Statistica Sinica Preprint No: SS-14-155R1	
Title	On some exact distribution-free one-sample tests for high dimension low sample size data
Manuscript ID	SS-14-155R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.2014.155
Complete List of Authors	Munmun Biswas, Minerva Mukhopadhyay and Anil K. Ghosh
Corresponding Author	Anil Ghosh
E-mail	akghosh@isical.ac.in
Notice: Accepted version subject to English editing.	

ON SOME EXACT DISTRIBUTION-FREE ONE-SAMPLE TESTS FOR HIGH DIMENSION LOW SAMPLE SIZE DATA

Munmun Biswas, Minerva Mukhopadhyay and Anil K. Ghosh

Indian Statistical Institute, Kolkata

Abstract: Several rank based tests for the multivariate one-sample problem are available in the literature. But, unlike univariate rank based tests, most of these multivariate tests are not distribution-free. Moreover, many of them are not applicable when the dimension of the data exceeds the sample size. In this article, we develop and investigate some distribution-free tests for the one-sample location problem, which can be conveniently used in high dimension low sample size (HDLSS) situations. Under some appropriate regularity conditions, we prove the consistency of these tests when the sample size remains fixed and the dimension grows to infinity. Some simulated and real data sets are analyzed to compare their performance with some popular one-sample tests.

Key words and phrases: HDLSS data, linear rank tests, run tests, shortest covering path, weak law of large numbers.

1. Introduction

Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n independent realizations of a d-dimensional random vector $\mathbf{X}$ having a continuous distribution F symmetric about $\boldsymbol{\theta} \in \mathbb{R}^d$ (i.e., $\mathbf{X} - \boldsymbol{\theta} \stackrel{d}{=} \boldsymbol{\theta} - \mathbf{X}$). In the one-sample problem, we test the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ is a pre-specified point in $\mathbb{R}^d$. This problem is well investigated in the literature, especially when $d = 1$. If F is assumed to be normal, one uses the student's t-statistic to perform the test. In general univariate set up, we use nonparametric tests like those based on linear rank statistics (see e.g., Gibbons and Chakraborty, 2003). These tests are distribution-free, and they outperform the t-test for a wide variety of non-Gaussian distributions. Several attempts have been made to generalize these rank based tests for multivariate data. Puri and Sen (1971) proposed tests based coordinate-wise signs and ranks. Randles (1989, 2000) developed one-sample location tests based on the idea of interdirections. Chaudhuri and Sengupta (1993)

generalized Hodges' (1955) bivariate sign test to higher dimension. Other nonparametric tests for the multivariate one-sample problem include Bickel (1965), Hettmansperger et al. (1994), Mottonen et al. (1997), Hettmansperger et al. (1997), Chakraborty et al. (1998) and Hallin and Paindaveine (2002). Some of these multivariate nonparametric tests are distribution-free under some specific types of symmetry conditions on F (e.g., the sign test of Puri and Sen (1971) is distribution-free under coordinate-wise sign symmetry, the sign tests of Randles (1989, 2000) and Chaudhuri and Sengupta (1999) are distribution-free under elliptic symmetry), but none of them are distribution-free when F is centrally symmetric (i.e., $\mathbf{X} - \boldsymbol{\theta} \stackrel{d}{=} \boldsymbol{\theta} - \mathbf{X}$ for some $\boldsymbol{\theta} \in \mathbb{R}^d$). In such cases, one either uses the test based on the large sample distribution of the test statistic or the conditional test based on the permutation principle. Moreover, most of these tests become computationally prohibitive even for moderately high dimensional data, and they usually yield poor performance in high dimensional problems. None of these tests can be used when the dimension exceeds the sample size.

In the recent past, several one-sample tests have been proposed for high dimension low sample size (HDLSS) data (see e.g., Bai and Saranadasa, 1996; Srivastava and Du, 2008; Srivastava, 2009; Chen and Qin, 2010; Park and Ayyala, 2013). However, these tests are concerned with the mean vector of a high dimensional distribution, and none of them are distribution-free in finite sample situations. These tests are based on the asymptotic distribution of the test statistic, where the dimension increases with the sample size.

In this article, we develop a general method for multivariate generalizations of univariate rank based one-sample tests retaining their distribution-free property. These tests are fairly simple and computationally efficient. They can be conveniently used even when the dimension of the data exceeds the sample size.

2. Construction of distribution-free tests for multivariate data

Let $X_1, \ldots, X_n$ be independent and identically distributed univariate random variables with a distribution F, which is continuous and symmetric about some $\theta_0 \in \mathbb{R}$. If we define $Y_i = sign(X_i - \theta_0)$ and R_i as the rank of $|X_i - \theta_0|$ in $\{|X_1 - \theta_0|, \ldots, |X_n - \theta_0|\}$ for all $i = 1, 2, \ldots, n$, it is easy to check that

(*a*) $P\{(Y_1,\ldots,Y_n)=(y_1,\ldots,y_n)\}=2^{-n}$ for all $(y_1,\ldots,y_n)\in\{-1,1\}^n$,
(*b*) $P\{(R_1,\ldots,R_n)=(r_1,\ldots,r_n)\}=1/n!$ for all permutations $(r_1,\ldots,r_n)$ of $\{1,\ldots,n\}$,
(*c*) $(Y_1,\ldots,Y_n)$ and $(R_1,\ldots,R_n)$ are independent.

So, if the test statistic is a function of $(Y_1,\ldots,Y_n)$ and $(R_1,\ldots,R_n)$ (e.g., linear rank statistic), the resulting test will be distribution-free in finite sample situations. To construct distribution-free one-sample tests for multivariate data, we extend the notions of signs $Y_1,\ldots,Y_n$ and ranks $R_1,\ldots,R_n$ in such a way that the results (a)-(c) hold under H_0.

For testing $H_0:\boldsymbol{\theta}=\boldsymbol{\theta}_0$ ($\boldsymbol{\theta}_0\in R^d$) against $H_1:\boldsymbol{\theta}\neq\boldsymbol{\theta}_0$ based on n independent multivariate observations $\mathbf{x}_1,\ldots,\mathbf{x}_n$ from F, we define $\mathbf{x}_i^*=2\boldsymbol{\theta}_0-\mathbf{x}_i$ for $i=1,\ldots,n$. Note that under H_0, $\mathbf{x}_1,\ldots,\mathbf{x}_n$ and $\mathbf{x}_1^*,\ldots,\mathbf{x}_n^*$ have the same distribution, while under H_1 they differ in their locations. Now, consider a complete graph $\mathcal{K}_{2n}$ on $2n$ vertices $\mathbf{z}_1,\ldots,\mathbf{z}_{2n}$, where $\mathbf{z}_i=\mathbf{x}_i$ and $\mathbf{z}_{n+i}=\mathbf{x}_i^*$ for $i=1,\ldots,n$. Also, assume that each edge of $\mathcal{K}_{2n}$ has a cost associated with it. For instance, the Euclidean distance between the two vertices of an edge can be considered as its cost. Now, consider a path $\mathcal{P}$ of length $n-1$ in $\mathcal{K}_{2n}$ such that for every $i=1,\ldots,n$, $\mathcal{P}$ covers either $\mathbf{x}_i$ or $\mathbf{x}_i^*$. Clearly, there are $2^n n!$ such paths in $\mathcal{K}_{2n}$. However, for every path, there is another path in the reverse order. Again, for any path and its reverse path, two other equivalent paths can be obtained if we replace all $\mathbf{z}_i$'s by $\mathbf{z}_{n+i}$ (or $\mathbf{z}_{i-n}$ if $i>n$). For these four paths, the total cost of the $n-1$ edges remains the same. If we consider these four equivalent paths as the same path, the number of distinct covering paths (i.e., the paths that cover either $\mathbf{x}_i$ or $\mathbf{x}_i^*$ for all $i=1,\ldots,n$) reduces to $2^{n-2}n!$. For each of these distinct covering paths, the sum of the costs corresponding to its $n-1$ edges is defined as its cost. Among these distinct paths, we choose the one having the minimum cost, and we call it the shortest covering path $\mathcal{P}_0$. This shortest covering path may not be unique, but if the costs corresponding to different edges come from continuous distributions, it becomes unique with probability one.

Figure 1 shows a complete graph on $2n=6$ vertices in two-dimension along with the costs corresponding to different edges. There are 12 distinct covering paths in this graph, where the path $\mathbf{z}_1\to\mathbf{z}_3\to\mathbf{z}_5$ (or $\mathbf{z}_5\to\mathbf{z}_3\to\mathbf{z}_1$, or equivalently, $\mathbf{z}_4\to\mathbf{z}_6\to\mathbf{z}_2$ or $\mathbf{z}_2\to\mathbf{z}_6\to\mathbf{z}_4$) is the shortest covering path.

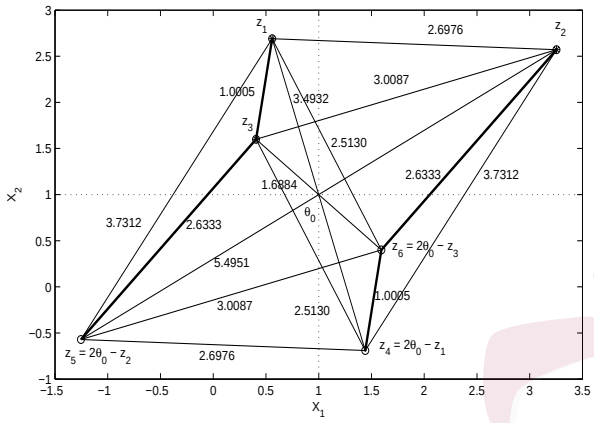

Figure 1: A complete graph on $2n = 6$ vertices and the shortest covering path.

We define $Y_1, \ldots, Y_n$ and $R_1, \ldots, R_n$ along $\mathcal{P}_0$. For each $i = 1, \ldots, n$, Y_i takes the value 1 (or -1, respectively) if $\mathbf{x}_i$ (or $\mathbf{x}_i^*$, respectively) appears on $\mathcal{P}_0$, and R_i is defined as the position of $\mathbf{x}_i$ (or $\mathbf{x}_i^*$) along $\mathcal{P}_0$. Between the two terminal nodes of $\mathcal{P}_0$, as a starting point, we choose the one which is closer to $\boldsymbol{\theta}_0$. Since $\mathbf{X}$ and $2\boldsymbol{\theta}_0 - \mathbf{X}$ has the same distribution under H_0, and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ form an exchangeable collection, it is easy to check that $Y_1, \ldots, Y_n$ and $R_1, \ldots, R_n$ defined in this way satisfy properties (a)-(c) mentioned earlier. So, if we construct a test statistic, which is a function of $Y_1, \ldots, Y_n$ and $R_1, \ldots, R_n$, the resulting test will be distribution-free. Like the univariate case, we can use the linear rank statistic of the form $T_0 = \sum_{i=1}^n I\{Y_i = 1\} a(R_i)$, where $I\{\cdot\}$ is the indicator function, and $a : \{1, \ldots, n\} \to \mathbb{R}$ is a score function. Using $a(i) = 1$ and $a(i) = i$ for $i = 1, \ldots, n$, one obtains the sign statistic $\sum_{i=1}^n I\{Y_i = 1\}$ and the signed-rank statistic $\sum_{i=1}^n R_i \; I\{Y_i = 1\}$, respectively. Under H_0, since $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$ have the same distribution, we expect almost equal numbers of $\mathbf{x}_i$'s and $\mathbf{x}_i^*$'s on $\mathcal{P}_0$. But, under H_1, one would expect a dominance of either the $\mathbf{x}_i$'s or the $\mathbf{x}_i^*$s on $\mathcal{P}_0$. So, we should reject H_0 for very small or very large values of T_0, or in other words, H_0 is to be rejected for large values of $T_0^* = \max\{T_0, \sum_{i=1}^n a(i) - T_0\}$.

One can also construct a test based on the number of runs or that based on the length of the longest run along $\mathcal{P}_0$. The number of runs can be expressed as $T_1 = 1 + \sum_{i=1}^{n-1} U_i$, where U_i is an indicator variable that takes the value 1 if and only if the i-th edge of $\mathcal{P}_0$ connects two observations with different Y-values. The length of the longest run is given by $T_2 = \max_{0 \le i < j \le n} \; (j - i) \; I\{U_i = 1, U_{i+1} =$

$\ldots = U_{j-1} = 0, U_j = 1\}$, where $U_0 = U_n = 1$ and for $i = 1, \ldots, n-1$, the U_is are defined as above. Under H_0, when two data clouds $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and $\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*$ are well mixed, T_1 is expected to be large, while T_2 is expected to be small. But under H_1, when they are well separated, we expect small values of T_1 and large values of T_2. So, here we use one-sided cut-offs.

In Fig. 1, along the path $\mathbf{z}_1 = \mathbf{x}_1 \to \mathbf{z}_3 = \mathbf{x}_3 \to \mathbf{z}_5 = \mathbf{x}_2^*$, both T_1 and T_2 take the value 2, while the values of the sign statistic and the signed rank statistic are 2 and 3, respectively. Barring the signed rank statistic, the values of the other three do not depend on the choice of the starting point of $\mathcal{P}_0$, and they remain the same if the path is traversed in the reverse order. The values of T_1 and T_2 also remain the same along an equivalent path, where all $\mathbf{z}_i$'s are replaced by $\mathbf{z}_{n+i}$ (or $\mathbf{z}_{i-n}$ if $i > n$). Along this path, though T_0 becomes $\sum_{i=1}^{n} a(i) - T_0$, the value of T_0^* remains the same. Therefore, all these tests lead to the same results if this equivalent path is chosen. If we use pairwise distances among the observations as costs corresponding to different edges of $\mathcal{K}_{2n}$, the resulting tests become invariant under location change and homogeneous scale transformation of the data. If the Euclidean distance is used, they become rotation invariant as well, but unfortunately they do not have the maximal invariance property for any of these transformations. Note that in the univariate case, $\mathcal{P}_0$ is obtained by joining the observations x_i (or x_i^*, if $x_i < \theta_0$) in increasing order of the magnitudes of their differences from θ_0. So, in that case, T_0 coincides with the univariate linear rank statistic. Usually, we do not use run tests for the univariate one-sample location problem. But, alternative distribution-free tests for that problem can also be constructed using univariate analogs of T_1 and T_2.

2.1 Distributions of test statistics and determination of cut-off values

The test statistics constructed in the previous section are the same functions of $(Y_1, \ldots, Y_n)$ and $(R_1, \ldots, R_n)$ as their univariate analogs. Since the joint distribution of $(Y_1, \ldots, Y_n)$ and $(R_1, \ldots, R_n)$ under H_0 remains the same as in the univariate set up, irrespective of the underlying distribution F and the data dimension d, null distributions of these test statistics exactly match with that of their univariate counter parts. In that sense, the tests constructed in this way can be viewed as most natural multivariate generalizations of the univariate distribution-free tests, and statistical tables available for the univariate tests can

be used to determine the cut-offs in multivariate cases as well. However, unlike univariate rank based procedures, our multivariate rank based tests may not have semi-parametric optimality discussed in Hallin and Werker (2003).

Under H_0, T_0 is distributed as $\sum_{i=1}^{n} W_i$, where $P_{H_0}(W_i = 0) = P_{H_0}(W_i = a(i)) = 1/2$ for each $i = 1, \ldots, n$, and they are independent (see e.g., Gibbons and Chakraborti, 2003). One can check that under H_0, $T_1 - 1$ follows a binomial distribution with parameters $n-1$ and $1/2$. The null distribution of T_2 is given in Fu and Koutras (1994). For construction of a linear rank test with the nominal level α $(0 < \alpha < 1)$, we consider a test function of the form $\phi_0(t) = I\{t > c_0\} + \gamma_0 I\{t = c_0\}$, where c_0 and γ_0 $(0 \leq \gamma_1 < 1)$ are chosen in such a way that $E_{H_0}(\phi_0(T_0^*)) = \alpha$. For run tests, we reject H_0 when T_1 is small or T_2 is large. Because of the discrete nature of T_1 and T_2, here also we need randomization at cut-off points so that the sizes of these tests match the level of significance α.

However, if m and n are large, one can also use the test based on the asymptotic null distribution of the test statistic. Asymptotic normality of T_1 under H_0 is obtained using normal approximation to the binomial distribution, and that of T_0 can be shown using a central limit theorem for independent random variables $W_1, W_2, \ldots, W_n$ (see e.g., Gibbons and Chakraborti, 2003). The large sample distribution of T_2 can be found in Gordon et al. (1986).

2.2 Computation of test statistics

Unless the sample size is very small, finding $\mathcal{P}_0$ is computationally difficult. In fact, it is equivalent to the well-known travelling salesman's problem, which is NP-complete (see e.g., Garey and Johnson, 1979). However, one can use good heuristic search algorithms (see e.g., Lawler et al., 1985). In this article, we use a method based on Prim's (1957) algorithm, where the distance between two observations is used as the cost of the edge connecting them. First we select the pair $\mathbf{z}_i$ and $\mathbf{z}_j$ $(|j - i| \neq n)$ having the minimum distance between them and define a set $S = \{i, j\}$. We join $\mathbf{z}_i$ and $\mathbf{z}_j$ by an edge to get a path of unit length with $\mathbf{z}_i$ and $\mathbf{z}_j$ as its two ends. From each of these two ends, we calculate the distance of $\mathbf{z}_k$, where $k \notin S$ and $|k-l| \neq n$ for any $l \in S$. If the minimum of these distances is observed between $\mathbf{z}_i$ and $\mathbf{z}_r$, we join $\mathbf{z}_i$ and $\mathbf{z}_r$ to get a path of length 2 $(\mathbf{z}_j \to \mathbf{z}_i \to \mathbf{z}_r)$ with $\mathbf{z}_j$ and $\mathbf{z}_r$ as its two terminal nodes. We also update S by adding r to it. Next, we consider the distances of all $\mathbf{z}_k$ $(k \notin S$ and $|k - l| \neq n$

for any $l \in S$) from these two terminal nodes and choose a new edge in the same way to get a path of length 3. The set S is also updated by adding the index of the new selected node. We proceed in this way until a path of length $(n-1)$ is chosen. Clearly, this path contains either $\mathbf{x}_i$ or $\mathbf{x}_i^*$ for all $i = 1, \ldots, n$, and it is considered as the shortest covering path. Test statistics are computed using the signs and the ranks (as defined before) of the observations along this path. Though this path finding algorithm sometimes leads to a sub-optimal solution in terms of cost, the test statistic computed along this path often remains the same as that computed along the actual $\mathcal{P}_0$, especially in high dimensions. As a consequence, the resulting tests generally perform well for HDLSS data. We will discuss it in detail in the next section to make it more transparent.

3. Power properties of constructed tests in HDLSS set up

We study the power properties of our tests in HDLSS asymptotic regime, where the sample size remains fixed, and the dimension grows to infinity. Suppose that we have n independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ on a d-dimensional random vector $\mathbf{X} = \left(X^{(1)}, \ldots, X^{(d)}\right)$ following the distribution F, which has location $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(d)})$ and the scatter matrix $\boldsymbol{\Sigma}$. For our tests, we consider a cost function of the form $\rho_\psi^h(\mathbf{x}, \mathbf{x}') = h\left(\sum_{i=1}^d \psi(|x^{(i)} - x'^{(i)}|)\right)$, where $h : \mathbb{R}_+ \to \mathbb{R}_+$ and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ are continuous, monotonically increasing functions with $h(0) = \psi(0) = 0$ such that ρ_ψ^h is a distance in $\mathbb{R}^d$. Clearly, this class of distance functions include all l_p distances with $p \geq 1$. In order to investigate the power of the distribution-free tests based on ρ_ψ^h, we consider the following assumptions.
Let $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *be two independent copies of* $\mathbf{X}$*, which follows the distribution* F*. For* $\mathbf{V} = \mathbf{X}_2$ *and* $\mathbf{V} = 2\boldsymbol{\theta}_0 - \mathbf{X}_2$*,*
(A.1) *second moments of* $\psi(|X_1^{(i)} - V^{(i)}|)$*'s are uniformly bounded.*
(A.2) $\sum_{i \neq j} Corr\{\psi(X_1^{(i)}, V^{(i)}), \psi(X_1^{(j)}, V^{(j)})\}$ *is of the order* $o(d^2)$*.*

Note that if ψ is bounded, (A.1) holds automatically. If ρ_ψ^h is the l_p distance, (A.1) holds when the $2p$-th moment of the $X^{(i)}$'s are uniformly bounded. The assumption (A.2) implies a form of weak dependence among the measurement variables. Hall et al. (2005) looked at the d-dimensional observations as infinite time series $\left(X^{(1)}, X^{(2)}, \ldots\right)$ truncated at length d and studied the high dimensional behavior of some popular classifiers assuming a form of ρ-mixing for the time series. Assumption (A.2) holds under that ρ-mixing conditions. Jung and

Marron (2009) assumed some weak dependence among measurement variables to study the high dimensional consistency of estimated principal component directions. Assumption (A.2) holds under those conditions as well.

Now, define $\tau_d(\boldsymbol{\theta},\boldsymbol{\theta}_0) = d^{-1}\sum_{i=1}^{d} E\Big[\psi(|X_1^{(i)}+X_2^{(i)}-2\theta_0^{(i)}|)-\psi(|X_1^{(i)}-X_2^{(i)}|)\Big]$ and $\tau = \liminf_{d\to\infty}\tau_d(\boldsymbol{\theta},\boldsymbol{\theta}_0)$. In the case of Euclidean distance (i.e., when $\psi(t) = t^2$), one can show that $\tau_d(\boldsymbol{\theta},\boldsymbol{\theta}_0) = d^{-1}\sum_{i=1}^{d}(\theta^{(i)}-\theta_0^{(i)})^2 \geq 0$, where the equality holds if and only if $\theta^{(i)} = \theta_0^{(i)}$ for $i = 1, 2, \ldots, d$. Also, for any ψ, where $\psi^{'}(t)/t$ is a non-constant monotone function in $(0,\infty)$, from Baringhaus and Franz (2010) (p. 1335-1336), it follows that $E\Big[\psi(|X_1^{(i)}+X_2^{(i)}-2\theta_0^{(i)}|)-\psi(|X_1^{(i)}-X_2^{(i)}|)\Big] \geq 0$, where the equality holds if and only if $\theta^{(i)} = \theta_0^{(i)}$ $(i = 1, 2, \ldots, d)$. So, the result $\tau_d(\boldsymbol{\theta},\boldsymbol{\theta}_0) \geq 0$ also holds for such functions (e.g., $\psi(t) = t$ or $\psi(t) = t/(1+t)$), and there also $\tau_d(\boldsymbol{\theta},\boldsymbol{\theta}_0) = 0$ implies $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Therefore, under H_0, while we have $\tau = 0$, τ is expected to be positive under H_1. In such cases, the powers of the constructed distribution-free tests converge to unity as d increases.

Theorem 1 *Assume that the underlying distribution F satisfies (A.1) and (A.2). If $\tau = \liminf_{d\to\infty}\tau_d(\boldsymbol{\theta},\boldsymbol{\theta}_0) > 0$ and 2^{n-1} is larger than $1/\alpha$, the powers of the level α tests based on T_0, T_1, and T_2 converge to unity as d grows to infinity.*

Proof: Consider two independent random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ from F and define $\mathbf{X}_i^* = 2\boldsymbol{\theta}_0 - \mathbf{X}_i$ for $i = 1, 2$. Under (A.1) and (A.2), the weak law of large numbers holds for the sequence $\{\psi(|X_1^{(i)} - V^{(i)}|);\ i \geq 1\}$ (the proof is straight forward, and hence it is omitted), where $\mathbf{V} = \mathbf{X}_2$ or $\mathbf{X}_2^*$. Therefore, $\Big|d^{-1}\sum_{i=1}^{d}\psi(|\mathbf{X}_1^{(i)} - \mathbf{X}_2^{*(i)}|) - d^{-1}\sum_{i=1}^{d}\psi(|\mathbf{X}_1^{(i)} - \mathbf{X}_2^{(i)}|) - \tau_d(\boldsymbol{\theta},\boldsymbol{\theta}_0)\Big| \xrightarrow{P} 0$ as $d \to \infty$. So, if we have n independent random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from F and $\tau > 0$, for all $j \neq k$, $P\left[\sum_{i=1}^{d}\psi(|\mathbf{X}_j^{(i)} - \mathbf{X}_k^{*(i)}|) > \sum_{i=1}^{d}\psi(|\mathbf{X}_j^{(i)} - \mathbf{X}_k^{(i)}|)\right] \to 1$ as $d \to \infty$. Since h is monotonically increasing and n is finite, as $d \to \infty$, all $\mathbf{XX}$ type and $\mathbf{X}^*\mathbf{X}^*$ type distances become smaller than all $\mathbf{XX}^*$ type distances with probability tending to one. So, the shortest covering path $\mathcal{P}_0$ will contain $n-1$ edges connecting either all $\mathbf{X}_i$s or all $\mathbf{X}_i^*$s. As a result, T_0 will take either its minimum value 0 or its maximum value $\sum_{i=1}^{n} a(i)$. Now, under H_0, it takes each of these extreme values with probability $1/2^n < \alpha/2$. Therefore, the tests based on T_0 will reject H_0 with probability tending to 1. Since the path $\mathcal{P}_0$ tends to cover either all $\mathbf{X}_i$s or all $\mathbf{X}_i^*$s, T_1 converges in probability to its minimum value 1 and T_2 converges to its maximum value n. Since

$P_{H_0}(T_1 \leq 1) = P_{H_0}(T_2 \geq n) = 1/2^{n-1} < \alpha$, the powers of these two tests also converge to unity as d grows to infinity. □

In classical asymptotic regime, consistency of a test is a rather trivial property. The power of any reasonable test converges to unity as n increases. But when n is fixed, and d tends to infinity, consistency of a test is no longer a trivial property, and many well known and popular tests fail to have consistency in this set up (see e.g., Biswas and Ghosh, 2014). Theorem 1 shows that our tests are consistent in this HDLSS asymptotic regime, and for a test of 5% level, it is enough to have six observations for its high dimensional consistency.

Though our path finding method based on Prim's algorithm may fail to select the actual shortest covering path $\mathcal{P}_0$ in some of the cases, but the above theorem holds even for the implemented versions of the tests based on that algorithm. We have seen that as $d \to \infty$, under H_1, all $\rho_\psi^h(\mathbf{x}_i, \mathbf{x}_j)$ distances $(i \neq j)$ become smaller than all $\rho_\psi^h(\mathbf{x}_i, \mathbf{x}_j^*)$ distances with probability tending to one (see the proof of Theorem 1). So, this algorithm first selects an edge connecting either two $\mathbf{x}_i$s or two $\mathbf{x}_i^*$s. Now, if it selects an edge connecting two $\mathbf{x}_i$s (or $\mathbf{x}_i^*$s, respectively), because of this ordering of distances, only $\mathbf{x}_i$s (or $\mathbf{x}_i^*$s, respectively) get selected in the subsequent stages. So, for large d, the covering path selected by the algorithm contains either all $\mathbf{x}_i$s or all $\mathbf{x}_i^*$s with probability tending to 1. Note that along this path, the arrangement of the $\mathbf{x}_i$s (or the $\mathbf{x}_i^*$s) can differ from that in actual $\mathcal{P}_0$, but that re-arrangement only changes the cost of the covering path, not the values of the resulting test statistics.

4. Results from the analysis of simulated and real data sets

We analyzed some simulated and real data sets to compare the performance of our distribution-free tests with some existing one-sample tests. For this comparison, we considered the Hotelling's T^2 test with the chi-square critical value, Puri and Sen's (1971) tests based on coordinate-wise signs and ranks, spatial signs and rank tests (see e.g., Mottonen et al., 1997) and Hallin and Paindavine's (2002) sign and rank tests based on interdirections and pseudo Mahalanobis distances. We will refer to these sign and rank tests as PS-sign, PS-rank, Sp-sign, Sp-rank, HP-sign and HP-rank tests, respectively. Codes for these tests are available in different R packages. However, these tests are not applicable when the

dimension exceeds the sample size. So, in addition to them, we also considered one-sample tests proposed by Srivastava (2009), Chen and Qin (2010) and Park and Ayyala (2013), which can be used even in HDLSS situations. We will refer to them as the SR test, the CQ test and the PA test, respectively. Throughout this article, all tests are considered to have 5% nominal level.

For our tests, we used three types of distance function, the Euclidean distance, the l_1 distance and a bounded distance function with $\psi(t) = t/(1+t)$ and $h(t) = t$. Among them, the tests based on the Euclidean distance had the best overall performance. Also, the tests based on T_1 and T_2 performed better than the linear rank tests based on sign and signed rank statistics. Note that under H_1, $\mathbf{XX}$ type and $\mathbf{X}^*\mathbf{X}^*$ type distances are expected to be smaller than $\mathbf{XX}^*$ type distances. So, our path finding algorithm is supposed to start with either an $\mathbf{XX}$ type edge or an $\mathbf{X}^*\mathbf{X}^*$ type edge. Also, if it starts with an $\mathbf{XX}$ type edge, in the subsequent steps, it is supposed to choose $\mathbf{XX}$ type edges with high probability. But, if an $\mathbf{XX}^*$ type edge is chosen in the middle, there is a high probability of choosing $\mathbf{X}^*\mathbf{X}^*$ type edges in the subsequent steps. As a result, even under H_1, sometimes the values of T_0^* do not become large enough to reject H_0. We observed it in our experiments with sign and signed rank statistics. However, the tests based on T_1 and T_2 did not get much affected by this phenomenon. Therefore, in this article, we have reported the results only for T_1 and T_2 when the Euclidean distance was used.

4.1 Analysis of simulated data sets, where d is smaller than n

We begin with some examples involving multivariate normal, t (with 2 degrees of freedom) and Cauchy distributions. These distributions were chosen for varying degrees of heaviness of their tails. In each case, we generated 50 observations from a distribution with location parameter $\Delta = (\delta, \ldots, \delta)'$ and scatter matrix $\mathbf{I}$ to test $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$. We considered two choices of d (30 and 40) and four choices of δ (0, 0.1, 0.2 and 0.3) to study the level and the power properties of different tests. Each experiment was repeated 500 times, and the powers (sizes in the case of $\delta = 0$) of different tests were estimated by the proportion of times they rejected H_0.

For nonparametric sign and rank tests, we used both, the test based on the large sample distribution of the test statistic and the conditional test based on

Table 4.1: Observed powers of one-sample tests (with 5% nominal level).

	d	δ	Hot. T^2	Sp-sign	Sp-rank	PS-sign	PS-rank	CQ	PA	SR	Run (T_1)	L.Run (T_2)
Normal	30	0.0	0.048	0.048	0.048	0.044	0.052	0.044	0.044	0.044	0.062	0.056
		0.1	0.244	0.264	0.248	0.126	0.212	0.520	0.488	0.474	0.172	0.136
		0.2	0.854	0.886	0.868	0.636	0.844	0.998	0.998	0.998	0.784	0.482
		0.3	0.998	1.000	0.998	0.978	0.996	1.000	1.000	1.000	1.000	0.928
	40	0.0	0.050	0.048	0.052	0.052	0.052	0.048	0.050	0.038	0.048	0.048
		0.1	0.174	0.176	0.170	0.108	0.178	0.586	0.556	0.526	0.176	0.146
		0.2	0.646	0.682	0.664	0.396	0.652	1.000	1.000	1.000	0.852	0.574
		0.3	0.964	0.980	0.972	0.880	0.954	1.000	1.000	1.000	1.000	0.972
$t_{(2)}$	30	0.0	0.032	0.044	0.044	0.048	0.058	0.046	0.036	0.004	0.064	0.052
		0.1	0.162	0.186	0.206	0.102	0.158	0.138	0.090	0.020	0.134	0.112
		0.2	0.666	0.654	0.732	0.378	0.696	0.476	0.348	0.236	0.562	0.416
		0.3	0.956	0.946	0.972	0.820	0.954	0.774	0.640	0.598	0.942	0.872
	40	0.0	0.050	0.038	0.054	0.040	0.052	0.054	0.050	0.004	0.042	0.048
		0.1	0.116	0.126	0.140	0.096	0.278	0.160	0.116	0.026	0.132	0.106
		0.2	0.540	0.424	0.534	0.252	0.530	0.546	0.410	0.238	0.628	0.536
		0.3	0.888	0.734	0.888	0.546	0.840	0.822	0.702	0.650	0.974	0.954
Cauchy	30	0.0	0.016	0.030	0.034	0.036	0.036	0.040	0.028	0.000	0.058	0.044
		0.1	0.106	0.150	0.164	0.100	0.138	0.048	0.036	0.000	0.142	0.116
		0.2	0.472	0.458	0.570	0.308	0.540	0.086	0.060	0.006	0.434	0.474
		0.3	0.840	0.796	0.890	0.588	0.864	0.142	0.096	0.024	0.846	0.876
	40	0.0	0.034	0.048	0.046	0.056	0.054	0.046	0.024	0.000	0.034	0.048
		0.1	0.098	0.116	0.148	0.078	0.142	0.056	0.024	0.000	0.142	0.116
		0.2	0.456	0.292	0.466	0.174	0.456	0.104	0.060	0.000	0.576	0.560
		0.3	0.836	0.494	0.782	0.350	0.778	0.192	0.110	0.026	0.904	0.930
Mixture	30	0.0	0.036	0.046	0.032	0.138	0.034	0.086	0.068	0.086	0.068	0.064
		0.1	0.036	0.050	0.034	0.140	0.042	0.098	0.086	0.098	0.218	0.116
		0.2	0.038	0.054	0.038	0.102	0.042	0.126	0.120	0.126	0.926	0.524
		0.3	0.038	0.062	0.034	0.094	0.050	0.212	0.184	0.212	1.000	0.958
	40	0.0	0.062	0.062	0.042	0.196	0.044	0.086	0.074	0.086	0.052	0.034
		0.1	0.062	0.060	0.046	0.174	0.044	0.094	0.086	0.094	0.262	0.102
		0.2	0.060	0.060	0.048	0.162	0.056	0.140	0.128	0.140	0.980	0.528
		0.3	0.062	0.066	0.040	0.112	0.050	0.208	0.180	0.208	1.000	0.984

the permutation principle. In each case, the best one (which happened to be the permutation test in almost all cases) has been reported in Table 1. However, we had some problems in running the available R codes for HP-sign and HP-rank tests in the case of 40-dimensional t and Cauchy distributions. In other cases, their performance was either similar or inferior to Sp-sign and Sp-rank tests. That is why, we have not reported the results for these two tests in Table 1.

Table 1 shows that in the examples involving normal distributions, all tests had sizes close to 0.05, but in the case of t (with 2 df) and Cauchy distributions, the SR test had size much below the nominal level. The Hotelling's T^2 test and the PA test also had sizes below 0.05 in the case of Cauchy distributions. All other tests rejected the true $H_0 : \delta = 0$ in nearly 5% of the cases.

In the examples involving normal distributions, CQ, PA and SR tests had much higher powers than their competitors, though all other tests performed

well. Among them, the test based on T_1 had the best performance for $d = 40$. In the examples involving 30-dimensional t distributions, the Sp-rank test had the best performance closely followed by PS-rank, Hotelling's T^2 and Sp-sign tests. The PS-sign test and two run tests also had competitive performance. However, in the case of $d = 40$, these run tests outperformed all other tests considered here. We observed similar results in examples with Cauchy distributions as well. For $d = 30$, Hotelling's T^2, Sp-rank, PS-rank and two run tests had comparable performance, but for $d = 40$, the run tests performed much better. CQ, PA and SR tests had very poor performance in these examples.

We considered another example, where F was an equal mixture of four normal distributions all having the same scatter matrix $\frac{1}{2}\mathbf{I}$. The locations of these normal distributions were $(-3+\delta)\mathbf{1}_d$, $(-1+\delta)\mathbf{1}_d$, $(1+\delta)\mathbf{1}_d$ and $(3+\delta)\mathbf{1}_d$. Here $\lambda\mathbf{1}_d$ denotes the d-dimensional column vector having all elements equal to λ. We carried out our experiment for two choices of d and four choices of δ as before. In this example, CQ and SR tests had sizes higher than 0.05. Because of near singularity of the estimated dispersion matrix of coordinate-wise signs, the PS-sign test failed to maintain the level property. All other tests had sizes close to the nominal level (see Table 1). However, the powers of the two run tests were substantially higher than those of all other tests considered here. In the case of $\delta = 0.3$, while the test based on T_1 rejected H_0 in all occasions, and that based on T_2 had power more than 0.95, all other tests had power less than 0.25.

4.2 Analysis of simulated data sets, where d is much larger than n

We again considered some examples with normal, t and Cauchy distributions. In each of these cases, we generated 20 observations from a d-dimensional distribution having the location parameter $(0.15, \ldots, 0.15)'$ and the scatter matrix $\mathbf{I}$. The powers of different tests were computed based on 500 trials as before. We repeated the experiment for values of d ranging from 3 to 3000, and the results are reported in Fig. 2(a)-2(c). In these examples, the location of each variable differs from the origin. So, one would expect the power of these tests to tend to 1 as d increases. We observed this phenomenon in most of the cases. In the case of normal distributions, the CQ test had the best overall performance followed by the PA test. Though the SR test had the highest power for small values of d, in high dimensions, it was outperformed by the CQ test, the PA

test and the test based on T_1. In the case of t-distributions, the CQ test and two run tests performed better than PA and SR tests, while the test based on T_2 had an edge in high dimensions. The SR test had very poor performance; its power dropped down to zero as d increased. In the example involving Cauchy distributions, our run tests substantially outperformed all other tests considered here. This is consistent with what we observed in Table 1. Again, the power of the SR test was close to zero in high dimensions.

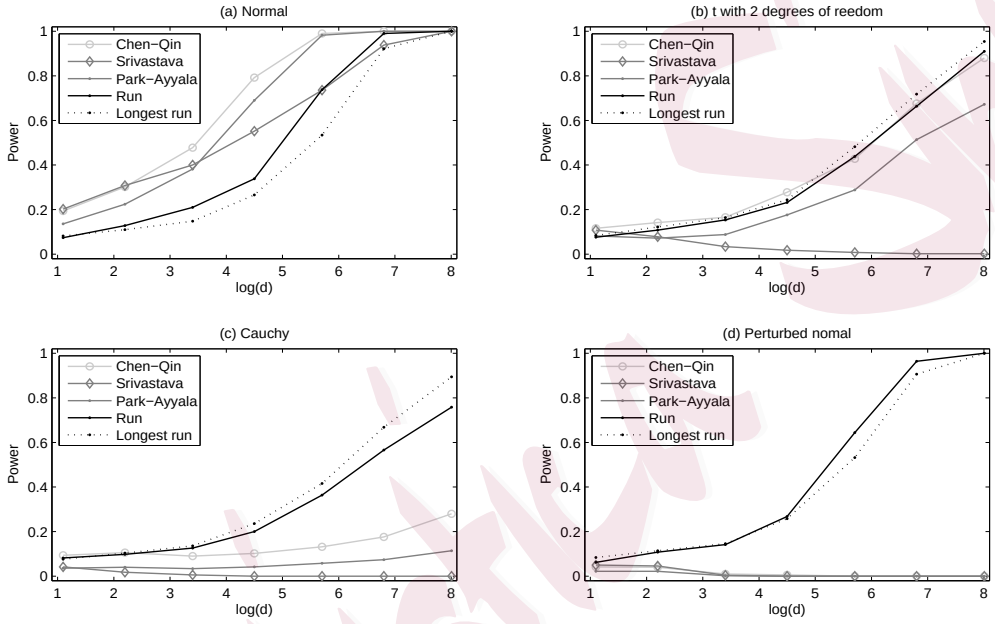

Figure 2: Powers of different tests for varying choices of d.

These examples show the robustness of our tests against heavy tailed distributions. In cases of multivariate t and Cauchy distributions, especially in the latter case, they had excellent performance when the other tests failed. However, in the example with the normal distribution, CQ and PA tests outperformed them. But, even in that case, the situation gets completely changed in the presence of contaminations. We carried out one such experiment, where we generated the observations from the normal distributions as before, but perturbed one out of the 20 observations by subtracting 2 from its each coordinate. This contamination heavily affected the performance of CQ, PA and SR tests. All of them had zero power for almost all values of d (see Fig. 2(d)). But, the tests based on

T_1 and T_2 did not get much affected. The powers of these two tests converged to 1 as before as the dimension increased.

4.3 Analysis of PEMS-SF data set

PEMS-SF data set describes the occupancy rate, between 0 and 1, of different car lanes of San Francisco bay area freeways during Jan. 01, 2008 to Mar. 30, 2009. For each day, there is a time series of dimension 963 (the number of sensors) and length $6 \times 24 = 144$ (measurement are sampled in every 10 minutes). This data set is available at the UCI machine learning repository, and there are separate training and test sets. For our analysis, we used the 126 observations in the test set after removing Saturdays and Sundays. Figure 3(a) shows average occupancy rates for different time points of a day computed over 126 days and 963 locations. In this figure, we observe two modes at 8:30 A.M. and 5:30 P.M. Corresponding to these two time points, we have two distributions of dimension 963. Here, we subtract one vector (corresponding to 5:30 P.M.) from the other (corresponding to 8:30 A.M.) and test whether the location of the difference differs from the origin. The distributions of the difference for different working days of the week are given in Fig. 3(b)-3(f). Clearly, for some of the sensors, the location differs the origin. So, one would expect the null hypothesis of no difference to be rejected.

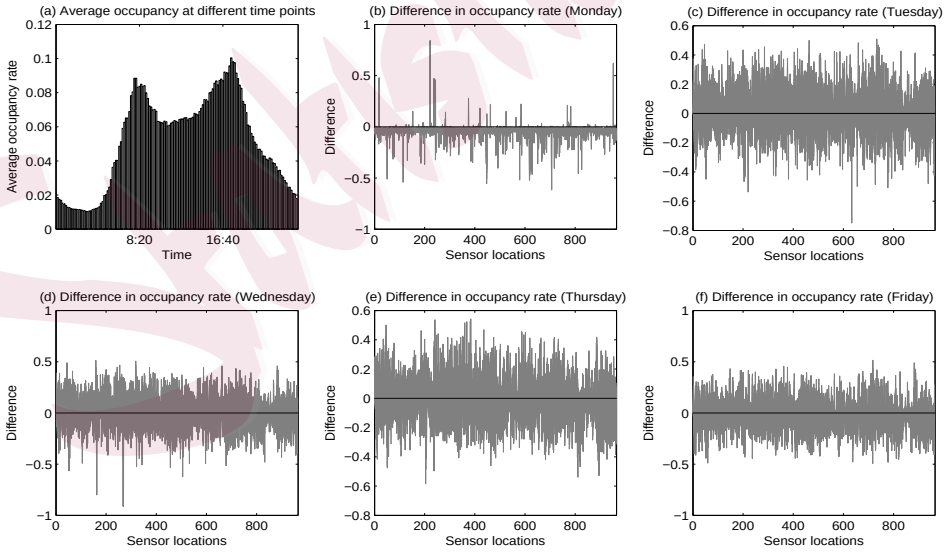

Figure 3: Occupancy rates of car lanes of San Francisco bay area freeways.

When we used 126 observations for testing, all five tests (CQ, PA, SR and two run tests) rejected H_0. Based on that single experiment, it was not possible to compare among different tests. So, we carried out our experiment using random subsets of size 5 and 10. Each experiment was repeated 500 times to estimate the powers of different tests. CQ and SR tests had the highest power 1 both for $n = 5$ and $n = 10$. The tests based on T_1 and T_2 also had power 1 for $n = 10$, but for $n = 5$, they had powers 0.812 and 0.806, respectively. The PA test had power 0.976 for $n = 10$, but in the case of $n = 5$, it could not reject H_0 even in a single occasion. To study the level properties of different tests, along with these 126 observations, we add their negatives to have a data cloud consisting of 252 observations, which is symmetric about the origin. We chose random samples of size 5 and 10 from this cloud to perform these tests, and each experiment was repeated 500 times. Both for $n = 5$ and $n = 10$, the tests based on T_1 (0.054 and 0.040, respectively) and T_2 (0.056 and 0.044, respectively) had sizes close to 0.05, but for the PA test, they were much below the nominal level (0.000 and 0.008, respectively). In the case of $n = 10$, CQ and SR tests also had sizes close to 0.05 (0.058 and 0.062, respectively). But, in the case of $n = 5$, they failed to maintain the level property and rejected H_0 in 13.6% and 15.8% cases, respectively. This bias towards H_1 could be the reason for their high powers for $n = 5$.

Acknowledgment The authors would like to thank the associate editor and the reviewer for their insightful comments and helpful suggestions.

References

Bai, Z. and Saranadasa, H. (1996) Effect of high dimension: by an example of a two-sample problem. *Statistica Sinica*, **6**, 311-–329.

Baringhaus, L. and Franz, C. (2010) Rigid motion invariant two two-sample tests. *Statistica Sinica*, **20**, 1333-1361.

Bickel, P. (1965) On some asymptotically nonparametric competitors of Hotelling's T^2. *Ann. Math. Statist.*, **36**, 160-173.

Biswas, M. and Ghosh, A. K. (2014) A nonparametric two-sample test applicable to high dimensional data. *J. Multivariate Anal.*, **123**, 160-171.

Chakraborty, B., Chaudhuri, P. and Oja, H. (1998) Operating transformation re-transformation on spatial median and angle test. *Statistica Sinica*, **8**, 767-784.

Chaudhuri, P. and Sengupta, D. (1993) Sign tests in multi-dimension: inference based on the geometry of the data cloud. *J. Amer. Statist. Assoc.*, **88**, 1363-1370.

Chen, S. X. and Qin, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, **38**, 808—835.

Fu, J. C. and Koutras, M. V. (1994) Distribution theory of runs: a Markov chain approach. *J. Amer. Statist. Assoc.*, **89**, 1050–1058.

Garey, M. and Johnson, D. (1979) *Computers and Intractability: A Guide to the Theory of NP Completeness*. W.H. Freeman and Co., San Francisco.

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric Statistical Inference*, CRC Press, Boca Raton, Florida.

Gordon, L., Schilling, M. and Waterman, M. (1986) An extreme value thoery for long head runs. *Probability Theory and Related Fields*, **72**, 279–287.

Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension low sample size data. *J. Royal Statist. Soc. Ser. B*, **67**, 427-444.

Hallin, M. and Paindaveine, D. (2002) Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Ann. Statist.*, **30**, 1103-1133.

Hallin, M. and Werker, B. J. M. (2003) Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli*, **9**, 137-165.

Hettmansperger, T., Mottonen, J. and Oja, H. (1997) Multivariate affine invariant one-sample signed rank tests. *J. Amer. Statist. Assoc.*, **92**, 1591-1600.

Hettmansperger, T., Nymblom, J. and Oja, H. (1994) Affine invariant multivariate one-sample sign tests. *J. Royal Statist. Soc., Ser. B*, **56**, 221-234.

Hodges, J. L. (1955) A bivariate sign test. *Ann. Math. Statist.*, **26**, 523-527.

Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.*, **37**, 4104-–4130.

Lawler, E., Lenstra, J., Kan, A. and Shmoys, D. (1985) *The Travelling Salesman Problem*, Wiley, New York.

Mottonen, J., Oja, H. and Tienari, J. (1997) On the efficiency of multivariate sign and rank tests. *Ann. Statist.*, **25**, 542-552.

Park, J. and Ayyala, D. N. (2013) A test for the mean vector in large dimension and small samples. *J. Statist. Plann. Inference*, **143**, 929–943.

Prim, R. C. (1957) Shortest connection networks and some generalizations. *Bell System Technical Journal*, **36**, 1389–1401.

Puri, M. L. and Sen, P. K. (1971) *Nonparametric Methods in Multivariate Analysis*, Wiley, New York.

Randles, R. (1989) A distribution-free multivariate sign test based on interdirections. *J. Amer. Statist. Assoc.*, **84**, 1045-1050.

Randles, R. (2000) A simpler, affine invariant, multivariate, distribution-free sign test. *J. Amer. Statist. Assoc.*, **95**, 1263-1268.

Srivastava, M. S. and Du, M. (2008) A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.*, **99**, 386–402.

Srivastava, M. S. (2009) A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.*, **100**, 518-–532.

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata.

E-mail: (munmun.biswas08@gmail.com)

Applied Statistics Unit, Indian Statistical Institute, Kolkata.

E-mail: (minervamukherjee@gmail.com)

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, Kolkata.

E-mail: (akghosh@isical.ac.in)