.1, part (2) again, we conclude that for \( x < 0 \), \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} < 0 \) if \( \lambda > 2 \) and \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} < 0 \) if \( \lambda < 2 \). When \( \lambda = 2 \), \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} = -\log^3(-x+1)/3 < 0 \) for \( x < 0 \). Hence \( \frac{\partial^2 \psi(\lambda, x)}{\partial \lambda^2} < 0 \) for all \( \lambda \) whenever \( x < 0 \). In summary,

\[
\begin{align*}
&= \frac{1}{\lambda^3} \left\{ (x+1)^{2-\lambda} \left[ \left( \log(x+1)^{2-\lambda} - 1 \right)^2 + 1 \right] - 2 \right\} \geq 0 \\
&= \log^3(x+1)/3 \geq 0 \\
&= \frac{-1}{(2-\lambda)^3} \left\{ (-x+1)^{2-\lambda} \left[ \left( \log(-x+1)^{2-\lambda} - 1 \right)^2 + 1 \right] - 2 \right\} < 0 \\
&= -\log^3(-x+1)/3 < 0
\end{align*}
\]

Therefore, \( \psi \) is convex in \( \lambda \) for \( x \geq 0 \) and concave in \( \lambda \) for \( x < 0 \). \( \square \)

**PROOF OF LEMMA A.1**

To establish the strong consistency of \( \hat{\theta}_n \), we introduce the notation \( \omega \) for generic outcome and \( A \) for the set where the convergence (B.7) below holds. The almost sure uniform convergence of \( T_n(\cdot) \) implies

\[
\lim_{n \to \infty} \left\{ \sup_{\theta \in \Theta} |T_n(\theta) - T(\theta)| \right\} = 0, \quad \text{with probability 1.} \tag{B.7}
\]

To obtain a contradiction, we assume that \( \hat{\theta}_n \) does not converge to \( \theta_0 \) almost surely so that there exists a set of outcomes \( B \) where \( \hat{\theta}_n(\omega) \) does not converge to \( \theta_0 \) almost surely and \( P(B) > 0 \). Without loss of generality, we can restrict our attention to the set \( C = A \cap B \) with \( P(C) > 0 \).

Since \( \Theta \) is compact, for each \( \omega \in C \), there exists a subsequence \( \{m\} \subset \{n\} \) and a limit point \( \theta_*(\omega) \) with \( \hat{\theta}_m(\omega) \xrightarrow{a.s.} \theta_*(\omega) \neq \theta_0 \). We now consider difference,

\[
|T_m(\hat{\theta}_m) - T(\theta_*)| \leq |T_m(\hat{\theta}_m) - T(\hat{\theta}_m)| + |T(\hat{\theta}_m) - T(\theta_*)| \leq \sup_{\theta \in \Theta} |T_m(\theta) - T(\theta)| + |T(\hat{\theta}_m) - T(\theta_*)| \tag{B.8}
\]

For \( \omega \in C \), we take the limit as \( m \to \infty \) in (B.8). Then the first term on the right hand side of (B.8) goes to zero by (B.7) and the second term also goes to zero by the continuity of \( T(\cdot) \) on \( C \) and the strong consistency of \( \hat{\theta}_m \). However, by definition of \( \hat{\theta}_m \), \( T_m(\theta_0) \geq T_m(\hat{\theta}_m) \) for each \( \omega \in C \) and taking the limit as \( m \to \infty \), we obtain

\[
T(\theta_0) \geq T(\theta_*) \quad \text{for each } \omega \in C,
\]

which is contradiction to the assumption (c.2) which states that \( \theta_0 = \arg\min T(\theta) \) is unique. Therefore,

\[
\hat{\theta}_n \xrightarrow{a.s.} \theta_0. \quad \square
\]

25