On robust and efficient estimation of the center of symmetry

Howard D. Bondell

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, U.S.A

(email: bondell@stat.ncsu.edu)

Abstract

In this paper, a class of estimators of the center of symmetry based on the empirical characteristic function is examined. In the spirit of the Hodges-Lehmann estimator, the resulting procedures are shown to be a function of the pairwise averages. The proposed procedures are also shown to have an equivalent representation as the minimizers of certain distances between two corresponding kernel density estimators. An alternative characterization of the Hodges-Lehmann estimator is established upon the use of a particularly simple choice of kernel.

Keywords: Characteristic function; Efficiency; Hodges-Lehmann estimator; Robustness; Symmetry.

1 Introduction

Many robust and nonparametric approaches to estimation of an unknown point of symmetry, θ, have been proposed in the literature. Among them are the Hodges-Lehmann estimator (Hodges and Lehmann, 1963) and the location M-estimators of Huber (1981).

The use of characteristic function based procedures as a technique for robust estimation in the fully parametric setting is discussed by Heathcote (1977), Feuerverger and McDunnough (1981), and others. The use of empirical characteristic function procedures to test the hypothesis of distributional symmetry appears in Heathcote, Rachev, and Cheng (1995) and Henze, Klar, and Meintanis (2003).

However, the use of a characteristic function approach to estimation under the simple assumption of distributional symmetry has not received attention. In this paper, a class of procedures based on minimizing a weighted norm of the imaginary part of the empirical characteristic function is investigated as an estimator of the center of symmetry. In the spirit of the Hodges-Lehmann estimator, the resulting procedures are shown to be a function of the pairwise averages. The proposed procedures are also shown to be equivalently represented as the minimizers of L^2 distances between two corresponding kernel density estimators. In view of this representation, an alternative characterization of the Hodges-Lehmann estimator is also established upon the choice of a particularly simple kernel.

In the next section, the resulting family of procedures is described and is shown to be based on the Walsh averages. In section 3 the equivalence with an intuitive density estimation problem is shown. The asymptotic distribution is derived via standard theory for U-statistics in section 4. Specific examples, including a generalization of the Hodges-Lehmann estimator, are discussed in section 5, along with a small simulation study.

2 The estimating procedure

Given a sample $(x_1, ..., x_n)$, the empirical characteristic function for each $t \in \mathbb{R}$ is computed as:

$$\hat{f}(t) = n^{-1} \sum_{j=1}^{n} \exp(itx_j) = n^{-1} \sum_{j=1}^{n} \{\cos(tx_j) + i \sin(tx_j)\}, \tag{2.1}$$

where $i \equiv \sqrt{-1}$ is the imaginary unit. If the underlying distribution were symmetric about zero, the characteristic function would be strictly real. This is the idea behind the tests of symmetry of Heathcote et al. (1995) and Henze et al. (2003), along with the family of estimators of this paper. Heuristically, one would like to find the location parameter whose shift minimizes the imaginary part of the resulting characteristic function in some sense. The choice of objective function to be used here is a weighted integral involving the imaginary part of the characteristic function.

Specifically, let $\sigma^2 = \text{Var}\ (X)$. The chosen quantity to minimize is:

$$D(\theta) \equiv \int \left[n^{-1} \sum_{j=1}^{n} \sin \{t(x_j - \theta)/\sigma\} \right]^2 w(t)\ dt, \tag{2.2}$$

where $w(t)$ is the density function corresponding to some distribution symmetric around zero. Using the identity

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

one obtains:

$$D(\theta) = \tfrac{1}{2} \int \textstyle\sum_{j=1}^{n} \sum_{k=1}^{n} \left[\cos \{t(x_j - x_k)/\sigma\} - \cos \{t(x_j + x_k - 2\theta)/\sigma\}\right] w(t)\ dt.$$

Using the symmetry of the distribution W, and the fact that the first term is independent of θ, the minimization problem can be expressed in the following form. The estimator, $\tilde{\theta}$ of the center of symmetry, θ, is given by:

$$\tilde{\theta} \equiv \arg\min_{\theta} \sum_{j=1}^{n} \sum_{k=1}^{n} \left[1 - \hat{f}_w \left\{ \frac{2}{\sigma} \left(\frac{x_j + x_k}{2} - \theta \right) \right\} \right], \tag{2.3}$$

where $\hat{f}_w(\cdot)$ represents the characteristic function of the distribution with density $w(t)$.

To derive the asymptotic distribution in section 4, it is useful to examine the estimating equation. The first order condition associated with (2.3) yields the estimating equation:

$$\sum_{j=1}^{n}\sum_{k=1}^{n}\Psi\left\{\frac{2}{\sigma}\left(\frac{x_j+x_k}{2}-\theta\right)\right\}=0, \tag{2.4}$$

where $\Psi(t)=(\partial/\partial t)\hat{f}_w(t)$, assuming that the derivative exists. Although it is possible that the estimating equation have multiple solutions, the estimator itself is defined as the minimizer of the objective function.

3 Representation via density estimation

The estimator defined as the minimizer of (2.2) derived via the characteristic function has an equivalent representation as the minimizer of the L^2 distance between two kernel density estimates for a corresponding choice of kernel. Specifically, the relationship is given by the following theorem.

THEOREM 1 *Let $g(x)$ be a square-integrable density function symmetric around zero with corresponding characteristic function $\phi(t)$. Suppose that $w(t)=k^{-1}\ \phi^2(t)$ for all t, with $k=\int \phi^2(t)\ dt$.*

Then the objective function based on the imaginary part of the empirical characteristic function using the weight function $w(t)$ can alternatively be expressed as

$$D(\theta)=\tfrac{\pi}{2k}\int\{\tilde{f}_n(x)-\tilde{f}_n(-x)\}^2\ dx,$$

where

$$\tilde{f}_n(x)=n^{-1}\textstyle\sum_{j=1}^{n} g(x-y_j),$$

and $y_j\equiv(x_j-\theta)/\sigma$.

Proof. First note that

$$\left\{n^{-1}\sum_{j=1}^{n}\sin(ty_j)\right\}^2 w(t) = \frac{1}{4k}|n^{-1}\sum_{j=1}^{n}\exp(ity_j)\ \phi(t) - n^{-1}\sum_{j=1}^{n}\exp(-ity_j)\ \phi(t)|^2.$$

Now $n^{-1}\sum_{j=1}^{n}\exp(ity_j)\ \phi(t)$ is the characteristic function of the convolution of the distribution whose density is given by $g(x)$ with the empirical distribution of $y_1, ..., y_n$, whereas $n^{-1}\sum_{j=1}^{n}\exp(-ity_j)\ \phi(t)$ is the characteristic function of the convolution with the empirical distribution of $-y_1, ..., -y_n$. The densities of these two convolutions are given by $\tilde{f}_n(x)$ and $\tilde{f}_n(-x)$, respectively. Now for square-integrable densities $f(x)$ and setting $\hat{f}(t) \equiv \int \exp(itx)f(x)\ dx$, the Plancherel-Parseval Relation, gives

$$\int |\hat{f}(t)|^2\ dt = 2\pi \int |f(x)|^2\ dx.$$

The theorem then follows.

REMARK: Note that $\tilde{f}_n$ is a nonparametric kernel density estimator with kernel $g(x)$ applied to the standardized data, whereas $\tilde{f}_n$ is the same density estimator applied to the standardized data after reflection about the origin. This representation in terms of an L^2 distance between the corresponding density estimates yields an alternative characterization of the estimator.

One instead could have started by defining an estimator as the point which minimizes the distance between the kernel density estimator and its reflected version, yielding the identical procedure for particular choices of kernels and the corresponding weight functions. Two examples of this dual representation are given in section 5, although others are possible.

4 Asymptotic Distribution

Using a first order Taylor expansion of the estimating equation in (2.4), one obtains the standard asymptotic representation:

$$n^{1/2}(\tilde{\theta} - \theta) \approx \; a^{-1} \; n^{1/2} V_n, \tag{4.1}$$

where

$$a \equiv -\frac{\partial}{\partial \theta} \; E \left[\Psi \left\{ \frac{2}{\sigma} \left(\frac{X_1 + X_2}{2} - \theta \right) \right\} \right], \tag{4.2}$$

and

$$V_n \equiv n^{-2} \sum_{j=1}^{n} \sum_{k=1}^{n} \Psi \left\{ \frac{2}{\sigma} \left(\frac{x_j + x_k}{2} - \theta \right) \right\}. \tag{4.3}$$

Under regularity conditions, one may interchange the derivative and expectation to compute the value of a. An example where this does not apply is given in section 5.2, where the direct computation of (4.2) is used. Meanwhile, V_n in (4.3) is a V-statistic of order two, so the distribution of $n^{1/2}V_n$ can be derived via general theory for the corresponding U-statistic (see, for example, Serfling, 1980).

Specifically, for a V-statistic, V_n, based on a symmetric kernel, $h(x_1, x_2)$, with

$$E\{h(X_1, X_2)\} = 0 \text{ and } E\{h(X_1, X_2)\}^2 < \infty,$$

one obtains that V_n is asymptotically normal with mean zero and

$$\text{Var } (V_n) = 4\xi_1/n + O(n^{-2}), \tag{4.4}$$

where $\xi_1 = E\{h(X_1, X_2)h(X_1, X_3)\} = E\{h_1(X)\}^2$ and $h_1(t) \equiv E_X\{h(t, X)\}$.
Hence the asymptotic distribution of the estimator follows directly as

$$n^{1/2}(\tilde{\theta} - \theta) \rightarrow N(0, a^{-2}4\xi_1).$$

Note that in this setting, the kernel is taken to be

$$h(x_1, x_2) \equiv \Psi\left\{\frac{2}{\sigma}\left(\frac{x_1+x_2}{2} - \theta\right)\right\}.$$

To conduct inference, a consistent estimate of ξ_1 is given by

$$n^{-3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} h(x_i, x_j)h(x_i, x_k). \tag{4.5}$$

Alternative estimators of ξ_1 are given by Sen (1960), or a jackknife version by Arvesen (1969).

In practice, σ^2 is typically unknown and must be estimated. If one chooses an auxiliary root-n consistent estimate then the asymptotic distribution is unaffected by the estimation of σ^2. A good choice due to its high degree of robustness would be the Median Absolute Deviation (MAD) as in Huber (1981).

5 Examples

5.1 Gaussian weight function

A particularly useful choice of $w(t)$ is to set, for $c > 0$,

$$w(t) = (c/\pi)^{1/2}\exp(-ct^2), \tag{5.1}$$

i.e. the density of $N(0, 1/2c)$. For this particular choice of weight function, one obtains

$$\begin{aligned} \hat{f}_w(t) &= \exp(-t^2/4c), \\ \phi(t) &= \exp(-ct^2/2), \\ g(t) &= (2\pi c)^{-1/2}\exp(-t^2/2c). \end{aligned} \tag{5.2}$$

Thus, for this choice of weight function on the frequencies, it is equivalent to comparing density estimates using a Gaussian kernel with variance c, which then, as in density estimation, plays the role of the bandwidth. One can choose c based on efficiency considerations, i.e. to achieve a desired asymptotic relative efficiency at a given model

such as the normal as in typical robust procedures involving a tuning constant. Alternatively, one can use bandwidth selection techniques from density estimation to choose an appropriate value. In this paper, the approach based on asymptotic efficiency is further discussed.

The score function involved in the estimating equation (2.4) for this choice of weight function is given by

$$\Psi(t) = (t/2c)\ \exp(-t^2/4c).$$

From this, one can interchange differentiation and integration in (4.2) to obtain

$$a = E\left\{\frac{Z^2 - c}{\sigma c^2}\exp(-Z^2/2c)\right\}, \tag{5.3}$$

where $Z = \sqrt{2}\left(\frac{X_1+X_2}{2} - \theta\right)/\sigma$.

For this choice of weight function, an explicit expression can be derived for the asymptotic variance of the estimator $\tilde{\theta}_1$ under the assumption of normality of the underlying distribution. This allows one to choose the constant c based on a desired efficiency level. Under the assumption that $X \sim N(\theta, \sigma^2)$, to compute the asymptotic distribution, the random variable Z in (5.3) is a standard normal, and by a standard calculation, one then obtains

$$a = \sqrt{\frac{c}{\sigma^2(c+1)^3}}. \tag{5.4}$$

After some straightforward, but tedious, algebra one also obtains,

$$\xi_1 = \frac{2c}{\{(2c+1)(2c+3)\}^{3/2}}. \tag{5.5}$$

Hence using (5.4) and (5.5), along with the representation given by (4.1), the asymptotic variance under normality has the following simple form:

$$n\ \mathrm{Var}\ (\tilde{\theta}_1 - \theta) \approx \left\{\frac{2c+2}{\sqrt{(2c+1)(2c+3)}}\right\}^3\ \sigma^2. \tag{5.6}$$

As expected, as $c \to \infty$ the variance converges to σ^2, as $\tilde{\theta}_1$ tends toward the mean of the pairwise averages, which is just the mean, $\bar{x}$, itself. To achieve an asymptotic efficiency level of $0 < \alpha < 1$, using (5.6), one would choose $c = (2\sqrt{1-\alpha^{2/3}})^{-1} - 1$. For 95% efficiency at the normal distribution, $c \approx 1.727$.

5.2 A generalization of the Hodges-Lehmann estimator

An alternate choice of $w(t)$ is given by

$$w(t) = (\pi c)^{-1}\{1 - \cos(ct)\}/t^2. \tag{5.7}$$

This distribution is sometimes known as Polya's distribution in reference to its use in the proof of Polya's criterion to construct characteristic functions (Durrett, 1996, chapter 2). Although this choice of weight function may seem strange at first, the reason for this choice shall become apparent as both its corresponding kernel and the connection with the Hodges-Lehmann estimator are demonstrated.

For this particular choice, using the fact that $1 - \cos(t) = 2\sin^2(t/2)$, one obtains

$$\begin{aligned} \hat{f}_w(t) &= \left(1 - \frac{|t|}{c}\right)^+, \\ \phi(t) &= \frac{\sin(ct/2)}{ct/2}, \\ g(t) &= c^{-1}\, I\{t \in [-c/2, c/2]\}. \end{aligned} \tag{5.8}$$

From (5.8), it follows that using the Polya weight function on the frequencies is equivalent to comparing density estimates using a Uniform kernel of width c. The fact that this simple choice of kernel for the density estimation view arises from the choice of Polya's distribution for a weight function makes it intuitively appealing. In addition, the form of the characteristic function of this distribution, $\hat{f}_w(t)$, is now shown to be intimately tied to the Hodges-Lehmann estimator.

From (2.3), the estimator, $\tilde{\theta}_2$, derived from using this choice of weight function or, equivalently, the uniform kernel density estimator, is then given as the solution of the

following minimization problem:

$$\tilde{\theta}_2 \equiv \arg\min_{\theta} \sum_{j=1}^{n} \sum_{k=1}^{n} \left\{ 1 - \left(1 - \frac{2\ |(x_j + x_k)/2 - \theta|}{\sigma c} \right)^{+} \right\}. \tag{5.9}$$

As the median of the pairwise averages, a variant of the Hodges-Lehmann estimator, θ_{HL} can be defined as

$$\tilde{\theta}_{HL} \equiv \arg\min_{\theta} \sum_{j=1}^{n} \sum_{k=1}^{n} |(x_j + x_k)/2 - \theta|. \tag{5.10}$$

Clearly, for sufficiently large c, the two estimators, $\tilde{\theta}_2$ and $\tilde{\theta}_{HL}$ will coincide exactly. Hence an alternative view of the Hodges-Lehmann estimator is taken as the limiting value of the family of estimators indexed by the constant c.

The estimator $\tilde{\theta}_2$ defined by (5.9) is the solution to the estimating equation as in (2.4) with

$$\Psi(t) = \ \text{sign}(t)\ I\{|t| \le c\} = I\{0 \le t \le c\} - I\{-c \le t \le 0\}. \tag{5.11}$$

Assuming the symmetry of the underlying distribution, direct use of (4.1) - (4.4) after some calculations then yields the asymptotic variance of $\tilde{\theta}_2$ as

$$n\ \text{Var}\ (\tilde{\theta}_2 - \theta) \approx \frac{\int \{F(x + c\sigma) + F(x - c\sigma) - 2F(x)\}^2\ \ f(x)\ dx}{4\ \left[\int \{f(x) - f(x + c\sigma)\}\ \ f(x)\ dx\right]^2}, \tag{5.12}$$

where f denotes the density function of the underlying random variable X, and F is the corresponding distribution function. Note that as expected, as $c \to \infty$, (5.12) converges to the variance of the Hodges-Lehmann estimator.

To be used in practice, estimation of the numerator of (5.12) can be done by substitution of the empirical distribution function for the distribution function, or can be done via the empirical moments as discussed at the end of section 4. To estimate the denominator, a kernel density estimate can be plugged in for the density. Alternatively, one could use a normal approximation or a bootstrap approach to the distribution of

the estimating function itself and construct confidence intervals and tests of hypothesis based on the distribution of the estimating function as discussed in Jiang and Kalbfleisch (2004).

5.3 Simulation study

A small simulation study was carried out to access the efficiency and robustness of the estimator, $\tilde{\theta}_1$, based on the Gaussian weight function, along with the accuracy of the asymptotic approximation. The Hodges-Lehmann estimator, which is equivalent to $\tilde{\theta}_2$ for large c, is compared along with the mean, median, and Huber location estimator. For the Huber estimator, as well as the estimator $\tilde{\theta}_1$, the tuning parameters are chosen to yield 95% asymptotic efficiency at the normal distribution, and the median absolute deviation is used as an auxiliary estimate of scale.

Five different choices of symmetric distributions are simulated. The distributions considered are:

1. normal
2. t-distribution with 3 degrees of freedom
3. two-sided exponential
4. a 90/10 mixture of two normals with identical means, but the smaller component having a standard deviation three times the larger component
5. an 80/20 mixture as in the previous case.

For each setup, 5000 samples of size 20 are generated and the variances are shown in Table 1. Note that the asymptotic variance formula would yield a value of $1/.95 = 1.053$ for $\tilde{\theta}_1$ under the normal distribution, which agrees very closely with the simulated value of 1.06 even for samples of size 20. Hence the asymptotic approximation appears to be

reasonable. Overall, the estimator $\tilde{\theta}_1$ performs similarly to both the Huber estimator and the Hodges-Lehmann estimator under the various symmetric distributions in terms of efficiency.

(**** TABLE 1 GOES HERE ****)

Additionally, contaminated asymmetric distributions are simulated by mixtures of normals with different means. In this situation a contamination bias is created in each of the estimators. Table 2 shows the mean squared error for each of the estimators under 5% contamination. Each setup contaminates a standard normal with a 5% proportion of a shifted normal for various shifts. Table 3 shows the corresponding results for 10% contamination. In both cases, the newly proposed estimator has smaller mean squared error than the other estimators, particularly for more extreme contamination. Note that $\tilde{\theta}_1$ has a redescending influence function so that as the contamination gets further from the true mean, the influence returns to zero. The other estimators in the study do not have this redescending property, but one can instead use alternative M-estimators that do, but their performance is not investigated in this study.

(**** TABLES 2 AND 3 GO HERE ****)

6 Discussion

This paper has proposed a new class of estimation procedures for estimating the center of a symmetric distribution. This class of estimators has been shown to be equivalently represented as a solution to an optimization problem based on either the

characteristic function or a seemingly unrelated kernel density estimate. The proposed estimators exhibit performance that compete with the currently implemented procedures, while their intuitive construction and dual interpretation makes them a interesting alternative.

As with the typical location estimators, the estimation procedures proposed in this paper have a straightforward extension to the (multiple) regression setting under the assumption of symmetry about zero for the distribution of the residuals. Setting $\theta = \boldsymbol{\beta}^T \mathbf{x}$, one can then minimize the objective function in terms of the regression coefficients. The use of these estimators in regression deserve further investigation.

References

Arvesen, J. N. (1969), Jackknifing U-statistics, *Ann. Math. Statist.* **40**, 2076-2100.

Durrett, R. (1996), *Probability: Theory and Examples*, Belmont, CA: Duxbury Press.

Feuerverger, A. and McDunnough, P. (1981), On some fourier methods for inference, *J. Am. Statist. Assoc.* **76**, 379-387.

Heathcote, C. R. (1977), The integrated squared error estimation of parameters, *Biometrika* **64**, 255-264.

Heathcote, C. R., Rachev, S. T. and Cheng, B. (1995), Testing multivariate symmetry, *J. Multivariate Anal.* **54**, 91-112.

Henze, N., Klar, B. and Meintanis, S. G. (2003), Invariant tests for symmetry about an unspecified point based on the empirical characteristic function, *J. Multivariate Anal.* **87**, 275-297.

Hodges, J. L. and Lehmann, E. L. (1963), Estimates of location based on rank tests, *Ann. Math. Statist.* **34**, 598-611.

Huber, P. J. (1981). *Robust Statistics*, New York: Wiley.

Jiang, W. and Kalbfleisch, J. (2004), Resampling methods for estimating functions with U-statistic structure, *University of Michigan Department of Biostatistics Working Paper*

Series, Working Paper 33.

Sen, P. K. (1960), On some convergence properties of U-statistics, *Calcutta Statist. Assoc. Bull.* **10**, 1-18.

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

Table 1: *Variances under various symmetric distributions for samples of size 20. The table entries denote n times the variance.*

	N (0,1)	t (3)	DE	.9 N (0,1) + .1 N (0,9)	.8 N (0,1) + .2 N (0,9)
$\tilde{\theta}_1$	1.06	1.59	1.47	1.29	1.64
$\bar{x}$	1.00	2.78	2.00	1.82	2.63
Median	1.45	1.83	1.37	1.67	1.94
HL	1.05	1.63	1.48	1.32	1.68
Huber	1.06	1.61	1.49	1.31	1.64

Table 2: *Mean squared error under various contaminating distributions for samples of size 20 with 5% contamination. The table entries denote n times the mean squared error.*

	N (1,1)	N (2,1)	N (3,1)	N (4,1)	N (5,1)
$\tilde{\theta}_1$	1.10	1.18	1.23	1.22	1.19
$\bar{x}$	1.04	1.17	1.41	1.74	2.18
Median	1.47	1.55	1.57	1.57	1.57
HL	1.08	1.17	1.23	1.26	1.27
Huber	1.09	1.17	1.22	1.24	1.24

Table 3: *Mean squared error under various contaminating distributions for samples of size 20 with 10% contamination. The table entries denote n times the mean squared error.*

	N (1,1)	N (2,1)	N (3,1)	N (4,1)	N (5,1)
$\tilde{\theta}_1$	1.22	1.54	1.72	1.67	1.50
$\bar{x}$	1.17	1.74	2.71	4.08	5.85
Median	1.61	1.86	1.94	1.94	1.94
HL	1.21	1.56	1.84	1.96	1.99
Huber	1.21	1.54	1.77	1.84	1.85