Estimation of a multivariate Box-Cox transformation to elliptical symmetry via the empirical characteristic function

Miguel Nakamura* Francisco J. Pérez Adolfo J. Quiroz*

Centro de Investigación en Matemáticas, Apdo. Postal 402, Guanajuato, Gto. 36000, Mexico

Abstract

Let $\mathbf{X} = (X_1, X_2, \ldots, X_d)^t$ be a random vector of positive entries, such that for some $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)^t$, the vector $\mathbf{X}^{(\lambda)}$ defined by $X_i^{(\lambda)} = (X_i^{\lambda_i} - 1)/\lambda_i$, $i = 1, \ldots, d$ is elliptically symmetric. We describe a procedure based on the multivariate empirical characteristic function for estimating the λ_i's. Asymptotic results regarding consistency of the estimators are given and we evaluate their performance in simulated data. In a one-dimensional setting, comparisons are made with other available transformations to symmetry.

Key words: Elliptically Contoured Distributions, Empirical Characteristic Function, Box-Cox Transformations.

Running head: Transformation to Elliptical Symmetry.

1. Introduction

The family of elliptically contoured distributions is a natural semi-parametric generalization of the multivariate Gaussian distribution. It is well known that many statistical procedures yield superior performance when data supports elliptical symmetry (see Nelson, Cox and Ndjuenga, 89). This is perhaps the main reason for the

*Research partially supported by CONACYT (Mexico) grant number 1858E9219

large amount of literature on the subject of elliptically contoured distributions. For an overview, we refer the reader to the book edited by Fang and Anderson (90). See also Cambanis, Huang and Simons (81), Devlin, Gnanadesikan and Kettenring (76), and Fang, Kotz and Ng (90).

Here we consider the problem of transforming a vector of positive entries $\mathbf{X} = (X_1, X_2, \ldots, X_d)^t$ to elliptical symmetry via a multivariate version of the Box-Cox transformation family. For positive X and real λ let

$$\begin{aligned} X^{(\lambda)} &= (X^\lambda - 1)/\lambda, \text{ for } \lambda \neq 0, \\ &= \log(X), \text{ for } \lambda = 0. \end{aligned} \tag{1.1}$$

For $\mathbf{X}$ as above and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_d)^t \in R^d$, let

$$\mathbf{X}^{(\boldsymbol{\lambda})} = (X_1^{(\lambda_1)}, X_2^{(\lambda_2)}, \ldots, X_d^{(\lambda_d)})^t. \tag{1.2}$$

We will assume that for some $\lambda_o \in R^d$, $\mathbf{X}^{(\lambda_o)}$ has a density $f = f_{\lambda_o}$ (with respect to Lebesgue measure, as usual) which is elliptically symmetric, that is, for some real positive definite symmetric matrix A and vector $\boldsymbol{\mu} \in R^d$, and for some function $g : [0, \infty) \to [0, \infty)$, f is of the form

$$f(x) = g(\|A(x - \boldsymbol{\mu})\|). \tag{1.3}$$

If $E\|\mathbf{X}^{(\lambda_o)}\|^2$ exists, then the matrix A can be taken to be (by redefining g) the square root inverse of the covariance matrix Σ of $\mathbf{X}^{(\lambda_o)}$, and $\boldsymbol{\mu}$ will correspond to the mean of $\mathbf{X}^{(\lambda_o)}$.

The case $d = 1$ has been addressed extensively in the literature. Estimation of the transformation parameter may be achieved by techniques based on the notion of residual skewness or other, more general measures of one-dimensional symmetry (Taylor, 85), minimum distance ideas (Nakamura and Ruppert, 90), or by comparing residual means and medians (Hinkley, 77). In the present paper, we discuss a method which is based on a characterization of elliptical symmetry through characteristic functions, which is defined from the outset for general d.

Maximum likelihood for the transformation parameter when f is the multivariate normal density, is considered in the work of Andrews *et al.* (71), and Velilla (93). The latter work compares maximum likelihood estimation and marginal estimation of each λ_i obtained by treating each coordinate separately as a one-dimensional problem, showing that the efficiency loss in the second method may be substantial when the components of $\mathbf{X}^{(\lambda_o)}$ are correlated. The method which we propose is non-parametric in that f is not assumed to be multivariate normal or even known, and is truly multidimensional in the sense that no marginalization takes place in estimating the vector λ. The method is described as follows.

Given an i.i.d. sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ from the distribution of $\mathbf{X}$ and $\lambda \in R^d$, let

$$\overline{\mathbf{X}}_\lambda = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^{(\lambda)}\ , \text{ and } S_\lambda = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i^{(\lambda)} - \overline{\mathbf{X}}_\lambda)(\mathbf{X}_i^{(\lambda)} - \overline{\mathbf{X}}_\lambda)^t. \tag{1.4}$$

Also let

$$\mathbf{e}_i(\lambda) = S_\lambda^{-1/2}(\mathbf{X}_i^{(\lambda)} - \overline{\mathbf{X}}_\lambda), \tag{1.5}$$

for $i = 1, \ldots, n$ be the *standardized residual vectors* obtained by the transformation being set at λ. We refer to $\mathbf{e}_i(\lambda)$ as the i-th λ-residual. The *empirical characteristic function* for the λ-residuals is the complex valued function $\hat{\varphi}_{n,\lambda}$ on R^d given by

$$\hat{\varphi}_{n,\lambda}(u) = \frac{1}{n}\sum_{i=1}^{n} \exp(i\, u^t \mathbf{e}_i(\lambda)).$$

Our estimator $\hat{\boldsymbol{\lambda}}$ of λ_o is the vector minimizing

$$W_n(\lambda) = \int_{R^d} \Im^2(\hat{\varphi}_{n,\lambda}(u))\, w(u)\, du, \tag{1.6}$$

where $\Im(.)$ denotes imaginary part and $w(u)$ is a weight function to be defined below. In the one-dimensional case, the statistic in (1.6) for any fixed λ corresponds to the Feuerverger and Mureika (77) statistic for testing a hypothesis of symmetry. For general dimension, our statistic is a member of (if w is chosen to be a *radial* function) the large family considered by Ghosh and Ruymgaart (92) for testing the hypothesis of spherical symmetry (see also Csörgõ, 86 and Csörgõ and Heathcote, 87). In the following section we discuss further analytical justification for this choice of estimator. In Section 3, the consistency of this estimator is established, using empirical processes results for U-statistics. Section 4 is non-technical, and contains a few examples and real data applications which illustrate the behavior of $\hat{\boldsymbol{\lambda}}$. Implementation issues are discussed and empirical comparisons with other estimators are made.

2. The empirical characteristic function and transformations to elliptical symmetry

A random vector $\mathbf{X} \in R^d$ with probability density f, is said to possess elliptical symmetry when f can be written as in (1.3). Let us say that $\mathbf{X}$ has *symmetry of marginals* or *marginal symmetry* when for every $u \in R^d$, the random variable $u^t\, \mathbf{X}$ is symmetric (about some center), in the usual univariate sense. It is not difficult to prove that

Proposition 2.1. Elliptical symmetry implies symmetry of marginals. □

Denote by φ and $\varphi_{u^t\mathbf{X}}$ the characteristic functions of $\mathbf{X}$ and $u^t\mathbf{X}$, respectively. Let $\boldsymbol{\mu} = E(\mathbf{X})$. For each coordinate X_j of $\mathbf{X}$ and $k \geq 1$, let $\mu_{j,k} = E(X_j^k)$ when this moment exists. We have

Proposition 2.2.

(a) The random vector $\mathbf{X}$ with density f, has symmetry of marginals if, and only if $\Im(\varphi(u)\exp(-iu^t\boldsymbol{\mu})) = 0$ for every $u \in R^d$.

(b) If, for each $j \leq d$ there exists an $\alpha_j > 0$ such that

$$\sum_{1\leq k<\infty} \frac{|\mu_{j,k}|\alpha^k}{k!} < \infty, \ \forall\, |\alpha| < \alpha_j, \tag{2.1}$$

then $\mathbf{X}$ has symmetry of marginals if, and only if $\Im(\varphi(u)\exp(-iu^t\boldsymbol{\mu})) = 0$ in a neighborhood of the origin.

Proof: The "only if" part of Prop. 2.2(a) is well known, while the "if" one can be obtained by an application of the formula

$$F_{u^t\mathbf{X}}(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \frac{\Im(e^{-izx}\varphi_{u^t\mathbf{X}}(z))}{z}dz,$$

where $F_{u^t\mathbf{X}}$ denotes the cumulative distribution function of $u^t\mathbf{X}$. (See, for instance, Parzen, 92, Section 9.3). The only fact which remains to be shown is the "if" part of 2.2(b). Let $u \in R^d$. (2.1) implies that all moments of $u^t\mathbf{X}$ exist. Then, using the formula for moments in terms of derivatives of the characteristic function at zero, we conclude that all odd moments of $u^t(\mathbf{X}-\boldsymbol{\mu})$ are zero. Now, for a distribution satisfying (2.1), Carleman's condition holds (Shohat and Tamarkin, 70 or Serfling, 80, Theorem 1.13) and therefore the problem of moments has a unique solution. That is, the distribution of $u^t(\mathbf{X}-\boldsymbol{\mu})$ is completely determined by its moments, and one can show by repeated application of Carleman's result, that the distribution of $u^t(\mathbf{X}-\boldsymbol{\mu})$ is symmetric about zero, completing the proof. □

Another elementary fact, more directly related to the transformations we are currently considering, is the following.

Proposition 2.3. With $\mathbf{X}$ and $\mathbf{X}^{(\lambda)}$ as defined in the previous section, if there exists a λ_o in R^d such that $\mathbf{X}^{(\lambda_o)}$ is marginally symmetric, then it is unique.

Proof: It will suffice to consider the univariate case. Suppose that for $\lambda \neq \beta$ both $X^{(\lambda)}$ and $X^{(\beta)}$ are symmetric. To simplify notations, without affecting the argument, assume that $\lambda = 1$. We will also assume $\beta > 0$. The other cases can be treated similarly. We have that X and X^β are symmetric, with non-zero centers of symmetry. Let t and s be the centers of symmetry of X and X^β, respectively, and let $\delta = s - t$. For $r > 0$ we must have, using the symmetry of X^β: $\Pr(X^\beta > s+r) = \Pr(X^\beta < s-r)$. This can be rewritten as

$$\Pr(X > t + ((t + \delta + r)^{1/\beta} - t)) = \Pr(X < t - (t - (t + \delta - r)^{1/\beta}))$$

which, using the symmetry of X, forces the equality $(t+\delta+r)^{1/\beta} - t = t - (t+\delta-r)^{1/\beta}$ for infinitely many values of r, an impossibility. $\square$

Denote by φ_λ the characteristic function of $\mathbf{X}^{(\lambda)}$. Let $w : R^d \to [0, \infty)$ be a weight function such that $w(0) > 0$, $\int_{R^d} w(u)\, du$ is finite, and is continuous at the origin. Let Λ be a compact set in R^d. Assume that for $\lambda_o \in \Lambda$, $\mathbf{X}^{(\lambda_o)}$ has an elliptically symmetric density. For each $\lambda \in R^d$ let $\boldsymbol{\mu}_\lambda$ and Σ_λ denote, the mean and covariance matrix of $\mathbf{X}^{(\lambda)}$, respectively. These are assumed to exist for all $\lambda \in \Lambda$. The three previous propositions have the following key result as a corollary.

Proposition 2.4.

(a) With the hypothesis of the previous paragraph, there exists only one value of λ in Λ, namely λ_o, such that

$$\int_{R^d} \Im^2(\exp(-i\, \Sigma_\lambda^{-1/2} u^t\, \boldsymbol{\mu}_\lambda) \varphi_\lambda(\Sigma_\lambda^{-1/2} u)) w(u)\, du = 0. \tag{2.2}$$

(b) If, in addition to the hypothesis above we assume that the coordinates of $\mathbf{X}^{(\lambda)}$ for $\lambda \in \Lambda$, satisfy the moment condition (2.1) (with constants α_j possibly depending on λ), then there exists only one value of λ in Λ, namely λ_o such that

$$\int_V \Im^2(\exp(-i\, \Sigma_\lambda^{-1/2} u^t\, \boldsymbol{\mu}_\lambda) \varphi_\lambda(\Sigma_\lambda^{-1/2} u)) w(u)\, du = 0, \tag{2.3}$$

where V denotes any neighborhood of the origin. $\square$

In the proof of 2.4(b) one can use the same argument of 2.2(b). Since the integral in (2.2) or (2.3) is being approximated by the integral of the λ-residual empirical characteristic function in (1.6), Prop. 2.4 justifies the proposed methodology. Prop. 2.4 will also be instrumental in Section 3 in establishing consistency of the estimator.

For computational reasons, in the applications we make a restriction to the context of Prop. 2.4 by setting the function w in (1.6) equal to zero outside a neighborhood of the origin. In general, certain care must be taken in the choice of the function w in (1.6). Because of the Riemann-Lebesgue Theorem, $\varphi_\lambda(u)$ will *always* approach zero when $\|u\| \to \infty$, even in the absence of elliptical symmetry. Thus it would not make sense to give much weight to large values of u. On the other hand, if integration were restricted to a very small vicinity of the origin, we could be loosing information present in data. We would also like to choose w in such way that the integral $W_n(\lambda)$ may be computed in closed form (see Section 4). These considerations (and some trial and error) have led us to the choice:

$$w(u) = \chi_B(u), \tag{2.4}$$

where χ denotes indicator function and B is a box $[-A, A]^d$, with the value of A to be specified in Section 4.

It is worth noting that in the case $\lambda_o = \mathbf{0}$ with $\mathbf{X}^{(\lambda_o)}$ having a Gaussian distribution, the moment conditions of Prop. 2.4(b) do not hold: If $\mathbf{X}$ is log-normal and we consider the moments of a coordinate of $\mathbf{X}^{(\lambda)}$ for any $\lambda \neq \mathbf{0}$, then the series in (2.1) diverges for every $\alpha > 0$. Still, in this case, direct calculation shows that the integrand in (2.3) is not zero for $\lambda \neq \mathbf{0}$, so that the thesis of 2.4(b) does hold, which is what is needed to prove consistency of $\hat{\lambda}$. In this case the proof of Theorem 3.3 goes through with some minor modifications. The moment conditions (2.1) do hold when the distribution of $\mathbf{X}^{(\lambda_o)}$ is truncated Gaussian (with elliptical symmetry) and for any $\mathbf{X}$ with support in a box of the form $[\epsilon, M]^d$, for some $0 < \epsilon < M < \infty$. This is not a stringent assumption in the present setting, since we are interested in modeling $\mathbf{X}$ with positive entries; transformation (1.3) would not be viable otherwise.

3. Consistency of $\hat{\lambda}$

Here we establish consistency of $\hat{\lambda}$ by means of empirical processes theory as described in Dudley (84) and Pollard (84), and empirical processes results for U-statistics as given in Nolan and Pollard (87). Let us begin by listing some assumptions needed to prove consistency.

Assumption 3.1. $\hat{\lambda}$ is selected from a compact set $\Lambda \subset R^d$ such that $\lambda_o \in \mathrm{int}(\Lambda)$ (the interior of Λ).

Assumption 3.2. The setting of Prop. 2.4(b) holds, that is, we have an i.i.d. sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of vectors in R^d, with the same distribution of $\mathbf{X}$. $\mathbf{X}$ is a vector of positive elements, and such that the distribution of $\mathbf{X}^{(\lambda_o)}$ is elliptically symmetric. The coordinates of $\mathbf{X}^{(\lambda)}$ for $\lambda \in \Lambda$, satisfy the moment condition (2.1).

Assumption 3.3. For λ in Λ, Σ_λ, as defined before Prop. 2.4, is non-singular.

Write P for the probability law of $\mathbf{X}$ and P_n for its empirical version: If C is a measurable set in R^d and g a measurable function,

$$P_n(C) = \frac{1}{n}\sum_{i=1}^{n} \chi_{(\mathbf{X}_i \in C)} \text{ and } P_n(g) = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i).$$

With Assumptions 3.1 and 3.2 we get what is called, in the empirical processes literature, an *envelope function*, namely, there exists a function F_1 on R^d such that F_1 is in $L^2(P)$ and satisfies $F_1(x) \geq \|x^{(\lambda)}\|$, for all $x \in \mathrm{supp}(f)$ and $\lambda \in \Lambda$. The following two propositions are needed for the proof of consistency.

Proposition 3.1. $\overline{\mathbf{X}}_\lambda$, as defined in (1.4) is uniformly consistent for $\lambda \in \Lambda$:

$$\sup_{\lambda\in\Lambda} \|\overline{\mathbf{X}}_\lambda - \boldsymbol{\mu}_\lambda\| \to 0, \text{ a.s. as } n \to \infty. \tag{3.1}$$

Proof: For $\lambda \in \Lambda$, write T_λ for the vector transformation $x \to x^{(\lambda)}$ and T_{λ_j} for the coordinate transformation $x_j \to x_j^{(\lambda_j)}$. Let π_j denote the projection $\pi_j(x) = x_j$. The j-th coordinate of the average $\overline{\mathbf{X}}_\lambda$ is $P_n(T_{\lambda_j} \circ \pi_j)$. Let $\Lambda_j = \{\lambda_j,\ \lambda \in \Lambda\}$. Without loss of generality we shall assume that Λ_j is an interval. Also let

$$C_{\lambda_j} = \{(x,t) \in R^d \times R : 0 \leq t \leq (T_{\lambda_j} \circ \pi_j)(x) \text{ or } (T_{\lambda_j} \circ \pi_j)(x) \leq t < 0\}, \text{ and}$$

$$\Theta_{(x,t)} = \{\lambda_j : (x,t) \in C_{\lambda_j}\}.$$

Rewrite $\Theta_{(x,t)} = \{\lambda_j : 0 \leq t \leq (x_j^{\lambda_j} - 1)/\lambda_j \text{ or } (x_j^{\lambda_j} - 1)/\lambda_j \leq t < 0\}$, to see that $\Theta_{(x,t)}$ is the union of at most two intervals in Λ_j. It follows that the dual density (Assouad, 83) of the class

$$\mathbf{C} = \{C_{\lambda_j} : \lambda_j \in \Lambda_j\}$$

is finite and, therefore, by Prop. 2.13 of Assouad (83) $\mathbf{C}$ is a Vapnik-Chervonenkis class. We conclude that the class

$$\mathbf{F} = \{T_{\lambda_j} \circ \pi_j\, ; \lambda_j \in \Lambda_j\}$$

is a VC-subgraph class (see Dudley, 84, 87 for this and related notions). Then, applying Lemma 25 and Theorem 24, Chapter 2 of Pollard (84) we obtain (3.1). $\square$

Proposition 3.2. With S_λ as defined in (1.4),

$$\sup_{\lambda\in\Lambda} \|S_\lambda - \Sigma_\lambda\| \to 0 \text{ a.s. as } n \to \infty, \tag{3.2}$$

where $\|.\|$ denotes euclidean norm with the matrices seen as vectors of length d^2.

Proof: Let $S_{\lambda,\infty} = \frac{1}{n-1}\sum_{i=1}^n(\mathbf{X}_i^{(\lambda)} - \boldsymbol{\mu}_\lambda)(\mathbf{X}_i^{(\lambda)} - \boldsymbol{\mu}_\lambda)^t$. By the previous proposition and the Cauchy-Schwarz inequality, we have

$$\sup_{\lambda\in\Lambda}\|S_\lambda - S_{\lambda,\infty}\| \to 0 \text{ a.s. as } n\to\infty.$$

Therefore we can replace S_λ in (3.2) by $S_{\lambda,\infty}$, which will simplify the proof. For a probability law Q on R^d, a collection of measurable functions $\mathbf{G}$ and $p = 1, 2$ let $N_p(\epsilon, \mathbf{G}, Q)$ denote the metric entropy of $\mathbf{G}$ with respect to the $L^p(Q)$ norm, as defined in Pollard (84, Chapter 2). Let $\mathbf{F}$ be the class of functions defined in the previous proof. By the fact that $\mathbf{F}$ is a VC-subgraph class, established above, and Lemma 25 of Pollard (84)

$$N_p(\epsilon, \mathbf{F}, Q) \le a_p(1/\epsilon)^{b_p} \tag{3.3}$$

for positive constants a_p, b_p depending only on the Vapnik-Chervonenkis index of the class $\mathbf{C}$ and on $\int F_1^p\, dQ$. From the moment assumption on the coordinates of $\mathbf{X}^{(\lambda)}$, we obtain, via dominated convergence, that the components of $\boldsymbol{\mu}_\lambda$ and the entries of $\Sigma_\lambda^{-1/2}$ are differentiable functions of $\lambda \in \Lambda$. It follows that the class of functions

$$\mathbf{F}_{\boldsymbol{\mu}} = \{g : g(x) = (T_{\lambda_j} \circ \pi_j)(x) - (\boldsymbol{\mu})_j : j \le d, \lambda \in \Lambda\}$$

also satisfies (3.3), for possibly different constants a_p and b_p. Then, we have (see, for example, Problem 24, Chapter 2, Pollard 84) that the class

$$\mathbf{F}_{\boldsymbol{\mu}} * \mathbf{F}_{\boldsymbol{\mu}} = \{g\,h\; : g, h \in \mathbf{F}_{\boldsymbol{\mu}}\}$$

satisfies

$$N_1(\epsilon, \mathbf{F}_{\boldsymbol{\mu}} * \mathbf{F}_{\boldsymbol{\mu}}, Q) \le N_2(\epsilon, \mathbf{F}_{\boldsymbol{\mu}} * \mathbf{F}_{\boldsymbol{\mu}}, Q) \le a_2^2(1/\epsilon)^{2b_2},$$

for the constants a_2 and b_2 mentioned above. An application of Theorem 24 in Pollard (84) now gives that for each pair i, j,

$$\sup_{\lambda\in\Lambda}|(S_{\lambda,\infty})_{ij} - (\Sigma_\lambda)_{ij}| \to 0, \text{ a.s.}$$

completing the proof. □

Now we can prove consistency of $\hat\lambda$.

Theorem 3.3. If the integrable, nonnegative weight function w in (6) is strictly positive in a neighborhood of the origin and zero outside a compact set, then under Assumptions 3.1 to 3.3, $\hat\lambda \to \lambda_o$, a.s., as $n \to \infty$.

Proof: Let $\mathbf{e}_i(\lambda),\ i = 1, \ldots, n$ be the λ-residual as defined in Section 1. Let

$$\mathbf{e}_i(\lambda)_\infty = \Sigma_\lambda^{-1/2}(\mathbf{X}_i^{(\lambda)} - \boldsymbol{\mu}_\lambda) \text{ and } U_n(\lambda) = \int (\frac{1}{n}\sum_{i=1}^{n} \sin(u^t \mathbf{e}_i(\lambda)_\infty))^2\, w(u)\, du. \tag{3.4}$$

Our proof of consistency relies on proving the following three facts, from which consistency of $\hat{\lambda}$ follows by a standard argument using Prop. 3.2:

$$(\mathrm{I}) : \sup_\Lambda |W_n(\lambda) - U_n(\lambda)| \to 0, \text{ a.s.}$$

$$(\mathrm{II}) : \sup_\Lambda |U_n(\lambda) - EU_n(\lambda)| \to 0, \text{ a.s., and}$$

(III) :$EU_n(\lambda)$ converges, uniformly in λ, to a continuous limit E_λ which is zero only at λ_o.

Note first that $W_n(\lambda)$ can be written as $\int(\frac{1}{n}\sum_{i=1}^{n} \sin(u^t e_i(\lambda)))^2\, w(u)\, du$. Now

$$|\frac{1}{n}\sum_{i=1}^{n} \sin(u^t \mathbf{e}_i(\lambda)_\infty) - \frac{1}{n}\sum_{i=1}^{n} \sin(u^t \mathbf{e}_i(\lambda))| \le$$

$$\frac{1}{n}\sum_{i=1}^{n} |u^t(S_\lambda^{-1/2} - \Sigma_\lambda^{-1/2})(\mathbf{X}_i^{(\lambda)} - \boldsymbol{\mu}_\lambda)| + \frac{1}{n}\sum_{i=1}^{n} |u^t \Sigma_\lambda^{-1/2}(\boldsymbol{\mu}_\lambda - \overline{\mathbf{X}}_\lambda)|. \tag{3.5}$$

The second term in (3.5) goes to zero uniformly in $\lambda \in \Lambda$ and $u \in \{u : w(u) > 0\}$ by Proposition 3.1 and the fact that the entries of $\Sigma_\lambda^{-1/2}$ being continuous functions of λ, are uniformly bounded. For a matrix M let $\|M\| = \sup_{\|x\|=1} \|M\,x\|$. The first term in (3.5) is bounded above by

$$\frac{1}{n}\sum_{i=1}^{n} \|u\| \|S_\lambda^{-1/2} - \Sigma_\lambda^{-1/2}\| \|\mathbf{X}_i^{(\lambda)} - \boldsymbol{\mu}_\lambda)\| \le$$

$$\frac{1}{n}\sum_{i=1}^{n} \|u\| \|S_\lambda^{-1/2} - \Sigma_\lambda^{-1/2}\| (F_1(\mathbf{X}_i) + \sup_\Lambda \|\boldsymbol{\mu}_\lambda\|),$$

and this approaches zero uniformly in λ by Proposition 3.2 and the strong law of large numbers applied to the function F_1. Therefore (I) is proved.

For $(\mathbf{X}_i, \mathbf{X}_j) \in R^d \times R^d$, and $\lambda \in \Lambda$ let

$$g_\lambda(\mathbf{X}_i, \mathbf{X}_j) = \int \sin(u^t \mathbf{e}_i(\lambda)_\infty) \sin(u^t \mathbf{e}_j(\lambda)_\infty) w(u)\, du, \tag{3.6}$$

with $\mathbf{e}_i(\lambda)_\infty$ and $\mathbf{e}_j(\lambda)_\infty$ as given in (3.4). Let

$$U_n'(\lambda) = \frac{1}{n(n-1)} \sum_{i \ne j} g_\lambda(\mathbf{X}_i, \mathbf{X}_j). \tag{3.7}$$

Since $U_n(\lambda)$ and $U_n'(\lambda)$ differ only in the *diagonal* terms, we have

$$\sup_{\Lambda} |U_n(\lambda) - U_n'(\lambda)| = O(\frac{1}{n}). \tag{3.8}$$

The stochastic process $\{U_n'(\lambda), \lambda \in \Lambda\}$ is a U-process, as those considered in Nolan and Pollard (87). It is indexed on the class of symmetric kernels

$$\mathbf{G} = \{g_\lambda : R^d \times R^d \to R; \lambda \in \Lambda\},$$

with g_λ as defined in (3.6). To apply the results of Nolan and Pollard we need to establish the following bounds on metric entropies for the class $\mathbf{G}$: For every $\epsilon > 0$:

$$\text{(i) } \log N_1(\epsilon, G, T_n) = o_p(n),$$

$$\text{(ii) } \log N_1(\epsilon, G, P_n \otimes P) = o_p(n), \text{ and}$$
$$\text{(iii) } N_1(\epsilon, G, P \otimes P) < \infty,$$

where T_n is a data dependent discrete measure, defined in Nolan and Pollard (87). We will only prove (ii). The other two bounds are found similarly. First notice that

$$\int |g_\lambda(\mathbf{X}_i, \mathbf{X}_j) - g_\beta(\mathbf{X}_i, \mathbf{X}_j)|\, dP_n(\mathbf{X}_i)\, dP(\mathbf{X}_j) \leq$$

$$\int\int |\sin(u^t \mathbf{e}_i(\lambda)_\infty) - \sin(u^t \mathbf{e}_i(\boldsymbol{\beta})_\infty)| w(u) du\, dP_n(\mathbf{X}_i) +$$
$$\int\int |\sin(u^t \mathbf{e}_j(\lambda)_\infty) - \sin(u^t\ \mathbf{e}_j(\boldsymbol{\beta})_\infty)| w(u) du\, dP(\mathbf{X}_j). \tag{3.9}$$

Given $\epsilon > 0$, take $M > 0$ such that $P_n(F_1(\mathbf{X}_i) > M)$ and $P(F_1(\mathbf{X}_j) > M)$ are both less than $\epsilon/4$. By the Glivenko-Cantelli Theorem and Tchebychev's inequality, $M = O_p(\epsilon^{-1/2})$. As noted above, the components of $\boldsymbol{\mu}_\lambda$ and $\mathbf{X}^{(\lambda)}$, and the entries of $\Sigma_\lambda^{-1/2}$ are differentiable with respect to the components of λ. It follows that on the set $\{(u, \mathbf{X}_i) :\ w(u) > 0 \text{ and } F_1(\mathbf{X}_i) \leq M\}$, $\sin(u^t \mathbf{e}_i(\lambda)_\infty)$ has partial derivatives, with respect to the components of λ, uniformly bounded by $C_3\, M$, for some positive constant C_3. Make a regular grid on Λ of points at distance $\epsilon/(4\, M\, C_3)$. It follows from (3.9) that this grid is an ϵ-cover: for each $\lambda \in \Lambda$ there is a β in the grid, such that

$$\int |g_\lambda(\mathbf{X}_i, \mathbf{X}_j) - g_\beta(\mathbf{X}_i, \mathbf{X}_j)|\, dP_n(\mathbf{X}_i) dP(\mathbf{X}_j) \leq \epsilon.$$

Since Λ is compact, the grid has $O_p(\epsilon^{-3d/2})$ points, showing that $\log N_1(\epsilon, \mathbf{G}, P_n \otimes P) = O_p(1)$, which is better than what we need for (ii). With (i), (ii) and (iii) Theorem 7 of Nolan and Pollard (87), gives us a uniform (in λ) strong law of large numbers for $U_n'(\lambda)$, proving (II).

To prove (III) observe that $E(U_n(\lambda) - U'_n(\lambda)) = O(\frac{1}{n})$, uniformly in λ and

$$E(U'_n(\lambda)) = Eg_\lambda = \int g_\lambda(\mathbf{X}_i, \mathbf{X}_j) dP(\mathbf{X}_i)\, dP(\mathbf{X}_j), \text{ for all } n,$$

where P is the probability law of $\mathbf{X}$. It is clear that $E(g_\lambda)$ is continuous in λ and, by the strong law of large numbers and Proposition 2.4 it reaches its only minimum, namely zero, at λ_o, finishing the proof. $\square$

4. Behavior of $\hat{\lambda}$ on simulated and real data

Although the asymptotic distribution of $\hat{\lambda}$ is not studied in the present paper, asymptotic normal theory does seem plausible once consistency has been established. We plan to address this topic in future research. One important, immediate application of this theory, would be to provide analytical tools for optimal selection of the constant A. In this section, a completely empirical approach is used to gain some insight into other properties of the estimator.

We consider the weight function as defined by (2.4). Algebraic manipulation shows that the following explicit formula for $W_n(\lambda)$ holds:

$$W_n(\lambda) = \int_{R^d} \Im^2(\widehat{\varphi}_{n,\lambda}(u))\, w(u)\, du =$$

$$\frac{2^{d-1}\,|S_\lambda|^{1/2}}{n^2} \sum_{j=1}^{n}\sum_{k=1}^{n}\left[\prod_{m=1}^{d} \frac{\sin\{(Z_{jm} - Z_{km})A\}}{Z_{jm} - Z_{km}} - \prod_{m=1}^{d} \frac{\sin\{(Z_{jm} + Z_{km})A\}}{Z_{jm} + Z_{km}}\right], \qquad (4.1)$$

where $\mathbf{Z}_j = \mathbf{X}_j^{(\lambda)} - \overline{\mathbf{X}}_\lambda$. If $Z_{jm} - Z_{km} = 0$ or if $Z_{jm} + Z_{km} = 0$, the terms in the products above are interpreted to be equal to A. This formula enables us to easily minimize $W_n(\lambda)$ using numerical methods, without actually having to compute an integral for each trial value of λ.

4.1. Symmetry in one dimension

In this subsection we assume that $X^{(\lambda_o)}$ is a symmetric random variable having mean μ_o and variance σ_o^2. This simple setting provides interesting comparisons with other available methods for estimating transformations to symmetry.

Let $\theta_o^t = (\lambda_o, \mu_o, \sigma_o^2)$. Let us firstly analyze the issue of choice of the constant A via simulation. We consider samples of size 100 under the four configurations $\theta_o^t =$

(0.51, 4, 0.81), (−0.24, −5, 1), (0.95, 4, 1), and (0, 2, 0.25), when the distribution of $X^{(\lambda_o)}$ is standard normal. Table 1 shows results which describe the distribution of $\hat{\lambda}$, based on 1000 trials, letting A take on the values 0.6, 0.8, 1.0, 1.5, and 2.0.

An interesting feature in Table 1 is that minimum variance within the proposed class of estimators is achieved selecting A near 1.0, regardless of the value of θ_o. It is also apparent that values of A which are either too small or too large do not favor estimating the transformation; this fact was pointed out earlier when discussing the weight function. On the other hand, the values obtained for sample skewness and kurtosis suggest asymptotic normality of the estimator.

We conducted a second small Monte Carlo exploratory study aimed at comparing the estimator based on the empirical characteristic function with other estimators of λ defined for one dimensional data. Some of these alternative estimators have appeared previously in the literature, while others are adapted here by minimizing existing tests for symmetry. In describing these alternative estimators, we use the notation $e_k(\lambda)$ for the k-th λ-residual. One family of tests for symmetry which we consider is

$$R_{n,p}(\lambda) = \sum_{j=1}^{[\frac{n}{2}]} \mid 2 \cdot \mathrm{med}\{e_k(\lambda), 1 \leq k \leq n\} - e_{(j)}(\lambda) - e_{(n-j+1)}(\lambda) \mid^p,$$

with $p > 0$ (suggested by David Ruppert in personal communication). The list of estimators included in our study is the following:

1. The estimator due to Berry (87), which selects λ to minimize

$$K_n(\lambda) = \mid \frac{1}{n}\sum_{j=1}^{n} e_j^3(\lambda) \mid + \mid \frac{1}{n}\sum_{j=1}^{n} e_j^4(\lambda) - 3 \mid .$$

2. The minimizer of the test due to Boos (82), which is

$$B_n(\lambda) = n\left\{\frac{\frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{n} \mid e_j(\lambda) + e_k(\lambda) - 2\hat{m} \mid}{\sum_{j<k} \mid e_j(\lambda) - e_k(\lambda) \mid} - 1\right\},$$

where $\hat{m} = \mathrm{med}\{[e_j(\lambda) + e_k(\lambda)]/2, 1 \leq j \leq k \leq n\}$.

3. The estimator based on the empirical characteristic function, setting $A = 0.8$ and minimizing $W_n(\lambda)$ defined by (4.1).

4. The minimizer of $R_{n,1}(\lambda)$.

5. The minimizer of $R_{n,2}(\lambda)$.

6. The minimizer of $R_{n,3}(\lambda)$.

7. The *skewness estimator*, obtained by minimizing

$$T_n(\lambda) = \mid \frac{1}{n}\sum_{j=1}^{n} e_j^3(\lambda) \mid .$$

   (Actually, the value of λ is found which makes $T_n(\lambda) = 0$). This estimator is investigated by Taylor (85). He shows that this skewness estimator has an optimality property for normal errors, and compares it with other estimators of λ as well.

8. A minimum distance type of estimator (Nakamura and Ruppert, 90) obtained by minimizing

$$V_n(\lambda) = \int_{-\infty}^{\infty} \{F_n(x,\lambda) + F_n(-x,\lambda) - 1\}^2 x^2 dx,$$

   where $F_n(x,\lambda)$ is the empirical distribution function of $\{e_j(\lambda), 1 \le j \le n\}$.

9. The maximum likelihood estimator assuming f is standard normal. Although this estimator actually attempts to transform to normality, not general symmetry, it was included in this study for comparative reasons. The estimate of λ minimizes

$$L_n(\lambda) = n \log\left\{\hat{\sigma}^2(\lambda)\right\} + \sum_{j=1}^{n} e_j^2(\lambda) - 2(\lambda - 1)\sum_{j=1}^{n} \log\{X_j\}.$$

We simulated one-dimensional samples with transformation (1.1) using three symmetric densities for f: the standard normal, the t-distribution with 3 degrees of freedom, and a double exponential distribution. Sample sizes considered were $n = 30, 50$ and 80, again generating 1000 Monte Carlo replications in each setting. The values used for λ_o were 0.0, 0.51, and 1.0, with $(\mu_o, \sigma_o^2) = (2,\ 0.5^2)$ for the normal and double exponential settings, and $\theta_o^t = (0,\ 2,\ 0.4^2)$, $(.51,\ 2,\ 0.45^2)$, and $(1,\ 2,\ 0.35^2)$ for the t-distribution. Table 2 summarizes Monte Carlo results for a few representative cases, where estimators within a fixed case have been ranked according to empirical mean squared error. A number of conclusions are noteworthy:

For the normal distribution and large sample sizes, the method based on the empirical characteristic function is amongst the most efficient after (normal theory) maximum likelihood.

In the non-normal symmetric situations, none of the estimators considered is uniformly best across all distributions and sample sizes. Which method occupies the top spot seems to depend strongly on the particular setting as well as the sample size. That is, a winner in one situation may rank last in another. However, the empirical characteristic function estimator never ranks worse than fourth (according to mean squared error). No other estimator maintains a high ranking across all situations.

Some of the transformation estimators that minimize a test for symmetry are comparatively inefficient. Examples are estimators numbered 2 and 4. There could be a connection between these results and the power these statistics would exhibit in testing symmetry.

There is a substantial improvement in the estimators based on $R_{n,p}(\lambda)$ when $p = 1$ is switched to $p = 2$, and then to $p = 3$. This point motivates further investigation of this estimator for other values of p, but we do not address this issue or any of the asymptotic properties of this estimator here.

Our results also suggest that the sample skewness, $T_n(\lambda)$, yields a very competitive estimator in this one-dimensional setting. However, our interest here lies in multivariate statistics and, in that setting, the skewness estimator is not directly applicable.

4.2. Symmetry in two and three dimensions

Let us turn to a bivariate scenario, and exemplify estimation of λ in simulated data sets of size 200, letting $\lambda_o = (0.5,\, 0.1)$ and setting the density f to be one of four elliptically symmetric bivariate distributions, labeled D1-D4. D1 is $\begin{pmatrix} 8 & 2 \\ 2 & 1 \end{pmatrix} (U + \begin{pmatrix} 5 \\ 8 \end{pmatrix})$, where U has a uniform distribution on the unit circle on the plane. D2 is a bivariate normal having covariance $\begin{pmatrix} 1 & .8 \\ .8 & 2 \end{pmatrix}$ and mean $\begin{pmatrix} 5 \\ 5 \end{pmatrix}$. D3 is a bivariate normal having covariance $\begin{pmatrix} 1 & .1 \\ .1 & 2 \end{pmatrix}$ and mean $\begin{pmatrix} 5 \\ 5 \end{pmatrix}$. D4 is $\begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix} (V + \begin{pmatrix} 13 \\ 13 \end{pmatrix})$, where V has the (spherically symmetric) bivariate distribution corresponding to the density characterized by $\|V\| \sim \chi^2_4$ and $\arctan(V_1/V_2) \sim \mathrm{Uniform}(0, 2\pi)$. D4 is intended to represent a distribution with "heavier tails" than the bivariate normal. D2 and D3 differ only in that one has a greater covariance term; for likelihood-based methods, this may induce a difference when compared to marginal symmetrization (see Velilla, 93).

The following display reports results obtained by minimizing (4.1) as a function of the bivariate transformation parameter, where we have set $A = 1$.

distribution	$\hat{\lambda}_1$	$\hat{\lambda}_2$
D1	0.308	-0.150
D2	0.483	0.146
D3	0.293	0.120
D4	0.677	0.128

For each setting, Figures 1a and 1b show scatter plots of original simulated data $\mathbf{X}$, and data after having applied the estimated transformation, $\mathbf{X}^{(\hat{\lambda})}$, respectively. These figures suggest good performance of the empirical characteristic function estimator when elliptical symmetry is achievable by a Box-Cox transformation.

We found that an application on a real-life bivariate data set is also interesting to look at. The set regards industrial pollution in 558 water samples in León, Mexico. Two selected variables, VAR1 and VAR2, are considered here for illustrative purposes; both are concentrations of chemical substances, *i.e.* are positive variables. Figure 2 shows scatterplots and marginal distributions of original data and after an estimated transformation obtained using the empirical characteristic function. It is striking how similar these are in appearance to some of the plots in Figures 1a and 1b. For comparisons, bivariate transformations are here not only estimated by using the empirical characteristic function, but also by maximum likelihood assuming bivariate normality (Andrews *et al.*, 77). In addition, we also estimate both transformation parameters using *marginal* likelihood, that is, by individually transforming VAR1 and VAR2 to normality using normal likelihoods. The three pairs of estimates are very similar in this case.

estimator	$\hat{\lambda}_1$	$\hat{\lambda}_2$
char. function	1.890	0.539
max. likelihood	1.949	0.462
marginal likelihoods	2.076	0.443

A final example regards a three-dimensional subset of Urology data (Andrews and Herzberg, 85, p. 249), depicted in Figure 3. Selected Variables are pH (PH), conductivity (MHO), and urea concentration (UREA), in 79 urine specimens. The three-dimensional implementation of equation (4.1), again setting $A = 1$, yields the estimate $\hat{\lambda} = (-1.45, 0.90, 0.68)$. Figure 3 also shows bivariate projections of the data after using this value in transforming all entries. We note that the marginal skewness displayed in original data no longer appears after transformation. In addition, judging by all two-dimensional scatterplots of transformed points, elliptical symmetry might have been achieved.

References

Andrews, D.F., Gnanadesikan, R. and Warner, J.L. (1971) Transformations of Multivariate Data. *Biometrics* **27**, 825–840.

Andrews, D.F., and Herzberg, A.M. (1985) *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* Springer Verlag, New York.

Assouad, P. (1983) Densité et Dimension. *Ann. Inst. Fourier Grenoble* **33**, 233–282.

Cambanis, S., Huang, S. and Simons, G. (1981) On the Theory of Elliptically Contoured Distributions. *Journal of Multivariate Analysis* **11**, 368–385.

Csörgõ, S. (1986) Testing for Normality in Arbitrary Dimension. *The Annals of Statistics* **14**, 708–723.

Csörgõ, S and Heathcote, C.R. (1987) Testing for Symmetry. *Biometrika* **74**, 177–184.

Devlin, S.J., Gnanadesikan, R. and Kettenring, J.R. (1976) Some Multivariate Applications of Elliptical Distributions, in *Essays in Probability and Statistics* (S. Ideka *et al.*, eds.), 365–395, Shink Tsusho Co., Ltd. Tokyo.

Dudley, R.M. (1984) A Course on Empirical Processes. *Ecole d'Eté de Probabilités de Saint-Flour, XII-1982. Lecture Notes in Math. 1097* 1–142. Springer, New York.

Dudley, R.M. (1987) Universal Donsker Classes and Metric Entropy. *The Annals of Probability* **15**, 1306–1326.

Fang, K.T. and Anderson, T.W. (eds.) (1990) *Statistical Inference in Elliptically Contoured and Related Distributions.* Allerton Press Inc., New York.

Fang, K.T., Kotz, S. and Ng, K.W. (1990) *Symmetric Multivariate and Related Distributions.* Monographs on Statistics and Applied Probability **36** Chapman and Hall, London.

Feuerverger, A. and Mureika, R.A. (1977) The Empirical Characteristic Function and its Applications. *The Annals of Statistics* **5**, 88–97.

Ghosh, S. and Ruymgaart, F. (1992) Applications of Empirical Characteristic Functions in some Multivariate Problems. *The Canadian Journal of Statistics* **20**, 4, 429–440.

Hinkley, D.V. (1977) On Quick Choice of Power Transformation. *Applied Statistics* **26**, 67–68.

Nakamura, M. and Ruppert, D. (1990) Semi-parametric estimation of symmetrizing transformations with application to the shifted power transformation. Unpublished manuscript.

Nelson, C.H., Cox, D.D. and Ndjuenga, J. (1989) Mean Variance Portfolio Choice: A test for Elliptical Symmetry. Dept. of Statistics, Univ. of Illinois. Technical Report 41.

Nolan, D. and Pollard, D. (1987) U-Processes: Rates of Convergence. *The Annals of Statistics* **15**, 780–799.

Parzen, E. (1992) *Modern Probability Theory and its Applications.* Wiley Classics Library, Wiley and Sons, New York.

Pollard, D. (1984) *Convergence of Stochastic Processes.* Springer, New York.

Serfling, R.J. (1980) *Approximation Theorems of Mathematical Statistics.* Wiley and Sons, New York.

Shohat, J.A. and Tamarkin, J.D. (1970) *The Problem of Moments.* Mathematical Surveys. Number 1. American Mathematical Society, Rhode Island.

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

Velilla, S. (1993) A Note on the Multivariate Box-Cox Transformation to Normality. *Statistics and Probability Letters* **17**, 259–263.