The Annals of Statistics
1996, Vol. 24, No. 4, 1457–1473

ORTHOGONALIZATION OF MULTIVARIATE LOCATION ESTIMATORS: THE ORTHOMEDIAN

BY RUDOLF GRÜBEL

Universität Hannover

The coordinatewise median of a multivariate data set is a highly robust location estimator, but it depends on the choice of coordinates. A popular alternative which avoids this drawback is the spatial median, defined as the value that minimizes the sum of distances to the individual data points. In this paper we introduce and discuss another orthogonal equivariant version of the multivariate median, obtained by averaging the coordinatewise median over all orthogonal transformations. We investigate the asymptotic behavior of this estimator and compare it to the spatial median.

1. Introduction. In this paper we introduce and discuss a new multivariate location estimator which generalizes the familiar one-dimensional median; see Small (1990) for a recent review, interesting historical material and a wealth of references on this problem.

We will think of the median and its multivariate variants as quantities associated with distributions on $\mathbb{R}$ or $\mathbb{R}^d$, $d > 1$. By the median of a data set $x_1, \dots, x_n$ we mean the median associated with the distribution that assigns mass $1/n$ to each of the values $x_1, \dots, x_n$. To fix the notation we define the median associated with a one-dimensional distribution P with distribution function F by

$$\mathrm{Med}(P) := \tfrac{1}{2}\bigl(\sup\{x \in \mathbb{R}: F(x) < \tfrac{1}{2}\} + \inf\{x \in \mathbb{R}: F(x) > \tfrac{1}{2}\}\bigr).$$

Multivariate versions of this concept differ with respect to their equivariance properties. For a given class $\mathscr{D}$ of transformations of $\mathbb{R}^d$ we call T $\mathscr{D}$-equivariant if, for all distributions P,

$$T(P^D) = D(T(P)) \quad \text{for all } D \in \mathscr{D}.$$

Here P^D denotes the image of P under the transformation D, that is, $P^D(A) = P(D^{-1}(A))$. Shift equivariance, orthogonal equivariance and scale equivariance arise if $\mathscr{D}$ is specialized to the respective class of transformations.

Received February 1993; revised August 1995.

AMS 1991 *subject classifications*. 62H12, 62F35.

Key words and phrases. Multivariate median, orthogonal equivariance, robust estimation, asymptotic normality.

Perhaps the easiest multivariate version of the median is the *coordinatewise median* defined by

$$\mathrm{CoMed}(P) := \begin{pmatrix} \mathrm{Med}(P^{(1)}) \\ \vdots \\ \mathrm{Med}(P^{(d)}) \end{pmatrix},$$

where $P^{(l)}$ denotes the lth marginal of P. The coordinatewise median is shift and scale equivariant, but it is not orthogonal equivariant. Its dependence on the choice of coordinates is illustrated in Figure 1 (this figure will be used for illustration purposes throughout the paper): part (a) displays nine data points in a fixed coordinate system, part (d) shows the respective coordinatewise medians if this system is turned by an angle of magnitude $\theta = \pi i/180$, $i = 0, \dots, 179$. Usage of this estimator and recognition of its dependence on coordinates date back to the beginning of this century; see Small (1990) for historical details.

The coordinatewise median is obtained simply by applying the one-dimensional median separately to each of the marginals. Other multivariate versions of the median can be found via characterizing properties of the one-dimensional median that do not make use of special aspects of the real line such as its ordering. One such property is the fact that the median minimizes the expected absolute distance. This leads to the L^1-*median*,

$$\mathrm{L1Med}(P) := \arg\min_{\theta} \int \|x - \theta\| P(dx),$$

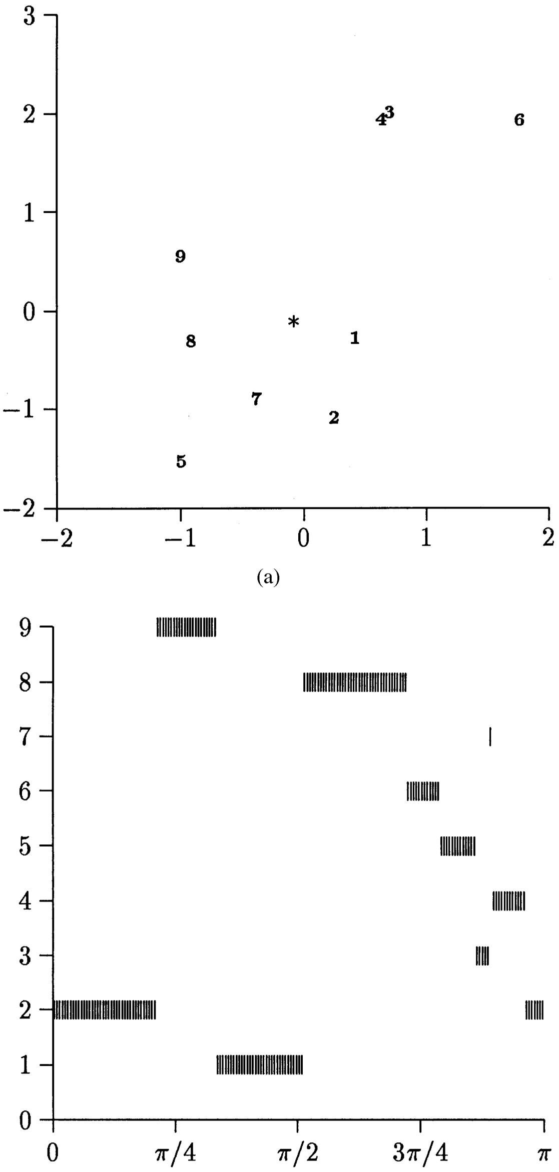
where $\|\cdot\|$ denotes Euclidean distance. As Euclidean distance does not change under orthogonal transformations, the resulting estimator is orthogonal equivariant. It is also shift equivariant, but not scale equivariant; see Brown (1983) and the references given there for uniqueness and other properties of the L^1-median. In the literature this estimator is often called the *spatial median*.

The starting point of the present paper is the observation that an orthogonal equivariant multivariate median can be constructed by averaging the coordinatewise median over all orthogonal transformations, this being possible because the group of orthogonal transformations is compact. To be precise, let $\mathscr{O}(d)$ be the group of orthogonal $d \times d$ matrices. We define the *orthomedian* associated with the d-dimensional distribution P by

$$\mathrm{OrMed}(P) := \int_{\mathscr{O}(d)} A' \, \mathrm{CoMed}(P^A) \, dA. \tag{1}$$

Here A' denotes the transpose of A and $\int \cdots dA$ refers to the (unique) Haar measure on $\mathscr{O}(d)$ with total mass 1. We can interpret this as the expected coordinatewise median in a randomly chosen coordinate system.

The orthomedian is orthogonal equivariant "by construction": from the coordinatewise median it inherits shift equivariance, but not scale equivariance. Roughly, the orthomedian shares the equivariance properties of the L^1-median. Also, both have the same, optimal, finite sample breakdown point.



(a)

(b)

FIG. 1. *A numerical example.*

For equivariance, breakdown points and general background on robust estimation of multivariate location, see Hampel, Ronchetti, Rousseeuw and Stahel (1986) and Lopuhaä and Rousseeuw (1991).

In the following sections we investigate the properties of the orthomedian and compare it to the spatial median. In the next section we first show that

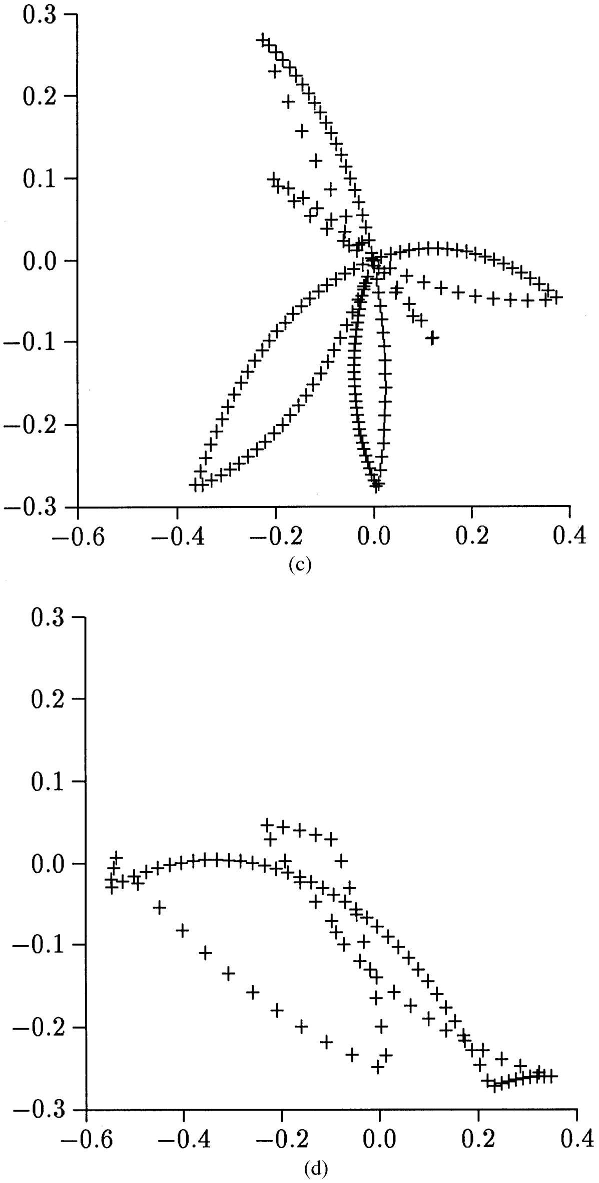


FIG. 1. *continued.*

the integral over $\mathscr{O}(d)$ in the above definition can be reduced to an integral over the boundary of the unit sphere in $\mathbb{R}^d$. Our main result in that section gives conditions for asymptotic normality of the orthomedian and obtains the limiting covariance matrix. In Section 3 this is applied to radially symmetric distributions. Interestingly, for this class of distributions the asymptotic variances of orthomedian and L^1-median coincide. Section 4 deals with a class of elliptic distributions. Computational issues and local robustness are considered in Section 5, and the final section collects some concluding remarks.

2. Asymptotic behavior of the estimator. Let $S(d) := \{x \in \mathbb{R}^d\colon \|x\| = 1\}$ be the boundary of the unit sphere in $\mathbb{R}^d$. For P a distribution on $\mathbb{R}^d$ and $a \in S(d)$, let P^a be the distribution on $\mathbb{R}$ defined by

$$P^a((-\infty, z]) := P(\{x \in \mathbb{R}^d\colon a'x \le z\}).$$

This is the image of P under the mapping $x \to a'x$ leading from $\mathbb{R}^d$ to $\mathbb{R}$. Let $a_1, \dots, a_d$ be the columns of A', $A \in \mathscr{O}(d)$. Then

$$\mathrm{CoMed}(P^A) = \begin{pmatrix} \mathrm{Med}(P^{a_1}) \\ \vdots \\ \mathrm{Med}(P^{a_d}) \end{pmatrix},$$

which implies

$$\mathrm{OrMed}(P) = \sum_{k=1}^{d} \int_{\mathscr{O}(d)} \mathrm{Med}(P^{a_k}) a_k \, dA.$$

If A is uniformly distributed on $\mathscr{O}(d)$, then $a_1, \dots, a_d$ are uniformly distributed on $S(d)$. Hence, writing $\int_{S(d)} \cdots da$ for integration with respect to the uniform distribution on $S(d)$, we obtain

$$\mathrm{OrMed}(P) = d \int_{S(d)} \mathrm{Med}(P^a) a \, da. \tag{2}$$

This formula will also be of interest in connection with computational issues to be discussed later. Figure 1(c) shows the vectors $\mathrm{Med}(P^a)a$ for the data in part (a) and $a = a(\theta) = (\cos(\theta), \sin(\theta))'$, $\theta = \pi i/180$ with $i = 0, \dots, 179$. The points in part (d) arise by adding two such vectors with angles θ and η, where $\eta = \theta + \pi/2$ modulo π.

Our main result in this section deals with the asymptotic behavior as $n \to \infty$ of the orthomedian if the data are realizations of n independent and identically distributed random vectors $X_1, \dots, X_n$. As explained in the Introduction, the orthomedian associated with $X_1, \dots, X_n$ is obtained by applying the definition to the empirical distribution associated with these values. We write $\mathrm{OrMed}(X_1, \dots, X_n)$ for the resulting random vector and $\to_{\mathrm{distr}}$ for convergence in distribution. Let $m_a := \mathrm{Med}(P^a)$ and define $\phi\colon S(d) \times S(d) \to \mathbb{R}$ by

$$\phi(a, b) := P(\{x \in \mathbb{R}^d\colon a'x \le m_a, b'x \le m_b\}).$$

THEOREM 1. *Assume that the distribution P is such that, for all $a \in S(d)$, P^a has a density f_a with the property that $(a, x) \to f_a(x)$ is continuous on*

$\{(a, m_a): a \in S(d)\}$. *Assume further that* $f_a(m_a) > 0$ *for all* $a \in S(d)$ *and let* $c(a) := 1/f_a(m_a)$. *Let* $X_1, X_2, \ldots$ *be a sequence of independent random vectors with distribution* P. *Then*

$$\sqrt{n}\,(\mathrm{OrMed}(X_1, \ldots, X_n) - \mathrm{OrMed}(P)) \to_{\mathrm{distr}} Z,$$

where Z *has a normal distribution with mean vector* 0 *and variance matrix*

$$EZZ' = d^2 \int_{S(d)} \int_{S(d)} \left(\phi(a, b) - \tfrac{1}{4}\right) c(a) c(b) ab' \, da \, db.$$

The proof of the theorem combines results from empirical process theory and a weak version of the delta method. We start with a sufficiently rich central limit theorem (CLT), then switch to an almost sure representation and investigate the behavior of individual "paths." This identifies an almost sure limit for the representation: the distribution of this limit is the desired limit distribution of the original quantities. Our basic reference for empirical process theory is Pollard (1984) to which book we refer for terminology not explained below. For the delta method, see Gill (1989) and the references given there; see also Grübel (1988), where an application to robust scale estimation is given.

For the proof we need the following three lemmas. A lemma similar to the first of these appears in many papers; in the form given here it follows from Lemma 2 in Gill (1989). From the literature I am aware of it seems that its basic idea should be attributed to Vervaat (1972).

LEMMA 1. *Let* $F, F_1, F_2, \ldots$ *be distribution functions on the real line. Assume that the derivative of* F *exists and is positive at* $\mathrm{Med}(F)$. *Then*

$$\sqrt{n}\,(F_n - F) \to g \quad \text{uniformly on } \mathbb{R} \text{ with } g: \mathbb{R} \to \mathbb{R} \text{ continuous}$$

implies

$$\sqrt{n}\,(\mathrm{Med}(F_n) - \mathrm{Med}(F)) \to -g(\mathrm{Med}(F))/F'(\mathrm{Med}(F)).$$

LEMMA 2. *Let* F *be a distribution function with the property that the derivative of* F *exists and is greater than or equal to* ε *on* $[\mathrm{Med}(F) - \delta, \mathrm{Med}(F) + \delta]$, $\varepsilon > 0$ *and* $\delta > 0$. *Then the following implication holds for all distribution functions* G:

$$\|F - G\|_\infty \le \delta\varepsilon/2 \quad \Rightarrow \quad |\mathrm{Med}(G) - \mathrm{Med}(F)| \le 2\|F - G\|_\infty/\varepsilon.$$

PROOF. We may assume $\mathrm{Med}(F) = 0$. Let $x = \|F - G\|_\infty/\varepsilon$. Then $0 \le x \le \delta$ so that $F(x) \ge 1/2 + \varepsilon x$, which gives

$$G(x) \ge 1/2 + \varepsilon x - \|F - G\|_\infty = 1/2;$$

hence $\mathrm{Med}(G) \le x$. For the lower bound let $x = -2\|F - G\|_\infty/\varepsilon$. From $-\delta \le x \le 0$ we now obtain $F(x) \le 1/2 - \varepsilon x$, which yields

$$G(x) \le 1/2 - \varepsilon x + \|F - G\|_\infty \le 1/2 - \|F - G\|_\infty.$$

If $\|F - G\|_\infty = 0$, then $\mathrm{Med}(G) = 0$. If $\|F - G\|_\infty > 0$, then $G(y) < 1/2$ for all $y \le x$; hence $\mathrm{Med}(G) \ge x$. □

LEMMA 3. *With P, f_a and m_a as in Theorem* 1, *the following holds*:

$$\exists \varepsilon > 0,\ \delta > 0\ \forall a \in S(d),\ x \in \mathbb{R}\colon |x - m_a| \le \delta \quad \Rightarrow \quad f_a(x) \ge \varepsilon.$$

PROOF. Assume that we have $f_{a_n}(x_n) \to 0$ and $x_n - m_{a_n} \to 0$ for some sequences $\{a_n\}$ in $S(d)$ and $\{x_n\}$ in $\mathbb{R}^d$. Since $S(d)$ is compact, we may assume that $a_n \to a_0$ for some $a_0 \in S(d)$. As P^{a_n} converges weakly to P^{a_0} and as $x \to P^{a_0}((-\infty, x])$ is strictly increasing in a neighbourhood of m_{a_0}, we have $m_{a_n} \to m_{a_0}$, that is, $(a_n, m_{a_n}) \to (a_0, m_{a_0})$. By the continuity assumption on $(a, x) \to f_a(x)$ this implies $f_{a_0}(m_{a_0}) = 0$, in contradiction to a previous assumption. □

PROOF OF THEOREM 1. We define a class of real functions on $\mathbb{R}^d$ by

$$\mathscr{F} := \{f(\cdot|z, a)\colon z \in \mathbb{R},\ a \in S(d)\}, \qquad f(x|z, a) := 1_{(-\infty, z]}(a'x).$$

This is a subset of the space $\mathscr{L}^2(P)$ of P-square integrable functions on $\mathbb{R}^d$. $\mathscr{F}$ inherits from $\mathscr{L}^2(P)$ the seminorm

$$\rho_P(f, g) = \bigl(P(f - g)^2\bigr)^{1/2},$$

where, following the convention in empirical process theory, we have written a measure applied to a function for the integral of the function with respect to the measure. Let $\mathscr{X}$ be the space of bounded real-valued functions on $\mathscr{F}$, endowed with the supremum norm and the σ-field generated by the open balls in this norm, and let $C_P(\mathscr{F})$ be the subset of ρ_P-continuous elements of $\mathscr{X}$. From our assumptions on P it follows that none of the one-dimensional projections P^a has an atom. Using this, it is easy to see that $(z, a) \to \Psi(f(\cdot|z, a))$ is continuous if $\Psi \in C_P(\mathscr{F})$.

The random vectors $X_1, X_2, \dots$ are defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$. The empirical distribution P_n is the random probability measure on $\mathbb{R}^d$ that assigns mass $1/n$ to each of the points $X_1, \dots, X_n$. We regard P_n as a mapping from Ω into $\mathscr{X}$. Formally,

$$(P_n(\omega))(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i(\omega)).$$

Similarly, P defines a (deterministic) function of $\mathscr{F}$.

The half-spaces $\{x \in \mathbb{R}^d\colon a'x \le z\}$, $a \in S(d)$, $z \in \mathbb{R}$, form a Vapnik–Cervonenkis class; $\mathscr{F}$ is the class of the associated indicator functions. Hence our setup satisfies the assumptions of the empirical CLT [Pollard (1984), page 157] which means that

$$Z_n := \sqrt{n}\,(P_n - P) \to_{\mathrm{distr}} Z,$$

where Z is a Gaussian process, indexed by $\mathscr{F}$, with mean function 0 and covariance function

$$\operatorname{cov}(Z(f), Z(g)) = P(fg) - P(f)P(g). \tag{3}$$

Moreover, all paths of Z are bounded and continuous with respect to ρ_P.

We now invoke the Skorohod–Dudley representation theorem [Pollard (1984), page 71]: there exist $\tilde{P}_n, \tilde{Z}$ on a suitable probability space $(\tilde{\Omega}, \tilde{\mathscr{A}}, \tilde{\mathbb{P}})$ equal in distribution to P_n and Z, respectively, such that $\tilde{Z}_n = \sqrt{n}(\tilde{P}_n - P)$ converges to $\tilde{Z}$ almost surely, that is,

$$\sup_{f \in \mathscr{F}} |\tilde{Z}_n(f) - \tilde{Z}(f)| \to 0 \qquad \tilde{\mathbb{P}}\text{-almost surely.} \tag{4}$$

[We skip some details from the theory of weak convergence of probability measures on nonseparable spaces such as $\mathscr{X}$; see Chapter 4 in Pollard (1984).] Fix an $\tilde{\omega}$ from the probability-1 set in (4) and fix some $a \in S(d)$. From now on, drop $\tilde{\omega}$ from the notation. The function $z \to \tilde{Z}(f(\cdot|z, a))$ is continuous and, by (4), it is the uniform limit of the functions

$$z \to \sqrt{n}\left(\tilde{P}_n(f(\cdot|z,a)) - P(f(\cdot|z,a))\right).$$

Note that the big brackets contain the difference of the distribution functions F_n^a and F^a associated with $\tilde{P}_n^a$ and P^a, respectively. We are therefore in a position to apply Vervaat's lemma, which gives

$$\sqrt{n}\left(\operatorname{Med}\left(\tilde{P}_n^a\right) - \operatorname{Med}(P^a)\right) \to -c(a)\tilde{Z}\left(f(\cdot|m_a, a)\right).$$

Lemmas 2 and 3 together imply the existence of constants $\varepsilon > 0$, $\delta > 0$ such that for all $a \in S(d)$ and all distribution functions G,

$$|\operatorname{Med}(G) - \operatorname{Med}(F^a)| \le \frac{2}{\varepsilon}\|G - F^a\|_\infty \tag{5}$$

if $\|G - F^a\|_\infty \le \delta\varepsilon/2$. From (4) we obtain

$$\sup_{a \in S(d)} \|F_n^a - F^a\|_\infty \to 0$$

as $n \to \infty$; hence (5) applies. Using (4) again and the fact that $f \to \tilde{Z}(f)$ is bounded, it follows that

$$\sup_{a \in S(d)} \sqrt{n}\left|\operatorname{Med}\left(\tilde{P}_n^a\right) - \operatorname{Med}(P^a)\right| = O(1).$$

Hence we can apply Lebesgue's dominated convergence theorem and obtain

$$\sqrt{n}\left(d\int_{S(d)} \operatorname{Med}\left(\tilde{P}_n^a\right) a\,da - d\int_{S(d)} \operatorname{Med}(P^a)\,a\,da\right)$$

$$\to Y := -d\int_{S(d)} c(a)\tilde{Z}\left(f(\cdot|m_a, a)\right) a\,da.$$

Remember the dropped $\tilde{\omega}$: the above limit refers to almost sure convergence with respect to $\tilde{\mathbb{P}}$.

The random vector Y is a bounded linear function of the Gaussian process $\tilde{Z}$; hence it is a normal random vector with mean 0 and variance matrix

$$EYY' = d^2 \int_{S(d)} \int_{S(d)} c(a)c(b) \operatorname{cov}\big(Z(f(\cdot|m_a, a)), Z(f(\cdot|m_b, b))\big) ab' \, da \, db.$$

Using (3) and the definition of ϕ we see that this is the asymptotic variance in the assertion of the theorem. It remains to go back to the untilded quantities where, due to the distributional equalities built into the construction, the same limit arises, now as the limit in distribution. □

The assumptions of the theorem are somewhat stronger than necessary. For example, it is enough if the densities exist in a neighbourhood of the respective median.

3. Symmetric distributions. In this section we take a closer look at the variance of the limit distribution in the case of radially symmetric distributions. Assume that P satisfies the following condition:

(S) P has a density f which can be written in the form $f(x) = h(\|x - x_0\|^2)$ for all $x \in \mathbb{R}^d$ with some $x_0 \in \mathbb{R}^d$, $h\colon [0, \infty) \to \mathbb{R}$, where $\sup_{0 \le x \le \varepsilon} h(x) < \infty$ for some $\varepsilon > 0$.

Multivariate normal distributions (with independent components) arise if h is a multiple of $r \to \exp(-r/2)$. If (S) holds, then the conditions of the theorem are satisfied. The symmetry centre x_0 is of no relevance to the variance matrix of the limiting normal distribution and may be taken to be 0 for notational convenience.

In this special case some explicit calculations can be carried out, resulting in a formula for the limit variance in terms of h and d.

Obviously, P^a does not depend on $a \in S(d)$. Let

$$c_0(k) := \frac{\pi^{k/2}}{\Gamma(k/2)}, \qquad c_1(h, k) := \int_0^\infty h(r) r^{k/2-1} \, dr.$$

Then $\int_{\mathbb{R}^k} h(\|x\|^2) \, dx = c_0(k) c_1(h, k)$ so that

$$f_a(0) = \int_{\mathbb{R}^{d-1}} h\big(x_2^2 + \cdots + x_d^2\big) \, dx_2 \cdots dx_d = c_0(d-1) c_1(h, d-1),$$

which gives a formula for the c-function associated with P. To obtain ϕ we first note that if X is a random vector with distribution P, then $X/\|X\|$ is uniformly distributed on $S(d)$. The function ϕ depends on P only through the distribution of this standardized random vector, which means that we may assume for the purpose of calculating ϕ that P is the d-dimensional standard normal distribution. If $|a'b| < 1$, then $(a'X, b'X)$ has a bivariate normal density to which Problem III.9.14 in Feller (1971) can be applied, resulting in

$$P((a'X)(b'X) > 0) = \frac{1}{2} + \frac{1}{\pi}\arcsin(a'b).$$

Also,

$$\phi(a,b) = P(a'X \le 0, b'X \le 0) = \tfrac{1}{2}P((a'X)(b'X) > 0)$$

in this situation, so we obtain

$$\phi(a,b) - \frac{1}{4} = \frac{1}{2\pi}\arcsin(a'b).$$

Putting together what has been obtained so far we arrive at the following expression for the asymptotic variance matrix:

$$\Sigma = \frac{d^2}{2\pi c_0(d-1)^2 c_1(h,d-1)^2}\int_{S(d)}\int_{S(d)} \arcsin(a'b)ab'\,da\,db. \tag{6}$$

The double integral can be further evaluated. Write a_i for the components of $a \in S(d)$ and let $\kappa_i := \int_{S(d)} a_i \arcsin(a_1)\,da$. Suppose that the random vector $X = (X_1,\dots,X_d)'$ is uniformly distributed on $S(d)$. Then $\kappa_i = EX_i \arcsin(X_1)$. The distribution of X remains unchanged if X_i is replaced by $-X_i$ which implies that $\kappa_i = 0$ for $i \ne 1$. Further, $Y := X_1^2$ has a beta distribution with parameters $1/2$ and $(d-1)/2$ so that

$$\begin{aligned}\kappa_1 &= \int_{S(d)} a_1 \arcsin(a_1)\,da = EY^{1/2}\arcsin Y^{1/2}\\ &= B(1/2,(d-1)/2)^{-1}\int_0^1 y^{1/2}\arcsin(y^{1/2})y^{(1/2)-1}(1-y)^{(d-1)/2-1}\,dy\\ &= (d-1)^{-1}B(1/2,(d-1)/2)^{-1}B(1/2,d/2).\end{aligned}$$

To evaluate the inner integral in (6) consider a fixed $b \in S(d)$ and let $A \in \mathcal{O}(d)$ be such that $Ab = e_1$. The uniform distribution on $S(d)$ is invariant under orthogonal transformations, so the substitution of a by Aa leads to

$$\begin{aligned}\int_{S(d)} \arcsin(a'b)ab'\,da &= A'\left(\int_{S(d)} a\arcsin(a'e_1)\,da\right)b'\\ &= \kappa_1 A'e_1 b'\\ &= \kappa_1 bb'.\end{aligned}$$

Using $\int_{S(d)} aa'\,da = (1/d)I_d$ we get

$$\int_{S(d)}\int_{S(d)} \arcsin(a'b)ab'\,da\,db = \frac{\kappa_1}{d}I_d$$

and combining these calculations we obtain the following result.

COROLLARY 1. *Assume that P satisfies* (S). *Then*

$$\sqrt{n}\,(\mathrm{OrMed}(X_1,\dots,X_n) - \mathrm{OrMed}(P)) \to_{\mathrm{distr}} Z,$$

where Z has a normal distribution with mean vector 0 *and variance matrix*

$$EZZ' = \frac{d\Gamma(d/2)^2}{(d-1)^2\pi^d c_d(h)^2} I_d \quad \textit{with} \quad c_d(h) := \int_0^\infty h(r) r^{(d-3)/2}\,dr.$$

We now compare the asymptotics of orthomedian and L^1-median for distributions satisfying (S). Obviously, we may again assume $\mathrm{L1Med}(P) = x_0 = 0$ by shift equivariance. First we note that orthogonal invariance of the distribution of the X-variates, together with orthogonal equivariance of the estimator, implies orthogonal invariance of the limit distribution. As a consequence the variance matrix of the limiting normal distribution will automatically be a multiple of the identity matrix in the situation considered in this section. This permits a very simple and direct comparison of the asymptotics of orthomedian and L^1-median (and sample mean, etc.) for distributions satisfying (S). For an elaboration of this argument in the more complicated context of affine equivariance and scale estimation, see Grübel and Rocke (1990).

The asymptotic behavior of the L^1-median has been investigated by Brown (1983), who assumed that either $d = 2$ or that the underlying distribution is normal. Brown's somewhat informal proof can be extended to general symmetric distributions and can be made precise as done by Pollard [(1984), page 152] in the two-dimensional standard normal case, resulting in the statement

$$\sqrt{n}\,(\mathrm{L1Med}(X_1,\dots,X_n) - \mathrm{L1Med}(P)) \to_{\mathrm{distr}} Z,$$

where Z is normal with mean 0 and variance matrix

$$EZZ' = \frac{d}{(d-1)^2}\left(E\|X\|^{-1}\right)^{-2} I_d.$$

How do the two multiples of the identity matrix compare? In the case of the orthomedian the dependence on h is via $f_1(0)$, the value of the density of the first component of the random vector in 0. For the L^1-median this dependence is via the moment of order -1 of the distance from the origin. However,

$$\begin{aligned} E\frac{1}{\|X\|} &= \int_{\mathbb{R}^d} \frac{1}{\|x\|} h(\|x\|^2)\,dx \\ &= \frac{\pi^{d/2}}{\Gamma(d/2)} \int_0^\infty h(r) r^{(d-3)/2}\,dr \\ &= \frac{\pi^{1/2}\Gamma((d-1)/2)}{\Gamma(d/2)} f_1(0), \end{aligned}$$

which shows that for distributions satisfying (S) these two quantities are closely related. Indeed, it follows from the above calculation that, if (S) holds, *orthomedian and L^1-median have the same limit distribution.*

4. Two-dimensional normal distributions. Let $d = 2$ and assume that P is the normal distribution on $\mathbb{R}^2$ with zero mean vector and variance matrix $\begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}$. These distributions have also been considered by Brown (1983) in connection with the L^1-median, which enables us to compare the behavior of orthomedian and L^1-median for a class of distributions that are not radially symmetric.

Let D be the diagonal matrix with diagonal elements $d_{11} = 1$ and $d_{22} = \sqrt{\lambda}$. Using the calculations performed in Section 3 we see that, for the above distribution P,

$$c(a)c(b)\left(\phi(a,b) - \frac{1}{4}\right) = \|Da\|\,\|Db\| \arcsin\left(\frac{a'D^2b}{\|Da\|\,\|Db\|}\right)$$

for all $a, b \in S(2)$. Further, $S(2)$ can be parametrized by an angle θ varying from $-\pi$ to π via

$$\theta \to a(\theta) := \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}.$$

This parametrization (which has also been used in Figure 1) preserves uniform distributions so that

$$EZZ' = \frac{1}{\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \begin{pmatrix} f_{11}(\theta,\eta) & f_{12}(\theta,\eta) \\ f_{21}(\theta,\eta) & f_{22}(\theta,\eta) \end{pmatrix} d\theta\, d\eta,$$

where

$$\begin{pmatrix} f_{11}(\theta,\eta) & f_{12}(\theta,\eta) \\ f_{21}(\theta,\eta) & f_{22}(\theta,\eta) \end{pmatrix} = \sigma(\theta,\eta)a(\theta)a(\eta)',$$

and, abbreviating $\sin(\theta), \cos(\theta)$ to S_θ, C_θ and so forth,

$$\sigma(\theta,\eta) = \sqrt{\left(C_\theta^2 + \lambda S_\theta^2\right)\left(C_\eta^2 + \lambda S_\eta^2\right)}\arcsin\left(\frac{C_\theta C_\eta + \lambda S_\theta S_\eta}{\sqrt{\left(C_\theta^2 + \lambda S_\theta^2\right)\left(C_\eta^2 + \lambda S_\eta^2\right)}}\right).$$

Inspection shows

$$f_{12}(-\theta,-\eta) = -f_{12}(\theta,\eta), \qquad f_{21}(-\theta,-\eta) = -f_{21}(\theta,\eta),$$

so that the asymptotic variance matrix is of diagonal form. I have not been able to find any further simplifications and resorted to numerical evaluation. Table 1 gives some results; it also contains the values obtained by Brown (1983) for the L^1-median.

In all cases considered, the variances for the L^1-median are smaller than the variances arising in connection with the orthomedian. However, for moderately elliptic distributions the difference is very small and should be

TABLE 1
Asymptotic variances for two-dimensional normal distributions

	First component		Second component	
λ	L1Med	OrMed	L1Med	OrMed
1	1.273	1.273	1.273	1.273
0.81	1.274	1.275	1.032	1.033
0.64	1.277	1.280	0.8175	0.8199
0.49	1.282	1.288	0.6293	0.6344
0.36	1.291	1.302	0.4668	0.4754
0.25	1.305	1.321	0.3295	0.3420
0.16	1.324	1.347	0.2168	0.2332
0.09	1.353	1.382	0.1280	0.1473
0.04	1.395	1.428	0.0624	0.0822
0.01	1.460	1.489	0.0195	0.0349
0.0025	1.507	1.527	0.00652	0.0164
0.0001	1.556	1.561	0.000624	0.003189

negligible for most practical purposes. This is confirmed by simulation experiments.

5. Computational aspects and local robustness. For two-dimensional data we can use the simple representation of $S(2)$ from Section 4 to approximate the integral in (2) by an associated Riemann sum to any desired degree of precision. However, a naive extension of this approach to dimensions greater than 2 can be very inefficient, and it is much better to use (2) directly with the uniform distribution on $S(d)$ replaced by the empirical distribution associated with N suitably chosen points $a_1, \dots, a_N$ from the sphere. The orthomedian of the data set $x_1, \dots, x_n$ can then be approximated by the mean of the projected medians $\mathrm{Med}(a_i' x_1, \dots, a_i' x_n) a_i$, $i = 1, \dots, N$, multiplied by d. To ensure that these approximations are shift equivariant for a given N, and not just in the limit as $N \to \infty$, it is advisable to center the data about the mean first.

There are two main methods for choosing $a_1, \dots, a_N$. The "number–theoretic method" (NTM) aims to choose these systematically and evenly spaced on $S(d)$ for given N and d, whereas simple Monte Carlo integration (MCI) uses a sequence $a_1, \dots, a_N$ of independent and uniformly distributed random elements of $S(d)$ (here and in the following "random" occasionally refers to pseudo-random quantities generated on the computer; what is meant should be clear from the context). The recent monograph by Fang and Wang (1994) explains NTM and provides sufficient detail to allow a relatively straightforward implementation for a variety of values of N and d. MCI can easily be implemented without any restriction on N or d. Uniformly distributed random elements Y of $S(d)$ can be obtained by generating d independent standard normal variates $X_1, \dots, X_d$ and using $Y = X/\|X\|$, $X = (X_1, \dots, X_d)'$. Both methods will give $\mathrm{OrMed}(x_1, \dots, x_n)$ in the limit as $N \to \infty$, MCI with probability 1. For NTM, which is also known as quasi Monte Carlo integration, theoretical results are available which predict that for sufficiently smooth integrands the approximation error is on the order N^{-1} [apart from logarithmic factors; see, e.g., Theorem 2.5 in Fang and Wang (1994)]. For MCI we only have the rate $N^{-1/2}$. Practical experience shows that both methods are viable. Computation times are similar for both methods and depend linearly on n, d and N. On a PC with Intel 90 MHz Pentium processor, 0.656 seconds were needed for MCI with $n = 200$, $d = 10$ and $N = 1000$.

The error introduced by these approximations will be negligible if N is of larger order of magnitude than n, a situation somewhat similar to choosing the number of resamples in a bootstrap procedure if we consider MCI. For MCI, guidelines of a more quantitative nature can be developed on noting that, conditionally on the data set, $(\xi_i)_{i\in\mathbb{N}}$ with $\xi_i := d\,\mathrm{Med}(a_i' x_1, \dots, a_i' x_n) a_i$, is a sequence of bounded, independent and identically distributed random vectors. The Monte Carlo approximation for $\mathrm{OrMed}(x_1, \dots, x_n)$ is the average of the first N of these, so an estimator for the covariance matrix of ξ_1 together with the central limit theorem leads to confidence bounds for the approximation error. The situation here is somewhat more complicated than in the familiar one-dimensional case since we would need the quantiles of a sum of d squared centered normals with nonunit variance, so we propose the following simple alternative: the trace of the empirical covariance matrix Σ_N associated with $\xi_1, \dots, \xi_N$ is an unbiased estimate of $E\|Z_n\|^2$, where Z_n denotes the limit in distribution of

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\xi_i - \mathrm{OrMed}(x_1, \dots, x_n)\right), \qquad N \to \infty.$$

This leads to $N^{-1}\,\mathrm{tr}(\Sigma_N)$ as an estimator for the expected squared distance between the orthomedian and its Monte Carlo approximation. To give a numerical example, for a sample of 200 ten-dimensional standard normal random vectors the value obtained after 1000 Monte Carlo repetitions was 0.000254. The actual deviations obtained in five additional runs on the same sample, with different sequences of random numbers, were 0.000332, 0.000430, 0.000244, 0.000172 and 0.000512. Here the result obtained with $N = 10^6$ was taken to be the true orthomedian for the data set in question. Incidentally, our theoretical results from the previous sections predict the value 0.082558 for the expectation of $\|\mathrm{OrMed}(X_1, \dots, X_n) - \mathrm{OrMed}(P)\|^2$ if $P = N_{10}(0, I_{10})$, that is, in this situation the error introduced by MCI is small as compared to the error inherent in the estimation procedure.

Indeed, for moderate values of sample size and dimension, $N = 1000$ should be adequate, but, as is obvious from the timing given above, larger values of N are entirely feasible. If interest is primarily in the detection of outliers or if the orthomedian is just the first step in a multistep procedure designed to combine robustness and high efficiency, then even lower values for N should suffice.

Again, comparison with the L^1-median is illuminating. The latter is defined, like many other multivariate variants of the median, as the minimizing value of some nondifferentiable function; here

$$\psi: \mathbb{R}^d \to \mathbb{R}, \qquad \psi(\theta) := \sum_{i=1}^{n} \|x_i - \theta\|.$$

One might think that the critical set $\{x_1, \dots, x_n\}$ of ψ is negligible for computational purposes, but this is not the case: the minimizing value will be x_i if

$$\left\| \sum_{j \neq i} \frac{1}{\|x_i - x_j\|} (x_i - x_j) \right\| < 1, \tag{7}$$

and this will happen with positive probability if, for example, the data values are a sample from an absolutely continuous distribution such as a nondegenerate multivariate normal distribution. As a result, calculation of L1Med($x_1, \dots, x_n$) is a nontrivial exercise; see Bedall and Zimmermann (1979) for a good solution. The dependence of the computation time on sample size n and dimension d of the data is of the form $(n \log n)d^2$ with this algorithm, but this does not seem to matter unless sample size or dimension become truely excessive. For example, computing the L^1-median for a sample of 200 ten-dimensional standard normal random vectors took 0.104 seconds (in contrast to the above algorithms for the orthomedian, computation times now depend on the sample configuration).

The fact that the L^1-median can be one of the data points also has statistical consequences: if the sample configuration is such that (7) holds, then configurations which arise by changing x_i slightly will also satisfy this inequality, that is, x_i will continue to be the L^1-median. This means that the local robustness properties of the L^1-median are poor in the sense that, with positive probability, a change of one of the data values will lead to a change in the estimate of the same magnitude [see Hampel, Ronchetti, Rousseeuw and Stahel (1986) for a discussion of local-shift sensitivity and related issues].

Although it is based on coordinatewise medians, which are similarly susceptible to small fluctuations in the data, the local robustness properties of the orthomedian are better than those of the L^1-median. This is due to the integration step—roughly, the orthomedian is a "mean of medians." To explain this, we invoke Figure 1 one last time: part (b) shows the index of the data value leading to the projected median in part (c) as a function of the angle θ of rotation. Obviously, the influence of any particular data point x_i is restricted to a small part of the integration range. To obtain a geometric understanding of the size of these intervals, imagine a line drawn through x_i. If the sample is from an absolutely continuous distribution, then it will be possible (with probability 1) to find such a line that splits the sample into two halves of equal size. The angle by which this line can then be rotated about x_i without hitting some other x-value gives the range of influence of x_i. This interpretation also makes it obvious that outlying observations have little influence on the estimator.

6. Concluding remarks.

In this section we indicate some possible extensions of the methods and results of the previous sections.

6.1. It has already been mentioned in Section 2 that our proof of asymptotic normality of the orthomedian is based on a (weak) differentiability property of the functional $P \to \mathrm{OrMed}(P)$. Properties of this type can also be used to show that the bootstrap "works," leading to associated confidence regions; see Gill (1989) for the one-dimensional case and Section 4 of Arcones and Giné (1992) and Chapter 3.9.3 of van der Vaart and Wellner (1996).

6.2. Averaging over $\mathscr{O}(d)$ as in (1) produces orthogonal equivariant estimators, even if the base estimator is not obtained by componentwise application of some one-dimensional estimator as in the case of the orthomedian. In this more general situation the reduction to $S(d)$ explained at the beginning of Section 2 might not be possible, but the basic idea for the Monte Carlo approximation of the estimator given in Section 5 still applies. Using (1) instead of (2) requires a sequence of independent and uniformly distributed elements of $\mathscr{O}(d)$ rather than $S(d)$. Techniques to generate such random orthogonal matrices are described in Heiberger (1978) [see also Tanner and Thisted (1982)] and Anderson, Ingram and Underhill (1987). If the base estimator is shift equivariant, then the resulting Monte Carlo approximations will also be shift equivariant.

6.3. The method introduced in the previous sections can be used quite generally to "lift" shift equivariant one-dimensional location estimators to higher dimensions, resulting in shift and orthogonal equivariant multivariate location estimators (O-estimators, if one so wishes). This was carried out above for the median, but other interesting possibilities exist. As a particular case consider trimmed means: in dimension 1 it is quite clear what is meant by removing the largest and smallest 10% of the data, but, as in the case of the median, there is no canonical generalization of this procedure to dimensions higher than 1, due to the lack of a suitable order structure. However, the transition to an "ortho-trimmed mean" can be carried out in complete analogy to the transition from one-dimensional median to orthomedian. The techniques used above in the latter context can be adapted to the analysis of estimators of this type. Asymptotic normality, for example, would again follow from a delta method based proof of the one-dimensional result if this can be made to hold uniformly in all projections (i.e., one would need the corresponding variants of the lemmas in Section 2; the proof of the theorem then carries over almost literally). The comments on numerical aspects in Section 5 apply without change.

Acknowledgments. I would like to thank Ludwig Baringhaus for helpful discussions and the referees for their stimulating comments on the first version of this paper.

REFERENCES

ANDERSON, T. W., INGRAM, I. and UNDERHILL, L. G. (1987). Generation of random orthogonal matrices. *SIAM J. Sci. Statist. Comput.* **8** 625–629.

ARCONES, M. A. and GINÉ, E. (1992). On the bootstrap of *M*-estimators and other statistical functionals. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 13–47. Wiley, New York.

BEDALL, F. K. and ZIMMERMANN, H. (1979). AS143: the mediancentre. *J. Roy. Statist. Soc. Ser. C* **28** 325–328.

BROWN, B. M. (1983). Statistical uses of the spatial median. *J. Roy. Statist. Soc. Ser. B* **45** 25–30.

FANG, K. T. and WANG, Y. (1994). *Number-Theoretic Methods in Statistics*. Chapman and Hall, London.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part I). *Scand. J. Statist.* **16** 97–128.

GRÜBEL, R. (1988). The length of the shorth. *Ann. Statist.* **16** 619–628.

GRÜBEL, R. and ROCKE, D. M. (1990). On the cumulants of affine equivariant estimators in elliptical families. *J. Multivariate Anal.* **35** 203–222.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

HEIBERGER, R. M. (1978). Algorithm 127. Generation of random orthogonal matrices. *Appl. Statist.* **27** 199–206.

LOPUHAÄ, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19** 229–248.

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

SMALL, C. G. (1990). A survey of multidimensional medians. *Internat. Statist. Rev.* **58** 263–277.

TANNER, M. A. and THISTED, R. A. (1982). A remark on AS 127. Generation of random orthogonal matrices. *Appl. Statist.* **31** 190–192.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

VERVAAT, W. (1972). Functional central limit theorems for processes with positive drift and their inverses. *Z. Wahrsch. Verw. Gebiete* **23** 245–253.

INSTITUT FÜR MATHEMATISCHE STOCHASTIK
UNIVERSITÄT HANNOVER
POSTFACH 6009
30060 HANNOVER
GERMANY
E-MAIL: rgrubel@stochastik.uni-hannover.de