A TWO-STAGE RANK TEST USING DENSITY ESTIMATION

WILLEM ALBERS

Department of Applied Mathematics, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands

Abstract. For the one-sample problem, a two-stage rank test is derived which realizes a required power against a given local alternative, for all sufficiently smooth underlying distributions. This is achieved using asymptotic expansions resulting in a precision of order m^{-1}, where m is the size of the first sample. The size of the second sample is derived through a number of estimators of e.g. integrated squared densities and density derivatives, all based on the first sample. The resulting procedure can be viewed as a nonparametric analogue of the classical Stein's two-stage procedure, which uses a t-test and assumes normality for the underlying distribution. The present approach also generalizes earlier work which extended the classical method to parametric families of distributions.

1. Introduction

Let $X_1, X_2, \ldots$ be independent identically distributed (iid) random variables (rv's) from a continuous distribution function (df) $F(x-\theta)$ with $F(-x) = 1 - F(x)$ for all x. The hypothesis $H_0 : \theta = 0$ is to be tested on the basis of a sample from the sequence $X_1, X_2, \ldots$. In the classical case the problem is specialized to $F(x) = \Phi(x/\sigma)$, where Φ is the standard normal df. Stein's two-stage procedure for this situation is well-known (see e.g. Lehmann (1986), pp. 258–260). An initial sample of size m is drawn and its sample variance S_m^2 is obtained. In the second stage a sample of size $N - m$ is selected, where

$$(1.1) \qquad N = \max(m, [S_m^2/c] + 1),$$

with $c > 0$ any given constant and $[y]$ denoting the largest integer $\leq y$. Then $N^{1/2}(N^{-1}\sum_{i=1}^{N} X_i - \theta)/S_m$ has a t_{m-1}-distribution and the constant c from (1.1) can be chosen such that the corresponding two-stage test has a prescribed power against a given alternative, regardless of the value of the scale parameter σ.

Now fixed sample size rank tests, being distributionfree, have levels independent of F, but powers directly dependent of F. Hence it seems worthwhile to try to

extend the property above of having powers independent (of certain aspects) of the unknown underlying df, to the rank case. In fact, results in this direction have already been presented in Albers (1991, 1992). In these papers attention is focussed on a straightforward extension of Stein's procedure: using asymptotic expansions for local alternatives, a two-stage rank test is derived which has to $o(m^{-1})$ a power independent of the scale parameter σ. Of course, if we are not constrained to the classical case $F(x) = \Phi(x/\sigma)$, there no longer is a special reason to let F belong to a scale family. Without essential difficulties, the results for the scale case can be extended to arbitrary parametric families of underlying df's. However, it does not seem particularly rewarding to pursue this: the resulting expressions become typically more complicated than in the scale case, while lacking the natural appeal of this latter case as a direct analogue of a well-known classical procedure.

A more appealing extension of the results for the scale case can be achieved by applying a nonparametric approach, using density estimation techniques. In this way, a two-stage rank test can be obtained which has, once more to $o(m^{-1})$, a power independent of F, for all F satisfying certain regularity conditions. The execution of this program will be the purpose of the present paper. Some comments to be made in advance are the following. Clearly, the idea is not new to estimate the expression depending on the unknown F which occurs in the asymptotic power, see e.g. Sen and Ghosh (1971, 1974) and Schweder (1975). Here we apply it to the two-stage case, and, more importantly, we use second order approximations based on asymptotic expansions to $o(m^{-1})$, rather than on mere first order asymptotic results. This entails that also second order terms, and hence derivatives of the density f, have to be estimated.

As a consequence, the results tend to become rather technical. To improve the accessibility and to allow implementation in practice, we therefore specialize to the case of Wilcoxon scores from a certain point on. In this situation, the first order term to be estimated depends on the unknown F through $\int_{-\infty}^{\infty} f^2(x)dx$, which we will denote by $\int f^2$ from now on. Fortunately, a multitude of, mostly quite recent, results are available on rates of convergence in estimating this functional. It occurs e.g. in Goldstein and Messer ((1992), p. 1318) and in Donoho and Liu ((1991), "an interesting example", p. 654). Both these papers refer to Ritov and Bickel (1990), where it is shown that without sufficient smoothness $\int f^2$ is not estimable at rate $m^{-\alpha}$ for any $\alpha > 0$. However, if the density has sufficient derivatives, the rate will be $m^{-1/2}$. This is the case we shall consider, as the use of asymptotic expansions to $o(m^{-1})$ already presupposes the existence of several derivatives of f. A useful reference for this situation is Hall and Marron (1987), together with subsequent papers such as Jones and Sheather (1991) and Sheather *et al.* (1992).

In Section 2 we present the asymptotic expansion for the two-stage rank test, which we apply in Section 3 to obtain tests with prescribed power. Specialization to Wilcoxon scores occurs in Sections 4 and 5. First we collect all density-type estimators involved in Section 4. Then in Section 5, we present and comment on explicit formulae for the total sample size N required to ensure a given power to $o(m^{-1})$ against a given alternative. An example and some simulation results conclude the paper.

2. The asymptotic expansion for the two-stage rank test

In this section we introduce the notation and quote the relevant results from Albers (1992). Let $X_1, X_2, \ldots$ be iid. rv's with continuous df $F(x - \theta)$, where F is such that $F(-x) = 1 - F(x)$ for all x. For a first sample of size m, let $Z_{(m)} = (Z_1, \ldots, Z_m)$ denote the vector of order statistics of $|X_1|, \ldots, |X_m|$. Moreover, for $j = 1, \ldots, m$, introduce V_j, which equals 1 if the X_i corresponding to Z_j is positive and which equals 0 otherwise. This leads to a rank statistic for testing $H_0 : \theta = 0$ given by

$$(2.1) \qquad T_1 = \sum_{j=1}^{m} a_{1j} V_j,$$

where $a_1 = (a_{11}, \ldots, a_{1m})$ is a vector of scores. We shall typically consider exact scores $a_{1j} = EJ(U_{j:m})$, where J is a continuous function on $(0, 1)$ and $U_{1:m} < \cdots < U_{m:m}$ are order statistics for a sample of size m from the uniform distribution on $(0, 1)$.

For the second stage of the test we select an additional sample of size $N - m$, where $N = N(X_1, \ldots, X_m)$ in general. We shall require that $N = N(Z_{(m)})$ in what follows. Since T_1 and $Z_{(m)}$ are independent under H_0, this ensures that the two-stage procedure remains distribution-free as well. The second sample, in combination with scores $a_{2j} = EJ(U_{j:N-m})$, $j = 1, \ldots, N - m$, leads to a rank statistic T_2, analogous to T_1, from (2.1). The two separate statistics are simply combined to the final statistic

$$(2.2) \qquad T = T_1 + T_2.$$

Of course, it would be slightly better to use a rank statistic of the form (2.1) based on the total sample of size N. But, as is argued in more detail in Albers (1992), it is less tedious to obtain an asymptotic expansion for the df of T from (2.2), whereas the loss due to using T is typically compensated by as little as a single additional observation.

Before introducing the asymptotic expansion in question, we shall first list the conditions involved on the estimator N, the df F and the score function J. Let $\mathcal{N}$ be the class of estimators $N = N(Z_{(m)})$ such that

$$(2.3) \qquad P((1 + \epsilon) \leq N/m \leq \epsilon^{-1}) = 1 - o(m^{-1}),$$

$$(2.4) \qquad E|N - EN|^{2\beta} = \mathcal{O}(m^{\beta}),$$

for some $\epsilon > 0$ and $\beta > 1$. Condition (2.3) ensures that $N - m$ will tend to infinity of the same rate as m, except for a sufficiently small probability, while (2.4) implies that $N - EN = \mathcal{O}_P(m^{1/2})$. Let $\mathcal{Q}$ be the class of twice continuously differentiable functions Q on $(0, 1)$ that satisfy

$$(2.5) \qquad \limsup_{t \to 0, 1} t(1 - t) \left| \frac{Q''(t)}{Q'(t)} \right| < \frac{3}{2}.$$

Let $\mathcal{F}$ be the class of df's on $\mathbb{R}^1$ with positive densities that are symmetric about zero, four times differentiable, and such that for $\psi_i = f^{(i)}/f$, $\Psi_i(t) = \psi_i(F^{-1}((1+t)/2))$, $m_1 = 6$, $m_2 = 3$, $m_3 = 4/3$, $m_4 = 1$, we have $\Psi_1 \in \mathcal{Q}$ (see (2.5)) and

$$(2.6) \qquad \limsup_{y \to 0} \int_{-\infty}^{\infty} |\psi_i(x+y)|^{m_i} f(x) dx < \infty, \qquad i = 1, \ldots, 4.$$

Let $\mathcal{J}$ be the class of nonconstant functions J on $(0,1)$ that satisfy $J \in \mathcal{Q}$ and $\int_0^1 J^4(t) dt < \infty$.

Next we give the expansion. Let ϕ denote Φ' and let $\phi^{(k)}$ denote the k-th derivative of ϕ. The Hermite polynomial of degree k is given by $H_k(x) = (-1)^k \phi^{(k)}(x)/\phi(x)$, $k = 0, 1, \ldots$. Using the convention that integration will be over $(0,1)$, unless stated otherwise, we define

$$\begin{aligned}
(2.7) \qquad
&\eta_m = -\frac{m^{1/2}\theta \int J\Psi_1}{(\int J^2)^{1/2}}, \\
&b_0 = -\frac{\int J(3\Psi_1^3 - 6\Psi_1\Psi_2 + \Psi_3) \int J^2}{6(\int J\Psi_1)^3}, \\
&b_1 = \frac{\int J^2\Psi_1^2 - \int\int J(s)\Psi_1'(s)J(t)\Psi_1'(t)(s \wedge t - st)ds dt}{2(\int J\Psi_1)^2}, \\
&b_2 = \frac{\int J^3\Psi_1}{3(\int J^2)(\int J\Psi_1)}, \\
&b_3 = \frac{\int J^4}{12(\int J^2)^2}, \\
&\tilde{b}_{0,m} = \frac{2\sum_{j=1}^m \mathrm{Cov}_j}{\int J\Psi_1} - \frac{\sum_{j=1}^m \sigma_j^2}{\int J^2},
\end{aligned}$$

where $\mathrm{Cov}_j = \mathrm{Cov}(J(U_{j:m}), \Psi_1(U_{j:m}))$ and $\sigma_j^2 = \sigma^2(J(U_{j:m}))$. In addition, for $r \geq (1+2\epsilon)m$, define

$$(2.8) \quad \bar{U} = (N/r)^{1/2} - 1,$$

$$(2.9) \quad \bar{H}(x) = \Phi(x) + \phi(x)\left\{ \frac{1}{2}r^{-1}\eta_r(\tilde{b}_{0,m} + \tilde{b}_{0,[r]-m}) \right.$$

$$+ r^{-1}\sum_{k=0}^3 \eta_r^{(3-k)} b_k H_k(x) - \eta_r E\bar{U}$$

$$- \frac{1}{2}\eta_r^2 E\bar{U}^2 H_1(x)$$

$$\left. + \frac{r^{-1}\eta_r E(\bar{U}\sum a_{1j}(\psi_1(Z_j) - E\psi_1(Z_j)))}{\int J\Psi_1} \right\}.$$

Then we have (see Albers (1992))

THEOREM 2.1. *Let $N \in \mathcal{N}$, $F \in \mathcal{F}$ and $J \in \mathcal{J}$. Suppose that $0 \leq \theta \leq Cm^{-1/2}$ for some $C > 0$. If r in (2.8) is chosen such that $r = EN + o(m^{1/2})$, we have for T from (2.2) that*

$$(2.10) \qquad \sup_x \left| P\left(\frac{2T - \sum a_{1j} - \sum a_{2j}}{(\sum a_{1j}^2 + \sum a_{2j}^2)^{1/2}} \leq x \right) - \bar{H}(x - \eta_r) \right| = o(m^{-1}),$$

while all terms in $\bar{H}(x - \eta_r)$ beyond $\Phi(x - \eta_r)$ are $o(m^{-1/2})$.

To illuminate the relevance of Theorem 2.1 for our purpose, we observe the following points. The two-stage procedure is performed conditionally. In fact, rejection of $H_0 : \theta = 0$ in favor of $H_1 : \theta > 0$ for large values of T standardized as in (2.10), leads to a critical value $\xi_\alpha = \tilde{\xi}_\alpha + o(N^{-1})$, where

$$(2.11) \qquad \tilde{\xi}_\alpha = u_\alpha - N^{-1} b_3 H_3(u_\alpha)$$

with $u_\alpha = \Phi^{-1}(1 - \alpha)$ (cf. once more Albers (1992)). But replacement of N^{-1} in this $\tilde{\xi}_\alpha$ by r^{-1} such that $r = EN + o(m^{1/2})$ with probability $1 - o(m^{-1})$, causes differences of $o(m^{-1})$. Hence the power of the unconditional test based on the standardized T from (2.10) will agree to $o(m^{-1})$ with the power of the two-stage procedure to be investigated. Consequently, it makes sense to use the unconditional expansion presented in Theorem 2.1.

To shed some light on the expansion itself, we remark that the result is obtained by first taking the expansion for a fixed sample size rank test, using the (to order $m^{1/2}$) expected sample size r. The expression for $\bar{H}$ in (2.9) then arises by adding three terms involving $\bar{U}$ from (2.8) to take into account that N rather than r is used, and by using $\tilde{b}_{0,m} + \tilde{b}_{0,[r]-m}$ instead of $\tilde{b}_{0,[r]}$ to account for the use of two separate rank statistics T_1 and T_2 (cf. (2.2)) instead of one overall statistic.

3. Two-stage tests with prescribed power

Using the results of the previous section we shall indicate for the case of general J how a prescribed power can be realized to $o(m^{-1})$ against a given alternative. Before going into some detail we shall give an overview of the steps involved. Let $\pi(\theta)$ denote the power of the two-stage test which rejects whenever the standardized T from (2.10) exceeds $\tilde{\xi}_\alpha$ given in (2.11). Suppose we require for certain κ and π_1 that

$$(3.1) \qquad \pi(\kappa m^{-1/2}) = \pi_1.$$

As Theorem 2.1 amply implies that $\pi(\theta) = 1 - \Phi(u_\alpha - \eta_r) + o(m^{-1/2})$, it follows through (2.7) that (3.1) is satisfied to first order for

$$(3.2) \qquad r = \frac{m(u_\alpha - u_\pi)^2 \int J^2}{\kappa^2 (- \int J\Psi_1)^2},$$

where $u_\pi = \Phi^{-1}(1 - \pi_1)$. Note that the requirement that $r \geq (1 + 2\epsilon)m$ for some $\epsilon > 0$ will be met if κ in (3.2) is sufficiently small, i.e. if $\theta = \kappa m^{-1/2}$ is sufficiently close to $H_0 : \theta = 0$.

The form of the leading term N_1 of the desired estimator is readily suggested by (3.2): if W is a suitable estimator for $-\int J\Psi_1$, then define

$$(3.3) \qquad N_1 = r\frac{(-\int J\Psi_1)^2}{W^2} = \frac{m(u_\alpha - u_\pi)^2 \int J^2}{\kappa^2 W^2}.$$

The next step consists of adding a correction term f_r to N_1, selected through (2.8)–(2.10) in such a way that all lower order terms in the expansion are precisely cancelled, thus ensuring that $N_1 + f_r$ will produce the desired realization of (3.1) to $o(m^{-1})$.

Obviously this f_r will involve unknown expressions like $\int J\Psi_1\Psi_2$ and $\int J^3\Psi_1$ (cf. (2.7)), which will have to be estimated in their turn. But as only lower order terms are involved, replacement of f_r by a consistent estimator $\hat{f}_r$ will lead to negligible differences. It remains to add the final step $N = \max(m, [N_1 + \hat{f}_r + \frac{1}{2}])$ (cf. (1.1)).

Hence the problem has now essentially been reduced to estimating a number of integrals involving Ψ_i, $i = 1, 2, 3$. Usually consistency will suffice, but for the estimator W of $-\int J\Psi_1$ in view of (3.3), (2.8) and (2.9), second order properties will be needed. Several approaches are possible for $-\int J\Psi_1$. For example, Sen and Ghosh (1971, 1974) use rank statistics based on shifted samples. Another possibility is to rewrite $-\int J\Psi_1$, using that $\Psi_1(t) = (f'/f)(F^{-1}((1 + t)/2))$, into $2\int J'(t)/q((1 + t)/2)dt$, where $q = 1/f \circ F^{-1} = (F^{-1})'$ is the so-called quantile density. Methods for estimating q are discussed in Jones (1992). Here we shall adopt a rather direct approach, used e.g. by Schweder (1975).

Define $H = 2F - 1$, $h = H' = 2f$ and introduce the (extended) empirical df for $|X_1|, \ldots, |X_m|$ and all y:

$$(3.4) \qquad H_m(y) = F_m(y) - F_m(-y) = \frac{1}{m} \times \#\{|X_i| \leq y\} \operatorname{sign} y.$$

Using that $\Psi_1 = \frac{h'}{h} \circ H^{-1}$, the integral $-\int J\Psi_1$ can be rewritten into $\int J'h \circ H^{-1} = \int_0^\infty J'(H(x))h(x)dH(x)$, which can be estimated by

$$(3.5) \qquad \frac{1}{m} \sum_{i=1}^m J'(H_m(|X_i|))\hat{h}(|X_i|),$$

where $\hat{h}$ is a density estimator for h. We shall use a kernel type choice for $\hat{h}$. To be more precise, let k be a density which is symmetric around zero and which has a finite, positive second moment. For bandwidth τ we then have

$$(3.6) \qquad \hat{h}(y) = \int\limits_{-\infty}^{\infty} k_\tau(y - z)dH_m(z),$$

where $k_\tau(z) = \tau^{-1}k(z/\tau)$. Using that $H_m(-y) = -H_m(y)$, we can transform (3.6) into $\hat{h}(y) = \int_0^\infty \{k_\tau(y-z) + k_\tau(y+z)\}dH_m(z) = m^{-1}\sum_{j=1}^m \{k_\tau(y-|X_j|) + k_\tau(y+|X_j|)\}$. Together with (3.5) this leads to

$$(3.7) \quad W = \frac{1}{m(m-1)}\sum_{i\neq j}\sum J'(H_m(|X_i|))\{k_\tau(|X_i| - |X_j|) + k_\tau(|X_i| + |X_j|)\}.$$

Note that we have omitted the non-stochastic diagonal term. The interesting discussion in itself about possible advantages of including such a term (cf. Jones and Sheather (1991)) is not relevant here, as its contribution in our approach would be automatically cancelled again by the second order term involving $E\bar{U}$ in (2.9). Such a cancellation (at least to first order) also occurs for the bias due to having X_j from $F(x - \theta)$, rather than from $F(x)$. In the next section we shall be more explicit about this point (see the remark following (4.2)).

Next we consider the lower order terms. Those involving b_i, $i = 0, 1, 2, 3$, can be dealt with in an analogous way as $-\int J\Psi_1$. For example, letting $J(0) = 0$ without loss of generality, the integral $\int J(3\Psi_1^3 - 6\Psi_1\Psi_2 + \Psi_3)$ in b_0 leads to $\int_0^\infty \{-4J'(H(x))h''(x) - 3J''(H(x))h'(x)h(x)\}dH(x)$, which can be estimated by

$$(3.8) \qquad \frac{1}{m}\sum_{i=1}^m \{-4J'(H_m(|X_i|))\hat{h}''(|X_i|) - 3J''(H_m(|X_i|))\hat{h}'(|X_i|)\hat{h}(|X_i|)\},$$

where $\hat{h}'$ and $\hat{h}''$ are suitable estimators of the derivatives h' and h''. To deal with the terms involving $\bar{U}$ in (2.9) we observe through (2.8) and (3.2) that $\bar{U} = (-\int J\Psi_1)/W - 1$. Hence $E\bar{U}$ and $E\bar{U}^2$ can be evaluated once we know bias and variance of W to $o(m^{-1})$. Results of this type can be found in Schweder ((1975), e.g. p. 115, Theorem 1; see Schweder (1981) for a correction note!).

However, we shall not bother to present all these estimators here explicitly. One reason is that, like the one in (3.8), they tend to be rather complicated. But, more importantly, (3.8) and (3.7) also serve to make clear that rather severe conditions on J will be needed to ensure that the estimators involved are well-behaved. Typically, boundedness of J, or even of the derivatives J' and J'' occurring in (3.8), will be imposed, thus ruling out the important choice $J = \Phi^{-1}$, corresponding to the normal scores test (cf. the remark to this effect in Schweder (1975), p. 115). There are ways of avoiding such assumptions and including $J = \Phi^{-1}$ (see e.g. Albers (1985)), but that would involve adding a lot of new technicalities on top of what we have already.

In view of the above we prefer to specialize from this point on to the case of Wilcoxon scores $J(t) = t$, as announced in the introduction. In this way we regain relatively simple estimators which will typically be well-behaved without additional effort or conditions.

4. Wilcoxon scores: the estimation

For the choice $J(t) = t$ the integral $-\int J\Psi_1$ reduces to $2\int_{-\infty}^{\infty} f^2$, while (3.7) simplifies to

$$(4.1) \qquad W = \frac{1}{m(m-1)} \sum_{i \neq j} \sum \{k_\tau(X_i - X_j) + k_\tau(X_i + X_j)\}.$$

Note that in (4.1) we have replaced the absolute values by the rv's themselves, which is allowed by virtue of the symmetry of k_τ. Nevertheless, it remains true that W, and hence N, depends on $X_1, \ldots, X_m$ only through $Z_{(m)}$ as required. Expressions for bias and variance of W are easily derived using results of Hall and Marron (1987). To be totally explicit, we specialize from now on, unless stated otherwise, to the Parzen-Rosenblatt kernel estimator, using the uniform kernel $k(s) = \frac{1}{2}$ for $|s| < 1$ and $k(s) = 0$ otherwise. Then we have

$$(4.2) \quad \begin{aligned} EW - 2\int f^2 &= \left(2\theta^2 + \frac{\tau^2}{3}\right) \int f f'' + \mathcal{O}(\theta^4 + \tau^4), \\ \operatorname{var} W &= \frac{16}{m}\left\{\int f^3 - \left(\int f^2\right)^2\right\} + \frac{2}{m^2\tau}\int f^2 + o(m^{-1} + m^{-2}\tau^{-1}), \end{aligned}$$

where we adopt the convention that integrals involving f will be from $-\infty$ to ∞, unless stated otherwise.

Some comments are useful here. In the first place, W contains two types of bias. As the X_j are from $F(x - \theta)$ rather than from $F(x)$, we in fact estimate $2\int f^2(x - \theta)dx$ rather than $2\int f^2$. This produces the term $2\theta^2 \int f f''$ in (4.2). A remedy might seem to be the use of $X_j - \kappa m^{-1/2}$ rather than X_j, as this would precisely eliminate this bias in the point of interest $\theta = \kappa m^{-1/2}$. However, this alignment destroys the property that W depends on $X_1, \ldots, X_m$ through $Z_{(m)}$ only, and is therefore not allowed. Obviously, the term $\tau^2 \int f f''/3$ represents the "genuine" bias, due to the estimation. As concerns the variance in (4.2), note that it is of order $m^{-1} + m^{-2}\tau^{-1}$ rather than the more customary order $m^{-1}\tau^{-1}$. This better rate is due to the fact that we are estimating integrated squared smooth densities, rather than densities themselves. Under these circumstances it is clear that Theorem 2.1 is applicable when using W from (4.1). To avoid lengthy expositions, we leave verification of the remaining minor technicalities to the reader.

Next we consider the choice of bandwidth τ in (4.1). The usual approach of minimizing the squared bias plus variance is not relevant here. By expanding $\bar{U} = 2\int f^2/W - 1$ in terms of $(W - 2\int f^2)/(2\int f^2)$, we observe from (2.9) that bias and variance should be balanced, rather than squared bias and variance (cf. Hall and Sheather (1988), p. 384 for the same phenomenon). However, for the leading terms in either case we shall compensate by estimating $\int f f''$ and $\{\int f^3 - (\int f^2)^2\}$, respectively. The actual error committed will thus be $\mathcal{O}(\theta^4 + \tau^4) + \mathcal{O}(m^{-2}\tau) + o(m^{-1})$. This will be minimized with respect to τ by letting $\tau \sim m^{-2/5}$. Just as in the mean integrated squared error case, we could try to estimate the appropriate

coefficient c in $\tau = cm^{-2/5}$. But in the present case this will even be harder than for MISE. More importantly, as m is the size of the first sample, it will typically be rather small, e.g. $10 \le m \le 20$. Referring once more to Hall and Sheather (1988), we prefer to use $m^{-2/5}$ as a qualitative guide only, as for $m = 10$, values of $m^{-2/5}$, $m^{-1/2}$ and $m^{-1/3}$ do not differ greatly (Hall and Sheather (1988), p. 382). As $\tau \sim m^{-2/5}$, while $\theta \sim m^{-1/2}$, an explicit proposal is to let $\tau = c\kappa m^{-1/2}$ with $c \in [1.5, 2]$.

Summarizing for a moment, we have introduced W in (4.1), found its bias and variance in (4.2) and also determined a reasonable choice of bandwidth τ. What remains concerning W is to find estimators for $\int ff''$ and $\int f^3$ (cf. (4.2)) and of course, to deal with the last term in (2.9) which is a mixed term involving $\bar{U}$ and thus W. Following Schweder (1975) or Hall and Marron (1987) we have for $\int ff''$ the estimator

$$(4.3) \qquad \frac{1}{m} \sum_{i=1}^{m} \hat{f}''(X_i),$$

where $\hat{f}''$ is a suitable estimator for f''. One possibility is to use $\hat{f}''(x) = m^{-1} \sum_{i=1}^{m} k_\tau''(x - X_j)$, with k sufficiently smooth, e.g. $k = \phi$. Here we shall simply use

$$(4.4) \qquad V_0 = \frac{1}{2m(m-1)} \sum \sum_{i \ne j} \{\tilde{k}_{\tau_0}(X_i - X_j) + \tilde{k}_{\tau_0}(X_i + X_j)\},$$

where $\tilde{k}_{\tau_0}(s) = \tau_0^{-3}\tilde{k}(s/\tau_0)$, $\tilde{k}(s) = -4$ for $|s| \le \frac{1}{2}$, $\tilde{k}(s) = 4$ for $\frac{1}{2} < |s| \le 1$ and $\tilde{k}(s) = 0$ otherwise. We use τ_0 rather than τ for the bandwidth to stress that for W from (4.1) and V_0 from (4.4) different choices of bandwidth are possible. Estimation of $\int f^3$ is more straightforward: use

$$(4.5) \qquad V_1 = \frac{1}{m} \sum_{i=1}^{m} \hat{f}^2(X_i),$$

with $\hat{f}(x) = (2m)^{-1} \sum_{j=1}^{m} \{k_{\tau_1}(x - X_j) + k_{\tau_1}(x + X_j)\}$ for suitable bandwidth τ_1.

To evaluate $I = E(\bar{U} \sum a_{1j}(\psi_1(Z_j) - E\psi_1(Z_j)))/\int J\Psi_1$ to first order, we note that the leading term in the expansion of $\bar{U}$ is $-(W - 2\int f^2)/(2\int f^2)$, whereas $m^{-1} \sum_{i=1}^{m} a_{1j}\psi_1(Z_j) = (m+1)^{-1} \sum_{j=1}^{m}(j/m)\psi_1(H_m^{-1}(j/m)) = (m+1)^{-1} \cdot \sum_{i=1}^{m} H_m(X_i)\psi_1(X_i)$. A straightforward but tedious computation shows that $\text{Covar}(W, \sum_{i=1}^{m} H_m(X_i)\psi_1(X_i)) = -12\int f^3 + 16(\int f^2)^2$ to first order. Combination of these facts leads to

$$(4.6) \qquad I = 4 - 3\int f^3 / \left(\int f^2\right)^2,$$

which quantity can be estimated using W from (4.1) and V_1 from (4.5).

It remains to deal with the terms involving the b's in (2.9). As concerns b_0, we already saw that $\int J(3\Psi_1^3 - 6\Psi_1\Psi_2 + \Psi_3)$ can be estimated through (3.8), which

reduces here to $-4m^{-1}\sum_{i=1}^{m}\hat{h}''(|X_i|) = -8V_0$ (cf. (4.3) and (4.4)). In b_1, we encounter $\tilde{I} = \int J^2\Psi_1^2 - \int\int J(s)\Psi_1'(s)J(t)\Psi_1'(t)(s\wedge t - st)dsdt$. For $J(t) = t$, the latter integral reduces after some tedious steps to $\int J^2\Psi_1^2 + 8\int f^3 - 16(\int f^2)^2$. Hence $\tilde{I}$ boils down to $16(\int f^2)^2 - 8\int f^3$, which can be estimated using W and V_1. Note that the specialization to Wilcoxon scores indeed heavily reduces matters! For b_2, we need $\int J^3\Psi_1$ which can be rewritten to $6\int H^2 f dF$, to be estimated by $6V_2$, where (cf. (3.4))

$$(4.7) \qquad V_2 = \frac{1}{m}\sum_{i=1}^{m}H_m^2(X_i)\hat{f}(X_i),$$

with $\hat{f}$ as in (4.5), but again with suitable bandwidth τ_2. As b_3 reduces to a constant, it remains to consider $\tilde{b}_{0,m}$. There we encounter $\sum_{j=1}^{m}\mathrm{Cov}_j$, which to first order equals $\int_{1/N}^{1-1/N}J'(t)\Psi_1'(t)t(1-t)dt$, which can be rewritten to $2f(0) - 4\int f^2$. Obviously, $\hat{f}(0)$ and W suffice for estimation. To summarize, after straightforward computation we arrive at

$$(4.8) \qquad \hat{b}_0 = -\frac{4}{9}\frac{V_0}{W^3}, \qquad \hat{b}_1 = 2 - \frac{4V_1}{W^2}, \qquad \hat{b}_2 = \frac{6V_2}{W},$$
$$\hat{b}_3 = \frac{3}{20} \qquad \text{and} \qquad \hat{\tilde{b}}_{0,\cdot} = \frac{7}{2} - \frac{4\hat{f}(0)}{W},$$

where W and V_k, $k = 0, 1, 2$ are given by (4.1), (4.4), (4.5) and (4.7) and, for a suitable bandwidth τ_3,

$$(4.9) \qquad \hat{f}(0) = \frac{1}{m}\sum_{i=1}^{m}k_{\tau_3}(X_i).$$

5. Wilcoxon scores: sample size determination

In Section 3 we observed that the total sample size N required to ensure that our two-stage test realizes $\pi(\kappa m^{-1/2}) = \pi_1$ to $o(m^{-1})$ for given κ and π_1, has the form

$$(5.1) \qquad N = \max\left(m, \left[N_1 + \hat{f}_r + \frac{1}{2}\right]\right).$$

Here N_1 is as given in (3.3), while $\hat{f}_r$ is a consistent estimator for the correction f_r, selected through (2.8)–(2.10) such that all lower order terms cancel. The step from f_r to $\hat{f}_r$ turned out in general to involve rather complicated estimators, valid under unduly restrictive conditions. Therefore we specialized in Section 4 to the case $J(t) = t$ and obtained explicit and relatively simple expressions for all estimators required.

What remains now is to use these explicit results from the previous section to evaluate $\hat{f}_r$ explicitly as well. However, this step completely parallels Theorem 3.1

in Albers (1992). Hence we omit this derivation and just present the resulting N. In fact, we have

$$(5.2) \qquad N = \max\left(m, \left[\frac{m(u_\alpha - u_\pi)^2}{3\kappa^2(W^*)^2} + \hat{f}_r^* + \frac{1}{2} \right] \right),$$

where

$$(5.3) \quad W^* = \left\{ W - \left(2\theta^2 + \frac{\tau^2}{3} \right) V_0 \right\} \left\{ 1 - \frac{2}{m}\left(\frac{4V_1}{W^2} - 1 \right)(u_\pi^2 - u_\alpha u_\pi - 2) \right\},$$

$$\hat{f}_r^* = -\frac{8}{9}(u_\alpha - u_\pi)^2 \frac{V_0}{W^3} + 4(u_\alpha - u_\pi)u_\pi \left(1 - \frac{2V_1}{W^2} \right)$$

$$+ 12(u_\pi^2 - 1)\frac{V_2}{W} - \frac{3}{10}(u_\alpha^2 + u_\alpha u_\pi + u_\pi^2 - 3)$$

$$+ \left(7 - \frac{8\hat{f}(0)}{W} \right) + \left(8 - \frac{24V_1}{W^2} \right),$$

in which $\theta = \kappa m^{-\frac{1}{2}}$ for the given κ and where W, $\hat{f}(0)$ and V_k, $k = 0, 1, 2$ are given by (4.1), (4.4), (4.5), (4.7) and (4.9).

Some explanatory remarks are the following. The term involving V_0 in (5.3) corrects for the bias, whereas the one involving V_1 corrects for the variance. Obviously, W^* in (5.2) can be replaced by W if we also replace $\hat{f}_r^*$ by $\hat{f}_r$, which is $\hat{f}_r^*$ supplemented with two terms to correct for bias and variance. The form involving W^* and $\hat{f}_r^*$ is given, as it seems easier to interpret. As concerns (5.4), we remark that half of the term $(7 - 8\hat{f}(0)/W)$ is the additional penalty due to splitting (cf. (2.2)), while $(8 - 24V_1/W^2)$ represents the interaction term occurring at the end of (2.9). As an example, consider the logistic df $F(x) = 1/(1 + e^{-x})$. Then

$$(5.5) \qquad EW = \frac{1}{3}, \quad EV_2 = EV_1 = -EV_0 = \frac{1}{30}, \quad f(0) = \frac{1}{4}.$$

It follows that $\hat{f}_r^*$ estimates $f_r^* = (5u_\alpha^2 - 3u_\alpha u_\pi + u_\pi^2 + 2)/10 + 1/2 + 4/5$, where we have kept separate $1/2$ due to splitting and $4/5$ due to the interaction term. Typically, f_r^* will attain values between 3 and 5. Comparison of (5.5) to (5.3) will also give some impression of the effect of the bias and variance corrections.

To conclude the paper, we present some simulation results. As underlying df's we shall use normal mixtures

$$(5.6) \qquad F(x) = (1 - \gamma)\Phi(x/\sigma_1) + \gamma\Phi(x/\sigma_2).$$

It is easily verified that for given $\rho = \sigma_2/\sigma_1$ the kurtosis in the family given by (5.6) is maximized for $\gamma = 1/(1 + \rho^2)$. Hence F is made as non-normal as possible in this respect for such γ. We shall consider the values $\rho = 1$ (i.e. $F = \Phi$), 2 and 3 here, leading to mixtures with $\gamma = .5$, .2 and .1, respectively. For σ_1 we choose 1, .8 and .75, respectively, leading to values of $\sigma^2 = (1 - \gamma)\sigma_1^2 + \gamma\sigma_2^2$ of 1, 1.02 and 1.01, respectively. This is by no means necessary, but it facilitates mutual comparison.

Hence we shall consider for $\rho = 1, 2, 3$ the triples $(\gamma, \sigma_1, \sigma_2) = (.5, 1, 1)$, $(.2, .8, 1.6)$ and $(.1, .75, 2.25)$, respectively in (5.6).

The initial sample sizes are $m = 10, 20, 30$ and 40, while the levels are $\alpha = .05$, .025 and .01. The required powers π_1 will range from .6 to .9 and the alternatives $\theta = \kappa m^{-1/2}$ are chosen such that $\bar{N}/m$ is at least about 2, where $\bar{N}$ is the average total sample size in a simulation run. For each $(m, \alpha, \pi_1, \theta, \rho)$ we use 10^4 simulations. In a simulation step, a sample of size m is drawn from (5.6) for the $(\gamma, \sigma_1, \sigma_2)$ under consideration. It is shifted over θ, upon which T_1 from (2.1) and N from (5.2) are computed. For the latter we need $W, V_k, k = 0, 1, 2$ and $\hat{f}(0)$ for suitable bandwidths τ and τ_k, $k = 0, 1, 2, 3$, respectively. On the choice of τ we have already commented extensively in Section 4 and we shall use τ/θ between 1.5 and 2, as advocated there. Comparable considerations can be given for the τ_k. Together with some experimentation in practice, this leads to the rule of thumb

$$(5.7) \qquad\qquad \tau_0 = \tau_2 = 2\tau, \qquad \tau_1 = \tau_3 = \tau,$$

which we used throughout the simulations.

The resulting N leads to an additional sample on which T_2 is based. Finally, we reject if $T = T_1 + T_2$ (cf. (2.2)) exceeds the critical value $N/4 + ((N-1)/3)^{1/2}\{u_\alpha - 3(u_\alpha^3 - 3u_\alpha)/(20N)\}/2$ (cf. (2.11)), thus leading in the end to a realized power $\hat{\pi}$. To check the performance of the procedure under H_0, each simulation run is repeated for $\theta = 0$, using the same starting point of the random generator, thus leading to the realized level $\hat{\alpha}$ as well. The results are collected in Table 1.

Inspection of Table 1 shows that the two-stage procedure behaves reasonably well. The realized levels are quite satisfactory: $\mathrm{ave}(\hat{\alpha} - \alpha)$, the average difference between $\hat{\alpha}$ and α, equals .08, $-.01$, $-.06$ and $-.06\%$ for $m = 10, 20, 30$ and 40, respectively. As concerns the power requirement, the procedure is slightly conservative, which effect decreases as m gets larger. In fact, the $\mathrm{ave}(\hat{\pi} - \pi_1)$ equals 2.2, 1.8, 0.8 and 0.4% for $m = 10, 20, 30$ and 40, respectively. The convergence is somewhat slower than in the parametric case considered in Albers (1992). This is according to expectation, as the density estimation approach is more ambitious. It covers a much wider range of df's, but requires more estimation. Due to the combined variation in the estimators involved, a small fraction of the simulations will produce values of $N_1 + f_r + \frac{1}{2}$ which fall below m, after which (5.1) leads to $N = m$ after all. The resulting positive bias is $o(m^{-1})$ asymptotically, but especially for $m = 10$ and 20 still noticeable in our simulation results. Indeed, closer inspection of these results reveals that $\hat{\pi} - \pi_1$ is heavily correlated with the number of times the second sample is empty during a run. A final remark about Table 1 is that $\bar{N}_\alpha$, the average sample size under $\theta = 0$, is clearly smaller than $\bar{N}_\pi$, the corresponding value under $\theta = \kappa m^{-1/2}$. To explain this phenomenon, we observe that the bias correction term for W involving V_0 in (5.3) is typically positive, as $EV_0 = -\int(f')^2 < 0$. Hence its effect on N from (5.2) is a decrease. But under H_0 there is no bias to correct for as far as θ is concerned and therefore N turns out lower than under $\theta = \kappa m^{-1/2}$.

Next we turn to a slight modification of the procedure above. As was mentioned after (2.2), it would be marginally better to use a single Wilcoxon statistic based on the total sample, rather than the combined statistic $T = T_1 + T_2$. The

Table 1.   The realized power $\hat{\pi}$ and realized level $\hat{\alpha}$ for the Wilcoxon two-stage procedure using density estimation, under normal mixtures from (5.6). We have: m=sample size, θ = shift, τ = bandwidth, $\rho = \sigma_2/\sigma_1$, α = level, π_1 = prescribed power and $\bar{N}_\alpha$ ($\bar{N}_\pi$) = realized average total sample size under $\theta = 0$ ($\theta = \kappa m^{-1/2}$). Each run for $\hat{\pi}$ consists of 10^4 simulations.

m	θ	τ	ρ	α	$\hat{\alpha}$	π_1	$\hat{\pi}_1$	$\bar{N}_\alpha$	$\bar{N}_\pi$
10	.5	1.0	1	.05	.0517	.750	.773	27.0	33.8
10	.4	.8	2	.05	.0504	.600	.635	25.0	29.8
10	.4	.6	3	.025	.0265	.700	.725	36.9	48.7
10	.6	.9	2	.025	.0255	.900	.906	35.5	55.0
10	.6	1.0	1	.01	.0104	.800	.817	34.3	48.4
10	.4	.8	3	.01	.0102	.700	.725	46.3	59.1
20	.5	1.0	3	.05	.0480	.900	.924	29.1	40.6
20	.4	.8	2	.05	.0502	.750	.778	33.6	40.3
20	.5	1.0	1	.025	.0255	.800	.814	35.3	44.3
20	.3	.6	3	.025	.0246	.700	.716	55.8	63.5
20	.6	1.0	1	.01	.0105	.900	.903	39.3	55.9
20	.4	.8	2	.01	.0105	.750	.777	55.5	67.1
30	.3	.5	3	.05	.0491	.800	.808	53.5	61.5
30	.4	.8	2	.05	.0497	.900	.908	51.5	62.7
30	.4	.6	1	.025	.0240	.800	.803	52.7	62.1
30	.3	.6	3	.025	.0218	.700	.717	53.1	61.1
30	.5	.8	1	.01	.0110	.900	.897	55.4	70.9
30	.4	.6	2	.01	.0107	.750	.763	52.0	63.5
40	.25	.5	1	.05	.0492	.700	.703	81.0	86.1
40	.3	.5	3	.05	.0493	.900	.897	73.1	84.4
40	.3	.6	2	.025	.0234	.750	.758	71.8	80.0
40	.35	.6	3	.025	.0251	.900	.908	65.3	79.5
40	.4	.6	2	.01	.0105	.900	.904	74.5	92.2
40	.3	.6	1	.01	.0095	.700	.704	96.1	105.2

reason we nevertheless preferred the latter approach was its better asymptotic tractability. However, it is interesting to investigate the other possibility empirically. Hence, we repeat the simulations leading to Table 1, but now we add the additional sample of size $N - m$ to the first one and compute the full Wilcoxon statistic T_F based on the sample of size N. Only two minor adjustments are required. As no splitting occurs, the corresponding penalty (cf. the discussion following (5.4)) can be lifted, which means that $(7 - 8\hat{f}(0)/W)$ in (5.4) is replaced by half this amount. In addition, as $\sum_{j=1}^{m}(j/(N+1))^2 + \sum_{j=1}^{N-m}(j/(N+1))^2 = (N-1)/3 + O(m^{-1})$ is replaced by $\sum_{j=1}^{N}(j/(N+1))^2 = N/3 - 1/6 + O(m^{-1})$, the corresponding replacement of $(N-1)/3$ by $(N/3 - 1/6)$ should be made in the critical value used above. Using the same simulation runs as in Table 1, we obtain realized powers $\hat{\pi}_F$ and realized levels $\hat{\alpha}_F$ as collected in Table 2.

In comparing the contents of Table 2 to those of Table 1, the following points

Table 2. Results as in Table 1, but now for the full statistic T_F, rather than for the combined statistic $T = T_1 + T_2$ from (2.2).

m	θ	τ	ρ	α	$\hat{\alpha}_F$	π_1	$\hat{\pi}_F$	$\bar{N}_{\alpha,F}$	$\bar{N}_{\pi,F}$
10	.5	1.0	1	.05	.0491	.750	.759	25.8	33.0
10	.4	.8	2	.05	.0472	.600	.621	24.6	29.5
10	.4	.6	3	.025	.0258	.700	.716	38.1	48.2
10	.6	.9	2	.025	.0229	.900	.897	35.2	54.3
10	.6	1.0	1	.01	.0083	.800	.816	33.3	48.0
10	.4	.8	3	.01	.0095	.700	.727	45.8	58.4
20	.5	1.0	3	.05	.0367	.900	.876	28.1	39.6
20	.4	.8	2	.05	.0433	.750	.744	32.6	39.5
20	.5	1.0	1	.025	.0227	.800	.804	34.4	43.7
20	.3	.6	3	.025	.0246	.700	.709	54.9	63.3
20	.6	1.0	1	.01	.0096	.900	.898	38.5	55.0
20	.4	.8	2	.01	.0100	.750	.767	54.4	66.4
30	.3	.5	3	.05	.0455	.800	.791	52.7	60.8
30	.4	.8	2	.05	.0449	.900	.892	51.1	61.8
30	.4	.6	1	.025	.0231	.800	.795	52.0	61.3
30	.3	.6	3	.025	.0222	.700	.712	52.3	60.5
30	.5	.8	1	.01	.0096	.900	.892	54.0	70.0
30	.4	.6	2	.01	.0095	.750	.751	51.2	62.9
40	.25	.5	1	.05	.0504	.700	.697	80.1	85.5
40	.3	.5	3	.05	.0476	.900	.886	72.3	83.7
40	.3	.6	2	.025	.0230	.750	.746	71.2	79.7
40	.35	.6	3	.025	.0226	.900	.896	64.8	78.8
40	.4	.6	2	.01	.0106	.900	.898	73.6	91.0
40	.3	.6	1	.01	.0088	.700	.702	95.9	104.7

can be noted. ave($\hat{\alpha}_F - \alpha$) equals $-.12$, $-.39$, $-.25$ and $-.12\%$ for $m = 10, 20,$ 30 and 40 respectively. Hence $\hat{\alpha}_F$ seems to be less close to α than $\hat{\alpha}$. On the other hand, the error typically occurs on the safe side, i.e. the resulting procedure is conservative. As concerns the power requirement, we obtain that ave($\hat{\pi}_F - \pi_1$) equals 1.4, .0, $-.3$ and $-.5\%$ for $m = 10, 20, 30$ and 40, respectively. Hence $\hat{\pi}$ in its turn is closer to π_1 than $\hat{\pi}$. Furthermore, ave($\bar{N}_\pi - \bar{N}_{\pi,F}) = 0.7$, which agrees well with e.g. the value $7/2 - 2\sqrt{2} = 0.672$ of $\frac{1}{2}(7 - 8f(0)/EW)$ in the normal case ($\rho = 1$). All in all, the full statistic procedure provides an interesting competitor to the original splitted proposal. In addition, it is more simple conceptually, as splitting is avoided.

At the end, let us return to the starting point: Stein's two-stage procedure. This procedure is of course applicable for $\rho = 1$. Moreover, it is known to be robust to certain forms of non-normality, see e.g. Ramkaran (1983). Hence it also makes sense to investigate its behaviour for $\rho = 2$ and $\rho = 3$. From Albers (1992),

p. 350, we obtain that c^{-1} in (2.2) satisfies to $o(1)$

$$(5.8) \qquad c^{-1} = \frac{m(u_\alpha - u_\pi)^2}{\kappa^2} \left(1 + \frac{(1 + u_\alpha^2 + u_\alpha u_\pi + u_\pi^2)}{2m} \right).$$

Using once more the same simulation runs as in Table 1, we collect realized powers $\hat{\pi}_{St}$ and realized levels $\hat{\alpha}_{St}$ for Stein's two-stage procedure in Table 3.

Table 3.   Results as in Table 1, but now for Stein's two-stage procedure.

m	θ	ρ	α	$\hat{\alpha}_{St}$	π_1	$\hat{\pi}_{St}$	$\bar{N}_{St}$
10	.5	1	.05	.0523	.750	.762	25.2
10	.4	2	.05	.0528	.600	.613	27.3
10	.4	3	.025	.0295	.700	.701	47.8
10	.6	2	.025	.0289	.900	.907	36.5
10	.6	1	.01	.0116	.800	.797	35.7
10	.4	3	.01	.0166	.700	.686	65.4
20	.5	3	.05	.0481	.900	.918	38.0
20	.4	2	.05	.0507	.750	.763	37.5
20	.5	1	.025	.0233	.800	.803	35.1
20	.3	3	.025	.0286	.700	.708	77.8
20	.6	1	.01	.0113	.900	.906	41.2
20	.4	2	.01	.0118	.750	.749	65.8
30	.3	3	.05	.0522	.800	.800	73.6
30	.4	2	.05	.0503	.900	.903	58.1
30	.4	1	.025	.0257	.800	.811	52.8
30	.3	3	.025	.0255	.700	.699	74.6
30	.5	1	.01	.0111	.900	.903	56.9
30	.4	2	.01	.0086	.750	.755	63.7
40	.25	1	.05	.0500	.700	.707	78.8
40	.3	3	.05	.0546	.900	.905	100.9
40	.3	2	.025	.0231	.750	.751	83.6
40	.35	3	.025	.0237	.900	.901	91.8
40	.4	2	.01	.0091	.900	.894	89.1
40	.3	1	.01	.0098	.700	.705	97.1

A summary as before in this case produces that $\text{ave}(\hat{\alpha}_{St} - \alpha)$ equals .35, .06, .06, .0% for $m = 10, 20, 30$ and 40, respectively. Hence, for smaller m, the procedure can be markedly anti-conservative. To investigate this phenomenon a bit further we introduce $t_{\alpha,F}$, the critical value of Student's test for underlying distribution F. From Albers *et al.* ((1976), p. 135, eq. (5.23)) it follows that, for sample size m,

$$(5.9) \qquad t_{\alpha,F} = u_\alpha + \frac{u_\alpha^3 - u_\alpha}{2m} - \frac{(\kappa_{4,F} + 3)}{12m}(u_\alpha^3 - 3u_\alpha) + o(m^{-1}),$$

where $\kappa_{4,F} = \int_{-\infty}^{\infty} x^4 dF(x)/(\int_{-\infty}^{\infty} x^2 dF(x))^2 - 3$, the fourth cumulant of F. This result in its turn readily implies that the error $\tilde{\alpha} - \alpha$ in α, due to using the normal critical value under nonnormal F, to first order equals

$$(5.10) \qquad \tilde{\alpha} - \alpha = \frac{\kappa_{4,F}}{12m}(u_\alpha^3 - 3u_\alpha)\phi(u_\alpha).$$

For the normal mixtures under consideration, $\kappa_{4,F}$ equals $3(\rho^2 - 1)^2/(4\rho^2)$. Together with (5.10) this e.g. produces for $m = 10$, $\alpha = 0.01$ and $\rho = 3$ a value 0.67% for $\tilde{\alpha} - \alpha$. This indicates that the value 0.0166 for $\hat{\alpha}_{St}$ is Table 3 for this configuration of m-, α- and ρ-values, is not an outlier. In fact, for this particular triple we performed 8 additional simulation runs, using varying π_1 and θ, resulting in values of $\hat{\alpha}_{St}$ which varied from .0134 to .0193, with an average of .0154.

As concerns the power requirement, we obtain that $\text{ave}(\hat{\pi}_{St} - \pi_1)$ equals .3, .8, .4 and .2% for $m = 10$, 20, 30 and 40, respectively. Hence in this respect, Stein's procedure performs very well. As far as power and level are concerned, the nonparametric procedures above are preferable mainly in the sense that they provide better control of the level, as should be the case.

The remaining aspect for comparison is the average sample size. As S_m^2 is translation invariant, we deal with a single value $\bar{N}_{St}$. Roughly speaking, the picture is as follows: for $\rho = 1$, we have that $\bar{N}_{St}$ is comparable to $\bar{N}_\alpha$ and substantially lower than $\bar{N}_\pi$, while for $\rho = 2$ we find that $\bar{N}_{St}$ is comparable to $\bar{N}_\pi$ and substantially larger than $\bar{N}_\alpha$. For $\rho = 3$, we even see that $\bar{N}_{St}$ is substantially larger than $\bar{N}_\pi$. An obvious explanation for this behavior is obtained by looking at the ARE of the Wilcoxon test with respect to the t-test under the normal mixtures from (5.6). For $\gamma = 1/(1 + \rho^2)$, this ARE equals

$$(5.11) \qquad \frac{6}{\pi}\{(1 - \gamma)^{5/2} + \gamma^{5/2} + [2\gamma(1 - \gamma)]^{3/2}\}^2,$$

which attains the values .955 ($=3/\pi$), 1.136 and 1.373 for $\rho = 1$, 2 and 3, respectively.

Summarizing, the simulation results confirm that under normality (i.e. for $\rho = 1$) Stein's two-stage procedure is best, but also strongly suggest that the picture changes as ρ increases. For $\rho = 3$, the nonparametric procedures are to be preferred, both with respect to control of level as to sample sizes required.

Acknowledgements

References

Albers, W. (1985). An exact test for scale under normality using Helmert's transformation, *J. Statist. Plann. Inference*, **12**, 331–346.

Albers, W. (1991). Second order analysis of two-stage rank tests for the one-sample problem, *Ann. Statist.*, **19**, 1042–1052.

Albers, W. (1992). Asymptotic expansions for two-stage rank tests, *Ann. Inst. Statist. Math.*, **44**, 335–356.

Albers, W., Bickel, P. J. and van Zwet, W. R. (1976). Asymptotic expansions for the power of distribution free tests in the one-sample problem, *Ann. Statist.*, **4**, 108–156.

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, *Ann. Statist.*, **19**, 633–667.

Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation, *Ann. Statist.*, **20**, 1306–1328.

Hall, P. and Marron, J. (1987). Estimation of integrated squared density derivatives, *Statist. Probab. Lett.*, **6**, 109–115.

Hall, P. and Sheather, S. J. (1988). On the distribution of a studentized quantile, *J. Roy. Statist. Soc. Ser. B*, **50**, 381–391.

Jones, M. C. (1992). Estimating densities, quantiles, quantile densities and density quantiles, *Ann. Inst. Statist. Math.*, **44**, 721–729.

Jones, M. C. and Sheather, S. J. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statist. Probab. Lett.*, **11**, 511–514.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed., Wiley, New York.

Ramkaran (1983). The robustness of Stein's two-stage procedure, *Ann. Statist.*, **11**, 1251–1256.

Ritov, Y. and Bickel, P. J. (1990). Achieving information bounds in non and semiparametric models, *Ann. Statist.*, **18**, 925–938.

Schweder, T. (1975). Window estimation of the asymptotic variance of rank estimators of location, *Scand. J. Statist.*, **2**, 113–126.

Schweder, T. (1981). Correction note, *Scand. J. Statist.*, **8**, p. 55.

Sen, P. K. and Ghosh, M. (1971). On bounded length sequential confidence intervals based on one sample rank order statistics, *Ann. Math. Statist.*, **42**, 189–203.

Sen, P. K. and Ghosh, M. (1974). Sequential rank tests for location, *Ann. Statist.*, **2**, 540–552.

Sheather, S. J., Hettmansperger, T. P. and Donald, M. R. (1992). Data-based bandwidth selection for kernel estimators of the integral of $f^2(x)$ (preprint).