Exact distribution-free two-sample tests applicable to high dimensional data

Anil K. Ghosh and Munmun Biswas

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203, Barrackpore Trunk Road, Kolkata 700108, India.

E-mail :akghosh@isical.ac.in, munmun_r@isical.ac.in.

Abstract

Multivariate two-sample testing problem is extensively studied in the literature, and both parametric and nonparametric tests are available for it. However, most of these tests perform poorly for high dimensional data, and they are usually not applicable when the dimension is larger than the sample size. In this article, we propose a procedure for multivariate generalization of well-known univariate distribution-free tests for two-sample problems involving independent as well as matched pair samples. This proposed procedure is based on ranks of real valued linear functions of multivariate observations. The linear function used to rank the observations is obtained by solving a classification problem that aims at finding an optimal discriminating hyperplane between the two multivariate distributions from which the observations are generated. Our tests perform well for high dimensional data even when the dimension exceeds the sample size. Besides, our tests are distribution-free under very general conditions. Asymptotic results on the powers of the proposed tests are derived when the sample sizes are fixed and the dimension of the data grows to infinity as well as for situations when the sample sizes grow while the dimension of the data remains fixed. We investigate the finite sample performance of our proposed tests by applying it to several high dimensional simulated and real data sets and compare them with several other tests available in the literature.

Keywords : Distance weighted discrimination, Komogorov-Smirnov statistic, Linear rank statistic, Sign test, Signed rank test, Support vector machines, Wilcoxon-Mann-Whitney statistic.

1 Introduction : the two-sample problem involving independent samples

Suppose that we have two independent samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1} \overset{i.i.d}{\sim} F$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2} \overset{i.i.d}{\sim} G$. Let $\mathcal{F}$ be a family of pairs of distributions (F, G) and $\mathcal{F}_0$ be a subset of $\mathcal{F}$ defined as $\mathcal{F}_0 = \{(F, G) \in \mathcal{F} : F = G\}$. Here, we want to test the null hypothesis $H_0 : (F, G) \in \mathcal{F}_0$ against the alternative hypothesis $H_A : (F, G) \in \mathcal{F} - \mathcal{F}_0$. For univariate two-sample problems, the classical t-test, the Wilcoxon-Mann-Whitney (WMW) rank test and the Kolmogorov-Smirnov (KS) test are very well

known and widely used. While the t-test is optimal for normally distributed data under some appropriate conditions, WMW and KS tests, which are based on the ranks of the univariate observations in the combined sample, have distribution free property under the null hypothesis, and they outperform the two-sample t-test for a wide variety of non-Gaussian distributions. The WMW test is suitable when there is some stochastic ordering between the two distributions F and G under the alternative hypothesis, while the KS test is meant for more general nonparametric alternatives. The multivariate version of the t-test is the well-known Hotelling's T^2 test, which has several optimal properties for data having multivariate Gaussian distributions. On the other hand, several attempts have also been made in the literature to generalize univariate rank tests into the multivariate set up. Perhaps the simplest among those are the two-sample tests based on co-ordinatewise ranks (see e.g., Puri and Sen, 1971). Brown and Hettmansperger (1987, 1989) developed rank based methods for bivariate two-sample location models. Hettmansperger and Oja (1994) and Hettsmanperger et. al. (1998) extended them to general multivariate set up. Randles and Peters (1990) developed multivariate two-sample rank tests based on interdirections. Mottonen and Oja (1995) and Choi and Marden (1997) used spatial ranks for constructing multivariate generalizations of two-sample rank tests. Some good reviews of most of these tests can be found in Chakraborty and Chaudhuri(1999), Marden(1999), Oja and Randles (2004) and Oja (2010).

All of the above mentioned multivariate extensions of univariate rank tests as well as Hotelling's T^2 test for multivariate two-sample problems yield poor results when applied to high dimensional data sets, and none of them can be used when the dimension of the data is larger than the sample size. Unlike univariate rank based tests, none of them have the exact distribution-free property in finite sample situations. In some cases, those tests are asymptotically distribution-free, and in some cases, one can implement their conditional versions using permutation type techniques. Liu and Singh (1993) developed some distribution-free two-sample tests based on simplicial depth, but those are computationally infeasible in high dimensions, and they cannot be used when the dimension exceeds the sample size. Recall that for $\mathbf{x} \sim F$ and $\mathbf{y} \sim G$, the equality of the distributions F and G is equivalent to the equality of the distributions of $\boldsymbol{\beta}'\mathbf{x}$ and $\boldsymbol{\beta}'\mathbf{y}$ for all choices of $\boldsymbol{\beta} \in R^d$ (throughout this paper, all vectors will be assumed to be column vectors). Therefore, in the case of multivariate two-sample problems, $\mathcal{F}_0 \subset \mathcal{F}$ can be expressed as $\mathcal{F}_0 = \{(F, G) \in \mathcal{F} : F_{\boldsymbol{\beta}} = G_{\boldsymbol{\beta}} \ \ \forall \boldsymbol{\beta} \in R^d\}$, where d is the dimension of the data, and $\boldsymbol{\beta}'\mathbf{x} \sim F_{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}'\mathbf{y} \sim G_{\boldsymbol{\beta}}$ if $\mathbf{x} \sim F$ and $\mathbf{y} \sim G$, respectively. In other words, the hypothesis $H_0 : F = G$ can be viewed as an intersection of the hypotheses $H_{0,\boldsymbol{\beta}} : F_{\boldsymbol{\beta}} = G_{\boldsymbol{\beta}}$ for varying choices of $\boldsymbol{\beta} \in R^d$. In this article, we propose certain tests based on ranks of real valued linear functions of the multivariate observations. Our tests have exact

distribution-free property in finite sample situations under very general conditions. Also, they are conveniently applicable to high dimensional data even when the dimension is larger than the sample size, and they have good power properties as demonstrated in the subsequent sections.

2 Construction of distribution-free tests for two independent samples

From now on, we will assume that both F and G are absolutely continuous w.r.t. the Lebsegue measure. If the multivariate sample observations $\mathbf{x}_1, \ldots, \mathbf{x}_{n_1}, \mathbf{y}_1, \ldots, \mathbf{y}_{n_2}$are projected along the direction $\boldsymbol{\beta}$, we get two sets of univariate observations $\boldsymbol{\beta}^{'}\mathbf{x}_1, \ldots, \boldsymbol{\beta}^{'}\mathbf{x}_{n_1} \sim F_{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^{'}\mathbf{y}_1, \ldots, \boldsymbol{\beta}^{'}\mathbf{y}_{n_2} \sim G_{\boldsymbol{\beta}}$, and we can use any suitable univariate distribution-free two-sample test on these projected observations. Clearly, any test based on ranks of a fixed linear function of the multivariate observations will have the exact distribution-free property. However, in order to have good power properties of such a test based on linear projections, one should choose the direction vector $\boldsymbol{\beta}$ in such a way that the separation between the projected observations from the two populations is maximized along that direction in an appropriate sense. One possible way to achieve this is to use the direction vector of a suitable linear classifier that discriminates between the two multivariate populations. The motivation for this choice partially comes from the fact that for two multivariate normal distributions with a common dispersion and different means, if one computes the univariate two-sample t-statistic based on linear projections of the data points along the director vector used in Fisher's linear discriminant function, where the mean vectors and the common covariance matrix for the two normal distributions are estimated from the data, it leads to the Hotelling's T^2 statistic. Further, for two independent normal random vectors $\mathbf{x}$ and $\mathbf{y}$, with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and a common dispersion matrix $\boldsymbol{\Sigma}$, the power of the univariate t-test for testing $H_{0,\boldsymbol{\beta}} : \boldsymbol{\beta}^{'}\boldsymbol{\mu}_1 = \boldsymbol{\beta}^{'}\boldsymbol{\mu}_2$ based on $\boldsymbol{\beta}^{'}\mathbf{x}$ and $\boldsymbol{\beta}^{'}\mathbf{y}$ is a monotonically increasing function of $\{\boldsymbol{\beta}^{'}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\}^2/\boldsymbol{\beta}^{'}\boldsymbol{\Sigma}\boldsymbol{\beta}$. So, a simple application of Cauchy-Schwartz inequality shows that the power of the test is maximized when $\boldsymbol{\beta}$ is chosen to be a positive scalar multiple of $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, which is the coefficient vector of Fisher's linear discriminant function.

Even when the underlying distributions are not normal, we have some nice connections between classification and hypothesis testing problems. Consider a classification problem between two multivariate distributions F and G such that the prior probabilities of these two distributions are equal. Let us also consider a discriminating hyperplane $\{\mathbf{w} : \beta_0 + \boldsymbol{\beta}^{'}\mathbf{w} = 0,\ \mathbf{w} \in R^d\}$ between these two

distributions. Suppose that the classifier classifies $\mathbf{w}$ to be an observation from F if $\beta_0 + \boldsymbol{\beta}'\mathbf{w} > 0$ and to be an observation from G if $\beta_0 + \boldsymbol{\beta}'\mathbf{w} \leq 0$. Clearly, the average misclassification probability of this classifier is given by $\Delta(\beta_0, \boldsymbol{\beta}) = 0.5[1 - \{F_{\boldsymbol{\beta}}(-\beta_0) - G_{\boldsymbol{\beta}}(-\beta_0)\}]$. So, $\Delta(\beta_0, \boldsymbol{\beta})$ is minimized if and only if $\boldsymbol{\beta}$ maximizes the KS distance between $F_{\boldsymbol{\beta}}$ and $G_{\boldsymbol{\beta}}$. Further, when F and G are both elliptically symmetric unimodal distributions, which differ only in their locations, we have the following result, which yields an interesting insight into the connection between classifiers having the optimal misclassification rate and tests having the optimal power.

PROPOSITION 2.1. *Suppose that F and G are elliptically symmetric unimodal multivariate distributions, which differ only in their locations. Then the one sided KS test as well as the one sided WMW test based on the ranks of $\boldsymbol{\beta}'\mathbf{x}_1, \boldsymbol{\beta}'\mathbf{x}_2, \ldots, \boldsymbol{\beta}'\mathbf{x}_{n_1}, \boldsymbol{\beta}'\mathbf{y}_1, \boldsymbol{\beta}'\mathbf{y}_2, \ldots, \boldsymbol{\beta}'\mathbf{y}_{n_2}$ will have the maximum power if and only if $\boldsymbol{\beta}$ coincides with the direction vector that determines the Bayes discriminating hyperplane associated with the classification problem involving distributions F and G with equal prior probabilities.*

PROOF. Without loss of generality, let us assume that F and G have locations $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}$, respectively. Also, it is enough to consider only those $\boldsymbol{\beta}$'s for which $\boldsymbol{\beta}'\boldsymbol{\mu} > 0$ and $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} = 1$, where $\boldsymbol{\Sigma}$ is the common scatter matrix associated with F and G. For all such choices of $\boldsymbol{\beta}$, the distribution of $\boldsymbol{\beta}'\mathbf{x}$ remains the same with location 0 and scatter 1. The distribution $\boldsymbol{\beta}'\mathbf{y}$ also remains the same except for its location $\boldsymbol{\beta}'\boldsymbol{\mu} > 0$. Now, consider two direction vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ such that $\boldsymbol{\beta}_1'\boldsymbol{\mu} > \boldsymbol{\beta}_2'\boldsymbol{\mu} > 0$. Clearly, this implies that $\boldsymbol{\beta}_1{}'\mathbf{y}$ is stochastically larger than $\boldsymbol{\beta}_2{}'\mathbf{y}$. Consequently, the ranks of the corresponding linear functions of the data points will have a similar stochastic ordering. Hence, the powers of the one sided KS test and the one sided WMW test will be higher if the data are projected along $\boldsymbol{\beta}_1$ than the powers of these tests based on data projected along $\boldsymbol{\beta}_2$. Therefore, in order to maximize the power of any such test, one needs to maximize $\boldsymbol{\beta}'\boldsymbol{\mu}$ subject to $\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} = 1$. Since $(\boldsymbol{\beta}'\boldsymbol{\mu})^2 \leq (\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})(\boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$, this maximum is achieved when $\boldsymbol{\beta}$ is a positive scalar multiple of $\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, the direction vector corresponding to the Bayes classifier.

Note at this point that a linear classifier, which has its class boundary defined by the hyperplane $\{\mathbf{w} : \beta_0 + \boldsymbol{\beta}'\mathbf{w} = 0,\ \mathbf{w} \in R^d\}$, will classify $\mathbf{w}$ as an observation from the distribution F if it falls on one side of that hyperplane, and $\mathbf{w}$ will be classified as an observation from G if it falls on the other side of the hyperplane. Since the two sides of the hyperplane are defined by the inequalities $\boldsymbol{\beta}'\mathbf{w} > -\beta_0$ and $\boldsymbol{\beta}'\mathbf{w} < -\beta_0$, it is appropriate to consider one sided KS test and one sided WMW test based on the ranks of $\boldsymbol{\beta}'\mathbf{x}_1, \boldsymbol{\beta}'\mathbf{x}_2, \ldots, \boldsymbol{\beta}'\mathbf{x}_{n_1}, \boldsymbol{\beta}'\mathbf{y}_1, \boldsymbol{\beta}'\mathbf{y}_2, \ldots, \boldsymbol{\beta}'\mathbf{y}_{n_2}$.

2.1 Adaptive determination of the projection direction

It is well-known that Fisher's linear discriminant function yields an optimal separation between two classes of observations when the underlying distributions are Gaussian having a common dispersion but different means. However, when one needs to estimate the dispersion and the means from the data, the estimated discriminant function performs poorly for high dimensional data. If the dimension of the data exceeds the total sample size, the estimated dispersion becomes singular, and it becomes difficult to use it for construction of Fisher's linear discriminant function. If one uses the Fisher's linear discriminant function based on the Moore-Penrose generalized inverse of the pooled dispersion matrix in such situations, it usually leads to poor performance in high dimensions (see e.g., Bickel and Levina, 2004).

Support vector machine (SVM) (see e.g., Vapnik, 1998; Hastie et. al., 2009) is a well-known classification tool available in the literature, and it can be used to construct a linear classifier for discriminating between two distributions when the data are high dimensional. Suppose that we have a data set of the form $\{(\mathbf{w}_i, z_i);\ i = 1, 2, \ldots, n = n_1 + n_2\}$, where z_i takes the value 1 and -1 if the observation $\mathbf{w}_i$ comes from the first population (i.e., $\mathbf{w}_i = \mathbf{x}_j$ for some j) and the second population (i.e., $\mathbf{w}_i = \mathbf{y}_j$ for some j), respectively. When the data clouds from two distributions have *perfect linear separation*, SVM looks for two parallel hyperplanes $\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w} = 1$ and $\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w} = -1$ such that $(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)z_i \geq 1$ for all $i = 1, 2, \ldots, n$, and the distance between these two hyperplanes $2/\|\boldsymbol{\beta}\|$ is maximum. In practice, it finds the separating hyperplane $\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w} = 0$ by minimizing $\frac{1}{2}\|\boldsymbol{\beta}\|^2$ subject to $(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)z_i \geq 1\ \forall\ i = 1(1)n$. If the data clouds from the two distribution are *not perfectly linearly separable*, SVM introduces slack variables ζ_i $(i = 1, 2, \ldots, n)$ and modifies the objective function by adding a cost $C_0 \sum_{i=1}^{n} \zeta_i$ (C_0 is a cost parameter) to it. In such cases, SVM minimizes $\frac{1}{2}\|\boldsymbol{\beta}\|^2 + C_0 \sum_{i=1}^{n} \zeta_i$ subject to $(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)z_i \geq 1 - \zeta_i$ and $\zeta_i \geq 0\ \forall\ i = 1(1)n$, and it uses the quadratic programming techniques to solve this minimization problem. This optimization problem is often reformulated as the problem of minimizing $S_n(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[1 - z_i(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)]_+ + \frac{\lambda}{2}\|\boldsymbol{\beta}\|^2$, where $[t]_+ = \max\{t, 0\}$ and $\lambda = 1/C_0$ is a regularization parameter (see e.g., Hastie *et. al.*, 2004).

Marron et. al. (2007) proposed a classification technique called distance weighted discrimination (DWD), which can also be used for linear classification in high dimensions. If the data from the two distributions are perfectly linearly separable, DWD finds the separating hyperplane by minimizing $\sum_{i=1}^{n}\{(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)z_i\}^{-1}$ subject to $\|\boldsymbol{\beta}\| \leq 1$ and $(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)z_i \geq 0$ for all $i = 1, 2, \ldots, n$. When the data clouds of the two distributions are not linearly separable, like SVM, DWD also introduces slack variables ζ_i to modify the objective function by adding a cost function $C\sum_{i=1}\zeta_i$, where C

is a cost parameter. So, in such cases, DWD finds the separating hyperplane $\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w} = 0$ by minimizing $\sum_{i=1}^{n} 1/r_i + C\sum_{i=1}^{n}\zeta_i$ subject to $\|\boldsymbol{\beta}\| \leq 1$, $\zeta_i \geq 0$ and $r_i = (\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)z_i + \zeta_i \geq 0$ for all $i = 1, 2, \ldots, n$. This is equivalent to minimization of $D_n(\beta_0, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}[V\{z_i(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{w}_i)\}]$, where

$$V(t) = \begin{cases} 2\sqrt{C} - Ct & \text{if } t \leq 1/\sqrt{C} \\ 1/t & \text{otherwise,} \end{cases}$$

(see e.g., Qiao et. al., 2010). DWD uses the interior point cone programming to minimize the objective function and to find an estimate of $\boldsymbol{\beta}$.

Clearly, for any fixed and non-random $\boldsymbol{\beta}$, the random variables $\boldsymbol{\beta}^{'}\mathbf{x}_1, \ldots, \boldsymbol{\beta}^{'}\mathbf{x}_{n_1}, \boldsymbol{\beta}^{'}\mathbf{y}_1, \ldots, \boldsymbol{\beta}^{'}\mathbf{y}_{n_2}$ form an exchangeable collection if $F = G$, and the ranks of these variables have the distribution-free property under H_0. Let $T_{\boldsymbol{\beta}}$ be a statistic based on the ranks of $\boldsymbol{\beta}^{'}\mathbf{x}_1, \ldots, \boldsymbol{\beta}^{'}\mathbf{x}_{n_1}, \boldsymbol{\beta}^{'}\mathbf{y}_1, \ldots, \boldsymbol{\beta}^{'}\mathbf{y}_{n_2}$ Assume that, for any specified level $0 < \alpha < 1$, the test for the null hypothesis $H_0 : F = G$ based on $T_{\boldsymbol{\beta}}$ is described by the test function

$$\phi_\alpha(T_{\boldsymbol{\beta}}) = \begin{cases} 1 & \text{if } T_{\boldsymbol{\beta}} > t_\alpha \\ \gamma_\alpha & \text{if } T_{\boldsymbol{\beta}} = t_\alpha \\ 0 & \text{otherwise,} \end{cases}$$

where one can choose t_α and γ_α in such a way that we have $E_{F,G}\{\phi_\alpha(T_{\boldsymbol{\beta}})\} = \alpha$ for all $(F, G) \in \mathcal{F}_0$. Because of the distribution-free property of $T_{\boldsymbol{\beta}}$, t_α and γ_α depend neither on (F, G) nor on $\boldsymbol{\beta}$. Further, for standard nonparametric tests (e.g., KS and WMW), one can obtain t_α and γ_α from standard tables or softwares.

Note, however, at this point that if $\boldsymbol{\beta}$ is estimated based on the whole sample using some classification method like SVM or DWD, and then the multivariate observations are ranked after projecting them along that estimated direction, the resulting ranks will not have the distribution-free property. This is due to the fact that $\widehat{\boldsymbol{\beta}}$, which is constructed from a classification problem based on the two samples, is not a symmetric function of the observations in the combined sample, and the random variables $\widehat{\boldsymbol{\beta}}^{'}\mathbf{x}_1, \ldots, \widehat{\boldsymbol{\beta}}^{'}\mathbf{x}_{n_1}, \widehat{\boldsymbol{\beta}}^{'}\mathbf{y}_1, \ldots, \widehat{\boldsymbol{\beta}}^{'}\mathbf{y}_{n_2}$ do not form an exchangeable collection even if $F = G$. Therefore, in order to have a distribution-free test, we adopt a strategy, which is motivated by a fundamental idea lying at the root of cross-validation techniques (see e.g., Hastie et. al., 2009) used in statistical model selection. In cross-validation, one splits the whole sample into subsamples, and then uses one subsample to estimate the model by optimizing a suitable criterion, while another subsample is used to assess the adequacy of the estimated model. In a similar way, we randomly split each of the two samples into two disjoint subsamples. We use a suitable linear classification technique (e.g., SVM or DWD) to construct $\widehat{\boldsymbol{\beta}}$ based on one subsample containing

some of the $\mathbf{x}$'s and one subsample containing some of the $\mathbf{y}$'s . Then we project the observations in the remaining two subsamples using that $\widehat{\boldsymbol{\beta}}$ and compute the test statistic $T_{\widehat{\boldsymbol{\beta}}}$ based on the ranks of those projected observations. We repeat this procedure for different random splits, and our test function ϕ_α^* for a given level $0 < \alpha < 1$ is obtained by averaging the test functions $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$ over different random splits. Since ϕ_α^* may take a fractional value in the open interval $(0, 1)$ for a given data set, the implementation of the test will require randomization. The exact distribution-free property of ϕ_α^* in finite samples will hold as asserted in the following result.

THEOREM 2.2. *ϕ_α^* is a distribution-free test function in the sense that $E_{(F,G)}(\phi_\alpha^*) = \alpha$ for all $(F, G) \in \mathcal{F}_0$. For a given data set, The P-value for this test function can be defined as $p = \inf\{\alpha : H_0$ is rejected by $\phi_\alpha^*\}$, and p will have uniform distribution on $(0, 1)$ for any $(F, G) \in \mathcal{F}_0$.*

PROOF. For any split of the data into independent subsamples, $\widehat{\boldsymbol{\beta}}$ is independent of the data points from which $T_{\widehat{\boldsymbol{\beta}}}$ and $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$ are computed. As a consequence, given $\widehat{\boldsymbol{\beta}}$, the conditional size of the test, which is same as the conditional expectation of the test function $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$, will be α for any $(F, G) \in \mathcal{F}_0$ in view of the distribution-free property of the test statistic $T_{\boldsymbol{\beta}}$ for any fixed $\boldsymbol{\beta}$. Since this conditional expectation does not depend on $\widehat{\boldsymbol{\beta}}$ nor on the specific split involved, and the test function ϕ_α^* is the average of the test functions $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$ over different splits, we must have $E_{(F,G)}(\phi_\alpha^*) = \alpha$ for all $(F, G) \in \mathcal{F}_0$. Note at this point that for a given data set and a specific split of the data into subsamples, $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})$ is a non-decreasing function of α, and so is ϕ_α^* being an average of those non-decreasing functions. Now it is straight forward to verify that, for a given data set, the P-value defined in the statement of the theorem will be smeller than $u \in (0, 1)$ if and only if the test function ϕ_u^* leads to rejection of H_0 for that data set. Hence, $P_{(F,G)}(p < u) = E_{(F,G)}(\phi_u^*) = u$ for any $u \in (0, 1)$ and $(F, G) \in \mathcal{F}_0$.

3 Power properties of the proposed tests for high dimensional data

In this section, We first carry out some theoretical analysis of the power properties of our proposed multivariate tests for high dimensional data, and then we use some simulated and real data sets to compare the powers of these tests with some other multivariate two-sample tests. We have already mentioned that unlike most of the existing two-sample tests, our tests based on SVM and DWD can be used even when the dimension is much larger than the sample size. In the following subsection, we investigate the limiting behavior of our tests when n is fixed, and d diverges to infinity.

3.1 Asymptotic results for d growing to infinity when n_1 and n_2 remain fixed

Here we look at the d-dimensional observations on $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})$ and $\mathbf{y} = (y^{(1)}, \ldots, y^{(d)})$ as i.i.d realizations of infinite sequences $(x^{(1)}, x^{(2)}, \ldots)$ and $(y^{(1)}, y^{(2)}, \ldots)$ truncated at length d, and investigate the behavior of the power functions of our tests as the length of the truncated sequence increases. For this investigation, we consider the $\mathbf{x}$-sequences and $\mathbf{y}$-sequences to be independent, and following Hall et. al. (2005), we make the following assumptions

(A1) *Fourth moments of $x^{(k)}$ and $y^{(k)}$ are uniformly bounded.*

(A2) *Let $\mathcal{X}$ and $\mathcal{X}'$ be two independent copies of $(x^{(1)}, x^{(2)}, \ldots)$, while $\mathcal{Y}$ and $\mathcal{Y}'$ are two independent copies of $(y^{(1)}, y^{(2)}, \ldots)$. For $(\mathbf{u}, \mathbf{v}) = (\mathcal{X}, \mathcal{X}'), (\mathcal{X}, \mathcal{Y})$ and $(\mathcal{Y}, \mathcal{Y}')$, the sequence $\{(u^{(k)} - v^{(k)})^2;\ k \geq 1\}$ is a ρ-mixing i.e., $\sup_{k \geq 1, l \geq k+r} |corr\{(u^{(k)} - v^{(k)})^2, (u^{(l)} - v^{(l)})^2\}| \leq \rho(r)$, where $\rho(r) \to 0$ as $r \to \infty$.*

Under (A1) and (A2), the weak law of large number (WLLN) holds for the sequence $\{(u^{(k)} - v^{(k)})^2;\ k \geq 1\}$ (see e.g., Billingsley, 1995; Hall et. al., 2005). In fact, we have this weak law even when (A2) holds under some permutation of the $x^{(k)}$'s (and the same permutation of the $y^{(k)}$'s). Note that if the elements of the sequence are i.i.d, WLLN holds under the existence of second order moments of $x^{(k)}$ and $y^{(k)}$. We need (A1) and (A2) for WLLN of the sequence of dependent and non-identically distributed random variables. Several other sufficient conditions for WLLN of non i.i.d. sequences have also been worked out in the literature (see e.g., Andrews, 1988, Jung and Marron, 2009). Following Hall et. al. (2005) and Jung and Marron (2009), here we also assume that

(A3) *There exist constants $\sigma_1^2, \sigma_2^2 > 0$ and μ such that $d^{-1} \sum_{k=1}^{d} Var(x^{(k)}) \to \sigma_1^2$, $d^{-1} \sum_{k=1}^{d} Var(y^{(k)}) \to \sigma_2^2$ and $d^{-1} \sum_{k=1}^{d} \{E(x^{(k)}) - E(y^{(k)})\}^2 \to \mu^2$ as $d \to \infty$.*

Under assumptions (A1)-(A3), the pairwise distance between any two observations, when divided by $d^{1/2}$, converges in probability to positive constant as d tends to infinity. If both of them are from the same distribution, it converges to $\sigma_1\sqrt{2}$ or $\sigma_2\sqrt{2}$ depending on whether they are from F or G. If one of them is from F and the other one is from G, it converges to $\sqrt{\sigma_1^2 + \sigma_2^2 + \mu^2}$. So, for large d, after re-scaling (by a factor of $d^{-1/2}$), n sample observations tend to lie on the vertices of an n-polyhedron. Note that n_1 out of these n vertices are limits of n_1 i.i.d observations from F, and they form a regular simplex $\mathcal{S}_1$ of side length $\sigma_1\sqrt{2}$. The other n_2 vertices are limits of n_2 data points from G, and they form another regular simplex S_2 of side length $\sigma_2\sqrt{2}$. The rest of the edges of the polyhedron connect the vertices of $\mathcal{S}_1$ to those of $\mathcal{S}_2$, and they are of length $\sqrt{\sigma_1^2 + \sigma_2^2 + \mu^2}$. Under H_0, when we have $\sigma_1^2 = \sigma_2^2$ and $\mu^2 = 0$, and the whole polyhedron turns out to be regular simplex

on n points, while we may have $\mu^2 > 0$ under the alternative hypothesis H_A. In a sense, (A1)-(A3) and $\mu^2 > 0$ ensure that the amount of information for discrimination between the $\mathbf{x}$ and $\mathbf{y}$ sequences grows to infinity as the dimension increases (see Hall et. al., 2005 for further discussion about these conditions and convergence results). In conventional asymptotics, we get more information as the sample size increases, but here the sample size is fixed, and we expect the amount of information to diverge as the dimension d tends to infinity. The next theorem establishes the consistency of the proposed tests under these conditions.

THEOREM 3.1 *Consider splits of the two samples such that in each split, there are two subsamples with sizes m_1 and $n_1 - m_1$ that consist of observations on $\mathbf{x}$, and two subsamples with sizes m_2 and $n_2 - m_2$ that consist of observations on $\mathbf{y}$. Let $\widehat{\boldsymbol{\beta}}$ be computed using SVM or DWD applied to the two subsamples with sizes m_1 and m_2, and $T_{\widehat{\boldsymbol{\beta}}}$ be computed from the other two subsamples with sizes $n_1 - m_1$ and $n_2 - m_2$. Assume that $T_{\boldsymbol{\beta}}$ is either the one sided KS statistic or any one sided WMW statistic such that $P_{H_0}(T_{\boldsymbol{\beta}} = t_{\max}) < \alpha$, where $t_{\max}$ is the largest possible value of the statistic $T_{\boldsymbol{\beta}}$ computed based on two subsamples with sizes $n_1 - m_1$ and $n_2 - m_2$. Then under the assumptions* **(A1)-(A3)**, *if $\mu^2 > 0$, we have $E_{H_A}(\phi^*_\alpha) \to 1$ as $d \to \infty$.*

PROOF. Under (A1)-(A3), as $d \to \infty$, $\|\mathbf{x}_i - \mathbf{x}_j\|\sqrt{d} \xrightarrow{P} \sigma_1\sqrt{2}$ for $1 \le i < j \le n_1$, $\|\mathbf{y}_i - \mathbf{y}_j\|\sqrt{d} \xrightarrow{P} \sigma_2\sqrt{2}$ for $1 \le i < j \le n_2$, and $\|\mathbf{x}_i - \mathbf{y}_j\|/\sqrt{d} \xrightarrow{P} \sqrt{\sigma_1^2 + \sigma_2^2 + \mu^2}$ for $1 \le i \le n_1$ and $1 \le j \le n_2$. So, after re-scaling, m_1 observations from F tend to lie on the vertices of a regular simplex $\mathcal{S}_1$ and m_2 observations from G tend to lie on the vertices of another regular simplex $\mathcal{S}_2$, while each vertex of $\mathcal{S}_1$ is equidistant from all vertices of $\mathcal{S}_2$ and vice versa. Because of this symmetric nature of data geometry, the discriminating surface constructed by SVM applied to the subsamples with sizes m_1 and m_2 consisting of observations on $\mathbf{x}$ and $\mathbf{y}$ bisects each of the $m_1 m_2$ lines joining the vertices of $\mathcal{S}_1$ and $\mathcal{S}_2$. So, if $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the means of these m_1 and m_2 observations on $\mathbf{x}$ and $\mathbf{y}$, and $\widehat{\boldsymbol{\beta}}$ is the projection direction estimated by SVM, $\widehat{\boldsymbol{\beta}}$ tends to be proportional to $\bar{\mathbf{y}} - \bar{\mathbf{x}}$ in the sense that $\left\| \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|} - \frac{\bar{\mathbf{y}} - \bar{\mathbf{x}}}{\|\bar{\mathbf{y}} - \bar{\mathbf{x}}\|} \right\| \xrightarrow{P} 0$ as $d \to \infty$. Similar results hold if $\boldsymbol{\beta}$ is estimated using the DWD classifier as well (see proofs of Theorems 1 and 2 in Hall et. al., 2005). So, both for SVM and DWD, the linear transformations $\mathbf{w} \to \widehat{\boldsymbol{\beta}}'\mathbf{w}$ and $\mathbf{w} \to (\bar{\mathbf{y}} - \bar{\mathbf{x}})'\mathbf{w}$ asymptotically (as $d \to \infty$) lead to the same ranking among the $n - m$ projected observations of the second subsample. Since $\mathbf{w}_1'(\bar{\mathbf{y}} - \bar{\mathbf{x}}) > \mathbf{w}_2'(\bar{\mathbf{y}} - \bar{\mathbf{x}}) \Leftrightarrow \|\mathbf{w}_1 - \bar{\mathbf{x}}\|^2 - \|\mathbf{w}_1 - \bar{\mathbf{y}}\|^2 > \|\mathbf{w}_2 - \bar{\mathbf{x}}\|^2 - \|\mathbf{w}_2 - \bar{\mathbf{y}}\|^2$, the transformation $\mathbf{w} \to \|\mathbf{w} - \bar{\mathbf{x}}\|^2 - \|\mathbf{w} - \bar{\mathbf{y}}\|^2$ also leads to the same ranking.

Now, for any $\mathbf{x}_i$ from the subsample of size m_1, $\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 - \|\mathbf{x}_i - \bar{\mathbf{y}}\|^2 \xrightarrow{P} -(\sigma_1^2/m_1 + \sigma_2^2/m_2 + \mu^2)$, and for any $\mathbf{y}_i$ from the subsample of size m_2, $\|\mathbf{y}_i - \bar{\mathbf{x}}\|^2 - \|\mathbf{y}_i - \bar{\mathbf{y}}\|^2 \xrightarrow{P} (\sigma_1^2/m_1 + \sigma_2^2/m_2 + \mu^2)$.

Therefore, after finding $\widehat{\boldsymbol{\beta}}$ using SVM or DWD, we consider the one-sided alternative H_1 that suggests $G_{\widehat{\boldsymbol{\beta}}}$ to be stochastically larger than $F_{\widehat{\boldsymbol{\beta}}}$.

Consider now $T_{\widehat{\boldsymbol{\beta}}}$, which is chosen to be the one sided KS statistic or the one sided WMW statistic, and it is computed from the two subsamples consisting of $n_1 - m_1$ observations on $\mathbf{x}$ and $n_2 - m_2$ observations on $\mathbf{y}$ using the ranks of those observations projected along $\widehat{\boldsymbol{\beta}}$, where ranking is done after combining the two subsamples. Now, for each of these $n_1 - m_1$ observations on $\mathbf{x}$, we have $\|\mathbf{x} - \bar{\mathbf{x}}\|^2 - \|\mathbf{x} - \bar{\mathbf{y}}\|^2 \xrightarrow{P} (\sigma_1^2/m_1 - \sigma_2^2/m_2) - \mu^2$ and each of the $n_2 - m_2$ observations on $\mathbf{y}$, we have $\|\mathbf{y} - \bar{\mathbf{x}}\|^2 - \|\mathbf{y} - \bar{\mathbf{y}}\|^2 \xrightarrow{P} (\sigma_1^2/m_1 - \sigma_2^2/m_2) + \mu^2$. So, when $\mu^2 > 0$, $T_{\widehat{\boldsymbol{\beta}}}$ attains it maximum value $t_{\max}$. Since $P_{H_0}(T_{\widehat{\boldsymbol{\beta}}} = t_{\max}) < \alpha$, occurrence of such an event implies that $\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}}) = 1$. Since any test function is bounded, this proves that $E_{H_A}\{\phi_\alpha(T_{\widehat{\boldsymbol{\beta}}})\} \to 1$, and consequently $E_{H_A}(\phi_\alpha^*) \to 1$ as $d \to \infty$.

It is appropriate to mention here that not only for one sided KS or one sided WMW statistic, the above result holds for any one sided linear rank statistic (see e.g., Hajek, Sidak and Sen, 1999) of the form $\sum_{i=1}^{m_1} a(R_i)$, where the R_i's are the rank of the projected observations on $\mathbf{x}$ in the combined sample, and a is a monotonically increasing function. Also, in view of the results in Hall et. al. (2005), the convergence of the powers of our tests to one actually holds even when both d and $(m_1 + m_2)$ grow to infinity in such a way that $(m_1 + m_2)/d^2$ tends to zero. One should notice that depending on the values of σ_1^2, σ_2^2 and μ^2, both SVM and DWD need some additional conditions on m_1 and m_2 for perfect classification of future observations, otherwise they classify all observations to a single class (see Hall et. al., 2005). But, for our tests based on SVM and DWD directions, we do not need such conditions for the convergence of the power function to 1. Note also that the condition $\mu^2 > 0$ holds in the commonly used set up for two-sample testing problems, where the population distributions are assumed to have the same dispersion but different means. For the one sided KS statistics as well as any one sided linear rank statistic (as mentioned above), it is easy to see that $T_{\widehat{\boldsymbol{\beta}}}$ will take its maximum value $t_{\max}$ if and only if, in the combined sample, the rank of the linear function of any observation from F is smaller than the rank of the linear function of any observation from G. Hence, $P_{H_0}(T_{\widehat{\boldsymbol{\beta}}} = t_{\max}) = (n_1 - m_1)!(n_2 - m_2)!/(n_1 + n_2 - m_1 - m_2)!$, which will be smaller than α if $(n_1 + n_2 - m_1 - m_2)$ is suitably large.

3.2 Results from the analysis of simulated data sets

We carried out simulation studies to compare the power properties of our proposed tests with Hotelling's T^2, spatial sign and rank tests (see e.g., Mottonen, 1995; Choi and Marden, 1997), Puri

and Sen's (PS) (1971) co-ordinate wise sign and rank tests. For spatial sign and rank tests, we have reported (see Table 1) the results for the conditional tests based on the permutation distributions since their empirical performance was better than the tests based on large sample asymptotics. The codes for these tests are available in MNM (see Oja, 2010 for details) and other packages in R. In addition to these rank based tests, there are some two-sample multivariate tests based on inter-point distances available in the literature, which are conveniently applicable to high dimensional data even when the dimension is larger than the sample size. We have included four such tests in our empirical study. Among these tests, the codes for the test based on nearest neighbor (NN) type coincidences (see e.g., Schilling, 1986; Henze, 1988) are available in the R package 'MTSKNN', and those for the Cramer test (see Baringhaus and Franz, 2004) are available in the R package 'cramer'. For the multivariate run test (see Friedman and Rafsky, 1979) based on minimal spanning tree (MST) and Rosenbaum's test best based on adjacency (we used the Euclidean metric for distance computation), we used our own codes. For the test based on NN type coincidences, we used the test based on three nearest neighbors, which has been reported to perform well in the literature. Chen and Qin (2010) also proposed a test based on a Hotelling's T^2 type statistic that can be used even when the dimension exceeds the sample size. We have also considered it for our simulation study. However, because of computational difficulty in high dimensions, we could not use the two-sample tests proposed in Randles and Peters (1990), Hettmansperger and Oja (1994), Hettmansperger et. al. (1998) and Liu and Singh (1993).

For our tests, we used SVM and DWD for estimating $\boldsymbol{\beta}$, and for $T_{\boldsymbol{\beta}}$, we used the one sided KS and the one sided WMW statistics. In all cases, 50% of the observations were used to carry out the classification procedure to construct the estimate for $\boldsymbol{\beta}$, and the remaining 50% of the data points were used to compute the rank based statistic and the test function. Similar random half-spilt of the whole sample into two subsamples for high dimensional low sample size data was also considered in Yata and Aoshima (2010). The test function ϕ^* was computed by averaging the test functions obtained from 50 random splits. For SVM, we used the R program 'svmpath' (see Hastie *et. al.*, 2004), which automatically selects the regularization parameter to be used for classification. For DWD, we used the MATLAB codes of Marron *et. al.* (2007) with the default penalty function.

We generated 50 observations for each of the two samples from d-dimensional standard multivariate normal, t with 2 d.f. (denoted as t_2) and Cauchy distributions, and we have considerd $d = 30$, 60 and 90. The reason for choosing these distributions was to consider different distributions with varying degrees of heaviness of their tails. Multivariate normal distributions have all their moments finite, multivariate t distributions with 2 degrees of freedom have finite first order

Table 1: Observed powers of two-sample tests with 5% nominal level

	Normal			t with 2 d.f.			Cauchy		
	$d = 30$	$d = 60$	$d = 90$	$d = 30$	$d = 60$	$d = 90$	$d = 30$	$d = 60$	$d = 90$
	$\Delta = 1.5$	$\Delta = 1.75$	$\Delta = 2.0$	$\Delta = 1.5$	$\Delta = 1.75$	$\Delta = 2.0$	$\Delta = 1.5$	$\Delta = 1.75$	$\Delta = 2.0$
Hotelling's T^2	0.987	0.887	0.350	0.682	0.651	0.315	0.290	0.467	0.297
Chen-Qin	0.993	0.997	1.000	0.117	0.056	0.031	0.002	0.000	0.000
spatial-sign	0.986	0.894	0.375	0.905	0.659	0.232	0.737	0.528	0.174
spatial-rank	0.984	0.890	0.362	0.857	0.707	0.310	0.636	0.609	0.271
PS-sign	0.424	0.031	0.000	0.254	0.014	0.000	0.149	0.009	0.000
PS-rank	0.726	0.087	0.000	0.436	0.040	0.000	0.220	0.014	0.000
NN type coin.	0.751	0.775	0.818	0.819	0.765	0.766	0.756	0.710	0.660
Multivar. run	0.526	0.526	0.588	0.640	0.640	0.677	0.563	0.596	0.631
Cramer	0.986	0.982	0.991	0.238	0.094	0.049	0.006	0.005	0.005
Adjacency	0.456	0.428	0.503	0.322	0.311	0.334	0.259	0.252	0.275
WMW-SVM	0.803	0.880	0.939	0.569	0.690	0.752	0.389	0.487	0.577
WMW-DWD	0.937	0.955	0.984	0.672	0.722	0.786	0.399	0.518	0.582
KS-SVM	0.736	0.831	0.896	0.570	0.678	0.758	0.432	0.543	0.635
KS-DWD	0.908	0.926	0.957	0.671	0.733	0.779	0.448	0.564	0.646

moments but their higher order moments are not finite, and multivariate Cauchy distributions do not have finite first order moments. Here, F and G were chosen to be spherically symmetric (with the common scatter scatter matrix $\mathbf{I}$), and they differ only in their locations. Note at this point that our tests are invariant under a common location shift and a common orthogonal transformation of the data in the two samples in view of the invariance property of the classification methods SVM and DWD under location shifts and orthogonal transformations of the data. For any two-sample multivariate test, which is invariant under location change and orthogonal transformation of the data in the two samples, its power is a function of the norm of the difference between the locations of the spherically symmetric distributions F and G. We chose F to be symmetric around the origin and G to be symmetric around the point $(\Delta, 0, \ldots, 0)$. The value of Δ was chosen depending on the problem (see Table 1) such that all competing tests had powers appreciably different from one as well as 5%, the nominal level. In each of these cases, we carried out 1000 Monte-Carlo experiments, and for each test, we computed the proportion of times it rejected H_0. Observed powers for nominal 5% tests are reported in Table 1, which clearly show the superior or comparable performance of our tests when compared with other tests for high dimensional data in many situations. The overall performance of the tests based on DWD was better than those based on SVM.

3.3 Results from the analysis of real data sets

We analyzed three real data sets for further evaluation of our proposed methods. The sonar data set, the hill and valley data set and their descriptions are available at the UCI machine learning repository. The colon data set is available in the R package 'rda'. Descriptions of this microarray gene expression data set can be found in Alon *et. al.* (1999). Several researchers have extensively investigated these data sets, mainly in the context of classification. It is well known that in all these examples, we have reasonable separability between two competing classes. So, in each of these cases, we can assume the alternative hypothesis to be the true, and different tests can be compared on the basis of their power functions.

Note that if we use the whole data set for testing, any test will either reject H_0 or accept it. Based on that single experiment, it is difficult to compare among different test procedures. So, in each of these cases, we repeated the experiment 1000 times based on 1000 different subsets chosen from the data, and the results are reported in Table 2. Because of the problem in inverting the estimated dispersion matrix (due to its singularity) of co-ordinatewise ranks, PS tests could not be used in these real data sets.

Sonar data consist of 208 sixty-dimensional observations from two different classes 'Metal cylinders' and 'Rocks'. Taking equal number of observations from these two classes, we randomly generated our sample and used that to test $H_0 : F = G$. We considered samples of three different sizes $n = 100$, 70 and 40. In this example, the test based on NN coincidences had the best performance. For $n = 100$, all tests except the adjacency test had competitive performance, but in the case of $n = 40$, when the sample size was smaller than the dimension, our proposed tests based on SVM and the test based on NN type coincidences outperformed all other tests. It is to be noted here that in the case of $n = 40$, the Hotelling's T^2 test and the tests based on spatial ranks could not be used.

Next, we consider the 'hill and valley' dataset, which contains 100 dimensional observations from two different classes 'hill' and 'valley'. Here we used samples of two different sizes $n = 200$ and 100. Due to the problem in inverting a high dimensional matrix, spatial sign and rank tests could not be used in many occasions. So, here we could not report the performance of these methods. Hotelling's T^2 method could be used only in the case of $n = 200$, but it rejected H_0 only 4.2% cases. The performance of our proposed tests was substantially better, and they rejected H_0 in almost all cases. Even in the case of $n = 100$, they had excellent performance. Other two-sample tests that we have considered in this article had poor performance in this data set. We observed similar phenomenon also in the case of noisy version of this data set available in the UCI machine learning repository,

Table 2: Observed Powers of two-sample tests in real data sets

	Sonar			Hill & Valley		Colon
	$n = 100$	$n = 70$	$n = 40$	$n = 200$	$n = 100$	$n = 40$
	$d = 60$	$d = 60$	$d = 60$	$d = 100$	$d = 100$	$d = 2000$
Hotelling T^2	0.994	0.427	—	0.042	—	—
Chen-Qin	0.970	0.754	0.330	0.002	0.004	1.000
spatial-sign	0.993	0.426	—	—	—	—
spatial-rank	0.991	0.429	—	—	—	—
NN type coin.	1.000	0.996	0.711	0.089	0.059	1.000
Multivar. run	1.000	0.969	0.461	0.123	0.075	0.969
Cramer	0.999	0.855	0.339	0.061	0.055	1.000
Adjacency	0.429	0.452	0.188	0.034	0.039	0.769
WMW-SVM	0.988	0.931	0.709	0.998	0.943	0.982
WMW-DWD	0.986	0.898	0.601	1.000	0.932	0.988
KS-SVM	0.978	0.905	0.672	0.998	0.941	0.971
KS-DWD	0.974	0.869	0.562	1.000	0.922	0.983

but those have not been reported here.

The colon data set contains 62 microarray sequences each of length 2000. In this case, we randomly chose 20 observations from each class (tumor and normal tissues) to form the samples. Since the dimension of the data set was bigger than its size, the Hotelling's T^2 test and the tests based on spatial ranks could not be used. The adjacency test rejected the null hypothesis in 76.9% of the cases. But the powers of our proposed tests and those of other two-sample tests were much higher.

4 Tests for high dimensional matched pair data

Instead of having two independent sets of observations from F and G, we can have n observations $\binom{\mathbf{x}_1}{\mathbf{y}_1}, \binom{\mathbf{x}_2}{\mathbf{y}_2}, \ldots, \binom{\mathbf{x}_n}{\mathbf{y}_n}$ from a $2d$-variate distribution with d-dimensional marginals F and G for $\mathbf{x}$ and $\mathbf{y}$. In such cases, it is a common practice to consider it as a one-sample problem, where $\{\boldsymbol{\xi}_i = \mathbf{x}_i - \mathbf{y}_i;\ i = 1, 2, \ldots, n\}$ are used as the sample observations. Here, we assume that the distribution of $\boldsymbol{\xi}$ is symmetric about $\boldsymbol{\mu}$ and test $H_0 : \boldsymbol{\mu} = 0$ against $H_1 : \boldsymbol{\mu} \neq 0$. Note that if $F(\mathbf{x} + \boldsymbol{\mu}_1) = G(\mathbf{x} + \boldsymbol{\mu}_2)$ for all $\mathbf{x}$, the distribution of $\boldsymbol{\xi}$ is symmetric about $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and testing the equality of the locations of F and G is equivalent to test $H_0 : \boldsymbol{\mu} = 0$. In such situations, one needs to develop multivariate versions of sign and signed rank tests. Several multivariate generalizations of such tests have been proposed in the literature (see e.g., Puri and Sen, 1971; Randles, 1989; Chaudhuri and Sengupta, 1993; Mottonen, Oja and Tienari, 1997; Hallin and Paindaveine, 2002),

and for any fixed dimension, their large sample properties have also been well investigated. A brief overview of these methods can be found in Oja (2010). However, these tests may not yield good results in high dimensional problems and they cannot be used when the dimension exceeds the sample size. In this section, we will deal with these high dimensional problems. There are some multivariate generalizations of the sign test, which are distribution-free when the underlying distributions are elliptic (see e.g., Randles, 1989; Chaudhuri and Sengupta, 1993). But, we do not have such property for any existing multivariate generalization of the signed rank test. Here we propose some methods based on linear projection of observations that lead to multivariate generalizations of univariate sign, signed rank and other one-sample linear rank tests (see e.g., Hajek, Sidak and Sen, 1999), which are distribution-free not only under elliptic set up, but for all continuous distributions, and they can be conveniently used even when the dimension of the data is much larger than the sample size.

Here also, we split the whole sample into two subsamples. The first subsample of size m is used to estimate the projection direction. For this estimation, we consider a classification problem between two data clouds $\{\boldsymbol{\xi}_i,\ i = 1, 2, \ldots, m\}$ and $\{\boldsymbol{\eta}_i = -\boldsymbol{\xi}_i,\ i = 1, 2, \ldots, m\}$, and use SVM or DWD to find the separating hyperplane. The direction vector perpendicular to this hyperplane is used as $\widehat{\boldsymbol{\beta}}$. Note that if the distribution of $\boldsymbol{\xi}$ is elliptically symmetric (or the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ is elliptically symmetric), the separating hyperplane $\{\mathbf{w} : \beta_0 + \boldsymbol{\beta}^{'}\mathbf{w} = 0\}$ leads to the best classification between the distributions of $\boldsymbol{\xi}$ and that of $\boldsymbol{\eta}$ if and only if the expected values of the one sided univariate sign and signed rank statistic are maximized when the observations are projected along $\boldsymbol{\beta}$ (follows from arguments similar to that used in the proof of Proposition 2.1). Motivations for this classification approach also follow from the interesting result given below.

PROPOSITION 4.1. *Suppose that* Π *is an elliptically symmetric multivariate distribution having non-zero location. Then, the sign and the signed rank based on linear projections of the data will have maximum power if and only if the observations are projected along the direction vector of the Bayes discriminating hyperplane associated with the classification problem involving distributions* Π *and* Π^* *with equal prior probabilities, where* $\boldsymbol{\xi} \sim \Pi^*$ *if and only if* $-\boldsymbol{\xi} \sim \Pi$.

PROOF. If Π is elliptically symmetric with a non-zero location, so is Π^*, and it differs from Π only its location. So, the above result can be proved using the arguments based on stochastic ordering as we used in the proof of Proposition 2.1.

After finding $\widehat{\boldsymbol{\beta}}$ using SVM or DWD, the $n-m$ observations in the second subsample are projected

along $\widehat{\boldsymbol{\beta}}$ to compute the univariate test statistic (i.e., sign or signed rank statistic). The test function is constructed as before, and ϕ^*_α is obtained by averaging the test functions corresponding to different random splits. Following the same argument as used in the proof of Theorem 2.2, one can verify that the proposed tests are distribution-free, and ϕ^*_α is monotone with respect to α in the sense that $\alpha \geq \alpha'$ implies $\phi^*_\alpha \geq \phi^*_{\alpha'}$.

Now, we carry out a theoretical investigation on the power properties of our proposed tests. Here also, we look at $\boldsymbol{\xi} = (\xi^{(1)}, \xi^{(2)}, \ldots, \xi^{(d)})$ as a truncated version of the infinite sequence $(\xi^{(1)}, \xi^{(2)}, \ldots)$, and study the behavior of the power function as the dimension increases. We assume the conditions stated in (A1) and (A3). However, instead of (A2), we consider the following assumption.

(A4) *The sequence* $\{\xi^{(k)2};\ k \geq 1\}$ *has the* ρ*-mixing property. If* $(\xi'^{(1)}, \xi'^{(2)}, \ldots)$ *is an independent copy of* $(\xi^{(1)}, \xi^{(2)}, \ldots)$*, the* ρ*-mixing property also holds for the sequences* $\{(\xi^{(k)} - \xi'^{(k)})^2;\ k \geq 1\}$ *and* $\{(\xi^{(k)} + \xi'^{(k)})^2;\ k \geq 1\}$.

Note that Hall et. al. (2005) considered an assumption on ρ-mixing properties for functions dominated by quadratics. Both (A2) and (A4) hold under that condition. The following theorem shows that under these regularity conditions, the power of these sign and signed rank tests converges to 1 as d tends to infinity.

THEOREM 4.2. *Assume the conditions stated in (A1), (A3) and (A4). Also assume that the univariate distribution-free test statistic T (e.g., sign or signed rank statistic), when it is computed based on observations in the second sample, under* H_0*, it takes its maximum value with probability smaller than* α *(i.e., the size of the second subsample is not too small). Then the powers of the proposed sign and signed rank tests based on SVM and DWD converge to 1 as the dimension d diverges to infinity.*

PROOF. Under (A3), $\frac{1}{d}\sum_{i=1}^{d} Var(\xi^{(i)}) \to \sigma_1^2 + \sigma_2^2 = \sigma^2$, say and $\frac{1}{d}\sum_{i=1}^{d}[E(\xi^{(i)})]^2 \to \mu^2 > 0$ as $d \to \infty$. Now, under (A1) and (A4), WLLN holds for the sequence of $\{\xi^{(1)2}, \xi^{(2)2} \ldots\}$, and using (A3), one can show that $d^{-1/2}\|\boldsymbol{\xi}_i\| = d^{-1/2}\|\boldsymbol{\eta}_i\| \xrightarrow{P} (\mu^2 + \sigma^2)^{1/2}$ as $d \to \infty$ (here $\boldsymbol{\eta}_i = -\boldsymbol{\xi}_i$ for all i) for all $i = 1, 2, \ldots, m$. This implies $d^{-1/2}\|\boldsymbol{\xi}_i - \boldsymbol{\eta}_i\| \xrightarrow{P} 2(\mu^2 + \sigma^2)^{1/2}$ for $i = 1, 2, \ldots, m$. Again, for any $i \neq j$, we have $d^{-1/2}\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\| = d^{-1/2}\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_j\| \xrightarrow{P} (2\sigma^2)^{1/2}$ and $d^{-1/2}\|\boldsymbol{\xi}_i - \boldsymbol{\eta}_j\| \xrightarrow{P} (2\sigma^2 + 4\mu^2)^{1/2}$. So, as d tends to infinity, after re-scaling by a factor of $d^{-1/2}$, $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_m$ tend to lie on the vertices of a regular m-simplex $\mathcal{S}_1$ and $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_m$ tend to lie on the the vertices of another regular m-simplex $\mathcal{S}_2$, which is obtained from $\mathcal{S}_1$ by reflecting it using all the co-ordinate axes.

Define $\delta_1 = (2\sigma^2 + 4\mu^2)^{1/2}$, $\delta_2 = 2(\mu^2 + \sigma^2)^{1/2}$, and note that $\delta_1 < \delta_2$. Now, for any vertex $\boldsymbol{\xi}_i$ on

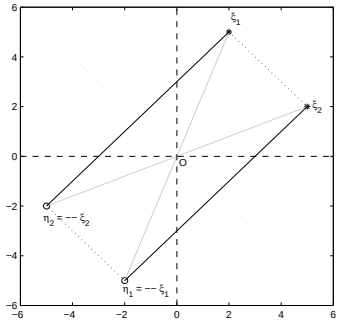

Figure 1: Geometry of high dimensional data.

$\mathcal{S}_1$, the re-scaled distances of the $(m-1)$ vertices of $\mathcal{S}_2$ are equal to δ_1 and that of one other vertex $\boldsymbol{\eta}_i$ is δ_2. This is easy to visualize for $m = 2$ (see Figure 1). Naturally, SVM chooses the hyperplane that passes through the origin and bisects each of these lines of length δ_1. Now, consider a new observation $\boldsymbol{\xi}_0$. As d tends to infinity, after rescaling by a factor of $d^{-1/2}$, it tends to be equi-distant from all vertices of $\mathcal{S}_1$, and that common distance is $(2\sigma^2)^{1/2}$. So, its squared distance from the centroid of the first set of observations is given by $2\sigma^2 - \sigma^2(1 - m^{-1})$. Similarly, its (re-scaled) distances from all vertices of $\mathcal{S}_2$ tend to be δ_1. Hence, its squared distance from the centroid of $\mathcal{S}_2$ turns out to be $\delta_1^2 - \sigma^2(1 - m^{-1})$. So, SVM classifies $\boldsymbol{\xi}_0$ to the correct population if $2\sigma^2 < \delta_1^2$ i.e. $\mu^2 > 0$. Also note that here we have same number of observations in each of the two classes. So, SVM and DWD have the same limiting behavior as $d \to \infty$ (see Hall et. al., 2005).

Let $\widehat{\boldsymbol{\beta}}$ be the direction perpendicular to the separating hyperplane chosen by SVM or DWD. Now, from the above discussion it is clear that $P\{\widehat{\boldsymbol{\beta}}^{'}\boldsymbol{\xi} > 0\} = P\{\widehat{\boldsymbol{\beta}}^{'}(\mathbf{x} - \mathbf{y}) > 0\} \to 1$ as $d \to \infty$. Now, using the same argument as used in the proof of Theorem 3.1, we can show that the powers of sign and signed rank tests converge to 1 as d tends to infinity.

We carried out simulation studies to compare the power properties of our proposed tests based on sign and signed rank statistics with other existing methods. In particular, we used Hotelling's T^2 test, Chen and Qin's (2010) test, tests based on coordinate-wise signs and signed ranks (see e.g., Puri and Sen, 1971), spatial sign and signed rank tests (see e.g., Mottonen, Oja and Tienari, 1997; Oja, 2010) and Hallin and Paindaveine's (HP) (2002) tests based on inter-directions and pseudo Mahalanobis distances. For HP tests and the tests based on spatial signs and ranks, here we have reported the results of the conditional tests based on the permutation principle since they performed better than the corresponding large sample tests in almost all cases. For our proposed methods based on DWD, we used the MATLAB codes as before, but due to singularity of matrices in the regularization path of SVM, the R program 'svmpath' could not be used. Instead we used the

Table 3: Observed powers of paired sample tests with 5% nominal level

	Normal			t with 2 d.f.			Cauchy		
	$p = 30$	$p = 60$	$p = 90$	$p = 30$	$p = 60$	$p = 90$	$p = 30$	$p = 60$	$p = 90$
	$\Delta = 0.75$	$\Delta = 1$	$\Delta = 1.25$	$\Delta = 0.75$	$\Delta = 1$	$\Delta = 1.25$	$\Delta = 0.75$	$\Delta = 1$	$\Delta = 1.25$
Hotelling's T^2	0.986	0.986	0.603	0.684	0.835	0.551	0.296	0.590	0.497
ChenQin	0.999	1.000	1.000	0.110	0.097	0.100	0.002	0.001	0.000
spatial-sign	0.984	0.984	0.622	0.877	0.841	0.366	0.771	0.647	0.237
spatial-rank	0.986	0.985	0.617	0.840	0.884	0.509	0.650	0.730	0.447
coord.-sign	0.816	0.487	0.000	0.559	0.166	0.000	0.385	0.053	0.000
coord.-SR	0.964	0.694	0.000	0.652	0.282	0.000	0.376	0.102	0.000
HP-sign	0.982	0.979	—	0.868	0.806	—	0.756	0.617	—
HP-rank	0.869	0.692	—	0.303	0.216	—	0.178	0.110	—
Sign-SVM	0.758	0.799	0.824	0.567	0.716	0.821	0.348	0.541	0.677
Sign-DWD	0.871	0.971	1.000	0.654	0.856	0.963	0.498	0.758	0.904
SR-SVM	0.888	0.905	0.926	0.576	0.712	0.825	0.309	0.467	0.600
SR-DWD	0.950	0.993	1.000	0.652	0.850	0.961	0.425	0.661	0.817

SVM toolbox in MATLAB, where the regularization parameter was chosen based on a pilot study.

For our simulation study, we considered examples with high dimensional (d= 30, 60 and 90) normal, t_2 and Cauchy distributions as before with $(0, 0, \ldots, 0)$ and $(\Delta, 0, \ldots, 0)$ as the locations for $\mathbf{x}$ and $\mathbf{y}$, respectively. We chose the value of Δ depending on the problem (Δ=0.75, 1.0, 1.25 for d=30, 60 and 90, respectively) such that almost all competing methods had power appreciably different from 1 as well as 0.05. In all these testing problems, we chose $Var(\mathbf{x}) = Var(\mathbf{y}) = 0.5\ \mathbf{I} + 0.5\ \mathbf{1}\mathbf{1}'$ and $Cov(\mathbf{x}, \mathbf{y}) = 0.5\ \mathbf{1}\mathbf{1}'$, where $\mathbf{1} = (1, 1, \ldots, 1)'$. We generated 100 observations from the joint distribution of $\mathbf{x}$ and $\mathbf{y}$ to constitute the sample, and each experiment is repeated 1000 times as before. We could not use HP tests for $d = 90$ because of the problem in convergence of Tyler's shape matrix. The overall performance of our proposed tests, particularly those based on DWD, was better than most of the existing methods, especially in cases of $d = 90$. For $d = 60$, if not better, spatial sign and rank tests also had competitive performance but for $d = 90$, our proposed methods clearly outperformed all other tests considered here. For instance, in the case of 90-dimensional normal distributions, while all other tests rejected H_0 nearly 60% of the cases, our multivariate versions of sign and signed rank tests based on DWD rejected H_0 in all the 1000 cases. We observed similar phenomenon also in the case of 90-dimensional t_2 and Cauchy distributions, where our proposed tests had much higher powers than their competitors.

5 Large sample properties of proposed tests when the dimension is fixed

So, far we have proved and demonstrated the exact distribution-free property of our proposed tests and we have also shown the convergence of their power function when the sample size is fixed and the dimension grows to infinity. In this section, we will study their power properties when the sample grows to infinity and the dimension of the data is not large. For studying this large sample properties of the proposed two-sample tests based on SVM and DWD, we assume that as the first subsample size m tends to infinity, m_1/m converges to $1/2$. Otherwise, one has to make some adjustment for the unbalancedness in the data (see e.g., Qiao et. al., 2010). However, if the dimension of the data is not large, especially relative to the sample size, in addition to SVM and DWD, there are many other ways to choose the projection direction $\boldsymbol{\beta}$. Unlike what happens for high dimensional data, we can use very simple classifiers like Fisher's linear discriminant rule to estimate $\boldsymbol{\beta}$. Alternatively, if $T_{\boldsymbol{\beta}}$ is the univariate KS or WMW statistic computed using the observations in the first subsample projected along $\boldsymbol{\beta}$, one can also find $\widehat{\boldsymbol{\beta}}$ by maximizing $T_{\boldsymbol{\beta}}$ over the set $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| = 1\}$. In cases of KS and WMW statistics, this maximization leads to linear classifiers based on regression depths and half-space depths, respectively (see, e.g., Ghosh and Chaudhuri, 2005). Clearly, the finite sample distribution-free property established in Theorem 2.2 will remain valid irrespective of the classification procedure used to construct $\widehat{\boldsymbol{\beta}}$ so long as it is computed from one subsample and the univariate distribution-free tests are implemented on linear projections of the data points in the other subsample. Note that in the case of two-sample multivariate location problems, where $F(\mathbf{x}+\boldsymbol{\mu}_1) = G(\mathbf{x}+\boldsymbol{\mu}_2)$ with $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, if $\widehat{\boldsymbol{\beta}} \notin Q = \{\boldsymbol{\beta} : \boldsymbol{\beta}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0\}$, the power of the univariate test (WMW or KS test) applied on the projected observations converges to 1 as the size of the second subsample tends to infinity. For instance, if the distribution of $\widehat{\boldsymbol{\beta}}$ is absolutely continuous, we have consistency of the resulting tests because the set Q has Lebesgue measure zero. Further, even in the case of general alternative $H_A : F \neq G$, if F and G satisfy Carleman condition, (i.e., $\mu_r = E(\|\mathbf{x}\|^r) < \infty \ \forall \ r \geq 1$ and $\sum_{r \geq 1} \mu_r^{-1/r} = \infty$), the set $Q_0 = \{\boldsymbol{\beta} : F_{\boldsymbol{\beta}} = G_{\boldsymbol{\beta}}\}$ has Lebesgue measure 0 (see Corollary 3.3 of Cuesta-Albertos, Fraiman and Ransford, 2007), and consequently, the power of our test constructed using the KS statistic converges to 1 as the size of the second subsample tends to infinity.

Suppose that F and G are elliptically symmetric, and they differ only in their locations $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. From Proposition 2.1, we know that in this case, KS, WMW or any other standard linear rank test for location based on linear projections of the data will have the maximum power if and only

if the observations are projected along $\boldsymbol{\beta}_* = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, where $\boldsymbol{\Sigma}$ is the common scatter matrix of F and G. Let $\widehat{\boldsymbol{\beta}}_D, \widehat{\boldsymbol{\beta}}_S, \widehat{\boldsymbol{\beta}}_F$ and $\widehat{\boldsymbol{\beta}}_M$ be the estimates of $\boldsymbol{\beta}$ obtained from the first subsample using DWD, SVM, Fisher's linear discrimination and maximization of $T_{\boldsymbol{\beta}}$ as described at the beginning of the section, respectively. Then, we have the following theorem that provides useful insights into asymptotic power properties of the proposed multivariate two-sample tests.

THEOREM 5.1 *For a fixed size of the second subsample, let $\gamma(\boldsymbol{\beta})$ be the power of the univariate KS (or WMW) test when the observations in the second subsample are projected along $\boldsymbol{\beta}$. Define $\gamma_0 = \sup_{\boldsymbol{\beta}} \gamma(\boldsymbol{\beta})$. If F and G are elliptically symmetric, and they differ only in their locations, $\gamma(\widehat{\boldsymbol{\beta}}_M)$ converges to γ_0 as the first subsample size m tends to infinity. If F and G have finite second order moments, we also have this convergence for $\gamma(\widehat{\boldsymbol{\beta}}_F)$, $\gamma(\widehat{\boldsymbol{\beta}}_D)$ and $\gamma(\widehat{\boldsymbol{\beta}}_S)$ when the regularization parameter λ used in SVM is of the order $o(m^{-1/2})$. Under the same set of conditions, the powers of these tests converge to 1 if the sizes of both of the first and the second subsamples tend to infinity.*

PROOF. If we can show the continuity of $\gamma(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we can say that $\exists\ \boldsymbol{\beta}_*$, such that $\gamma_0 = \gamma(\boldsymbol{\beta}_*)$. Then, the convergence of the power function to γ_0 follows from Lemma 3 (see Appendix). Consider any fixed sequence $\{\boldsymbol{\beta}_m, m \geq 1\}$ that converges to $\boldsymbol{\beta}$. So, for any fixed size of the second subsample, $(\boldsymbol{\beta}_m^{'}\mathbf{x}_1, \ldots, \boldsymbol{\beta}_m^{'}\mathbf{x}_{n_1-m_1}, \boldsymbol{\beta}_m^{'}\mathbf{y}_1, \ldots, \boldsymbol{\beta}_m^{'}\mathbf{y}_{n_2-m_2})$ converges to $(\boldsymbol{\beta}^{'}\mathbf{x}_1, \ldots, \boldsymbol{\beta}^{'}\mathbf{x}_{n_1-m_1}, \boldsymbol{\beta}^{'}\mathbf{y}_1, \ldots, \boldsymbol{\beta}^{'}$ almost surely and hence in distribution. Now, note that $Q_{\boldsymbol{\beta}} = \{(\mathbf{x}, \mathbf{y}) : \boldsymbol{\beta}^{'}(\mathbf{x} - \mathbf{y}) > 0\}$ is an open set in R^d with boundary having probability measure zero. Also, for any fixed $(\mathbf{x}, \mathbf{y})$, the set $Q^{\mathbf{x},\mathbf{y}} = \{\boldsymbol{\beta} : \boldsymbol{\beta}^{'}(\mathbf{x} - \mathbf{y}) > 0\}$ is open in R^d. Since $\boldsymbol{\beta}_m^{'}(\mathbf{x} - \mathbf{y}) \to \boldsymbol{\beta}^{'}(\mathbf{x} - \mathbf{y})$, for any $(\mathbf{x}, \mathbf{y}) \in Q_{\boldsymbol{\beta}}$, we have $\boldsymbol{\beta}_m^{'}(\mathbf{x} - \mathbf{y}) > 0$ for sufficiently large m. If $R(\boldsymbol{\beta})$ denotes the value of a univariate rank statistic (e.g., KS or WMW statistic) computed using the observations projected along $\boldsymbol{\beta}$, the event $\{R(\boldsymbol{\beta}) = r\}$ can be expressed in terms of finite unions and intersections of the sets $Q_{\boldsymbol{\beta}}^{ij} = \{(\mathbf{x}_i, \mathbf{y}_j) : \boldsymbol{\beta}^{'}(\mathbf{x}_i - \mathbf{y}_j) > 0\};\ 1 \leq i \leq n_1 - m_1,\ 1 \leq j \leq n_2 - m_2$. So, $P\{R(\boldsymbol{\beta}_m) = r\} \to P\{R(\boldsymbol{\beta}) = r\}$ for all r, and hence we have the continuity of $\gamma(\boldsymbol{\beta})$.

Since $\boldsymbol{\beta}_m$ converges to $\boldsymbol{\beta}_*$, and $\boldsymbol{\beta}_*^{'}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \neq 0$, for any $\epsilon > 0$, one can choose a sufficiently large M such that for all $m \geq M$, $P(\boldsymbol{\beta}_m \notin Q) > 1 - \epsilon$, where $Q = \{\boldsymbol{\beta} :\ \boldsymbol{\beta}^{'}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0\}$. Now, if $\boldsymbol{\beta}_m \notin Q$, the power of the test based on WMW and KS statistic converges to 1 as the size of the second subsample tends to infinity. So, as the sizes of the first and the second both tend to infinity, the powers of these tests convergence to 1.

In the case of paired sample problems, we may consider the situation when the $2d$-dimensional joint distribution of $(\mathbf{x}, \mathbf{y})$ is elliptically symmetric and the corresponding d-dimensional marginals

of $\mathbf{x}$ and $\mathbf{y}$ (F and G, respectively) differ only in their locations. Otherwise, we may assume the distribution of $\boldsymbol{\xi} = \mathbf{x} - \mathbf{y}$ to be symmetric about a non-zero location. In that case, we can either find $\widehat{\boldsymbol{\beta}}$ using classification methods like SVM and DWD applied to $\boldsymbol{\xi}$'s and $\boldsymbol{\eta}$'s as described in the previous section. As an alternative, one can also construct $\widehat{\boldsymbol{\beta}}$ using Fisher's linear discrimination or maximization of univariate sign $SG_m(\boldsymbol{\beta}) = \frac{1}{m}\sum_{i=1}^{m} I\{\boldsymbol{\beta}'(\mathbf{x}_i - \mathbf{y}_i) > 0\}$ or signed rank statistic $SR_m(\boldsymbol{\beta}) = \sum_{i=1}^{m}\sum_{j=i+1}^{m} I\{\boldsymbol{\beta}'(\mathbf{x}_i - \mathbf{y}_i) + \boldsymbol{\beta}'(\mathbf{x}_j - \mathbf{y}_j) > 0\}/\binom{m}{2}$ as described above in the case of two-sample problem. Results, which are analogous to Theorem 5.1, concerning asymptotic powers of multivariate paired sample tests constructed using sign and signed rank statistics based on data points projected along $\widehat{\boldsymbol{\beta}}$ can be derived under appropriate conditions. We omit the mathematical details as those details are very similar to those in two-sample problems (see our comments after the proof of Lemma 3 in Appendix).

6 Tests based on real valued functions of the data and related issues

Recall now the statement of Theorem 2.2 and discussion preceding the theorem. It is straight forward to verify that the distribution-free property asserted in Theorem 2.2 remain valid if the statistic T is computed based on any real valued function h of the data corresponding to second subsample, where such a h may be chosen based on the first subsample. Linearity of h is not required for Theorem 2.2 to hold. Consequently, if one constructs a nonlinear classifier based on the first subsample and use the corresponding discriminant function to form the test statistic based on the second subsample, one can get a distribution-free test with power properties depending on the choice of the discriminant function. If the distributions of the two multivariate samples are elliptic and unimodal differing only in their location, the optimal Bayes classifier discriminating between the two distributions happens to be linear when the prior probabilities are equal. So, in such cases, it is reasonable to construct tests based on only linear functions of the data. Also, if F and G both belong to the exponential family, the Bayes classifier turn out to be a linear function of the sufficient statistics. So, after finding the sufficient statistics for that family, the same method based on linear projection can be used there as well. However, in more general situations, the Bayes classifier may not be a linear function of the data, and there it will be more appropriate to consider tests based on suitable nonlinear functions of the data.

PROPOSITION 6.1. *Suppose that h is a real valued measurable transformation of d-dimensional observations, and it is chosen from the first subsample. Consider a univariate rank statistic T (e.g., KS or WMW statistic), which is computed on the transformed observations in the second subsample. Then the resulting multivariate two-sample test will also have the distribution-free property. Define $\gamma(h)$ as the power of the univariate test when it is implemented on the observations transformed using the transformation function h. If f and g are density function corresponding to the two distributions F and G, respectively, $\gamma(h)$ is maximized when $h(\cdot) = g(\cdot)/f(\cdot)$, which is the likelihood ratio.*

PROOF. The distribution-free property of the resulting test follows from the arguments used in the proof of Theorem 2.2.

Let $\mathbf{x} \sim F$, $\mathbf{y} \sim G$ and correspondingly $h(\mathbf{x}) \sim F_h$ and $h(\mathbf{y}) \sim G_h$. Since the most powerful tests are unbiased, we have G_h stochastically larger than F_h. Now, consider any other transformation $q(.)$, and let $q(\mathbf{x}) \sim F_q$ and $q(\mathbf{y}) \sim G_q$. Now, if $F_q(t) = F_h(t) \ \ \forall t$, from the properties of the most powerful test, we have $G_q(t) \leq G_h(t) \ \ \forall t$. So, the power of the resulting test gets maximized when the likelihood ratio is used as the transformation function (follows from arguments similar to that used in the proof of Proposition 2.1).

If $F_q(t) \neq F_h(t)$ for at least one t, one can find a monotone transformation $\psi(.)$, such that $\psi \circ q(\mathbf{x}) \sim F_h$. Since the power of the rank test remains invariant under monotone transformation, the transformations $q(\cdot)$ and $\psi \circ q(\cdot)$ lead to the same power. Now, since we have $F_{\psi \circ q}(t) = F_h(t) \ \ \forall t$, using the same argument as above, we can claim that the transformation $h(\cdot)$ leads to more power than the transformation $q(\cdot)$ or $\psi \circ q(\cdot)$.

Therefore, in practice, one can construct consistent estimates $\hat{f}$ and $\hat{g}$ for f and g from the first subsample, and transform the observations in the second subsample using the transformation function $\widehat{\mathcal{T}}(\cdot) = \hat{g}(\cdot)/\hat{f}(\cdot)$. Kernel density estimates or nearest neighbor density estimates can be used for this purpose. Results analogous to Theorem 5.1 can be proved for these transformations as well. However, nonparametric estimation of f and g makes the convergence of the estimates rather slow, especially in high dimension. Another option is to use nonlinear SVM based on radial basis or other suitably chosen basis functions. Note that these nonlinear SVM classifiers can be used even when the dimension of the data is larger than the sample size.

Acknowledgement

We would like to thank Professor Probal Chaudhuri for his active involvement in our academic discussions and providing us with several helpful ideas. His valuable comments and suggestions helped us to improve the manuscript substantially.

Appendix

LEMMA 1. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{m_1}$ and $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{m_2}$ are observations from two distributions F and G, respectively. Define the WMW statistic $U_{m_1,m_2}(\boldsymbol{\beta}) = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} I\{\boldsymbol{\beta}'(\mathbf{x}_i - \mathbf{y}_j) > 0\}$, the KS statistic $K_{m_1,m_2}(\boldsymbol{\beta}) = \sup_{\beta_0} |F_{m_1}^{\boldsymbol{\beta}}(\beta_0) - F_{m_2}^{\boldsymbol{\beta}}(\beta_0)|$, and their population analogs $U(\boldsymbol{\beta}) = P(\boldsymbol{\beta}'(\mathbf{x} - \mathbf{y}) > 0$ and $K(\boldsymbol{\beta}) = \sup_{\beta_0} |F_{\boldsymbol{\beta}}(\beta_0) - F_{\boldsymbol{\beta}}(\beta_0)|$, where $\mathbf{x} \sim F$ and $\mathbf{y} \sim G$. As $\min\{m_1, m_2\}$ tends to infinity, $\sup_{\boldsymbol{\beta}} |U_{m_1,m_2}(\boldsymbol{\beta}) - U(\boldsymbol{\beta})|$ and $\sup_{\boldsymbol{\beta}} |K_{m_1,m_2}(\boldsymbol{\beta}) - K(\boldsymbol{\beta})|$ both converge to 0 almost surely.

PROOF. Using the arguments based on Hoeffding's inequality (for U-statistic) and VC dimension, we can show the almost sure convergence of $\sup_{\boldsymbol{\beta}} |U_{m_1,m_2}(\boldsymbol{\beta}) - U(\boldsymbol{\beta})|$ to 0 (see Theorem 3.1(i) of Ghosh and Chaudhuri (2005) for details) as $\min\{m_1, m_2\} \to \infty$.

Now define $\Delta_{m_1,m_2}(\beta_0, \boldsymbol{\beta}) = \frac{1}{m_1} \sum_{i=1}^{m_1} I\{\boldsymbol{\beta}'\mathbf{x}_i < \beta_0\} + \frac{1}{m_2} \sum_{i=1}^{m_2} I\{\boldsymbol{\beta}'\mathbf{y}_i \geq \beta_0\}$ and $\Delta(\beta_0, \boldsymbol{\beta}) = P(\boldsymbol{\beta}'\mathbf{x} < \beta_0) + P(\boldsymbol{\beta}'\mathbf{y} \geq \beta_0)$. Using the arguments based on Hoeffding's inequality and VC dimension, one can show that $\sup_{\beta_0 \boldsymbol{\beta}} |\Delta_{m_1,m_2}(\beta_0 \boldsymbol{\beta}) - \Delta(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s.} 0$. (see Theorem 3.1 (ii) of Ghosh and Chaudhuri (2005) for details). Also, it can be shown that $\sup_{\beta_0} \Delta(\beta_0, \boldsymbol{\beta}) = 1 + K(\boldsymbol{\beta})$ and $\sup_{\beta_0} \Delta_{m_1,m_2}(\beta_0, \boldsymbol{\beta}) = 1 + K_{m_1,m_2}(\boldsymbol{\beta})$. So, we have $\sup_{\boldsymbol{\beta}} |K_{m_1,m_2}(\boldsymbol{\beta}) - K(\boldsymbol{\beta})| = \sup_{\beta_0 \boldsymbol{\beta}} |\Delta_{m_1,m_2}(\beta_0 \boldsymbol{\beta}) - \Delta(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s.} 0$ as $\min\{m_1, m_2\} \to \infty$.

REMARK. Lemma 1 holds even when d increases with the sample size at the rate of $\min\{m_1, m_2\}^{\rho}$ for some $\rho \in (0, 1)$ (see Section 3 in Ghosh and Chaudhuri, 2005).

LEMMA 2. Consider the objective function $D_m(\beta_0, \boldsymbol{\beta})$ used in DWD classification as discussed in Section 2. Suppose that m_1 (m_2, respectively) out of m observations are from F (G, respectively), and m_1/m tends to $1/2$ as $m \to \infty$. Define $D(\beta_0, \boldsymbol{\beta}) = 0.5\ E[V(\beta_0 + \boldsymbol{\beta}'\mathbf{x})] + E[V(-\beta_0 - \boldsymbol{\beta}'\mathbf{y})]$, where V is defined as in Section 2. Let $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ be a minimizer of $D_m(\beta_0, \boldsymbol{\beta})$, and $(\beta_0^D, \boldsymbol{\beta}^D)$ be the unique minimizer of $D(\beta_0, \boldsymbol{\beta})$. If F and G have finite second moments, $\widehat{\boldsymbol{\beta}}_m^D$ converges to $\boldsymbol{\beta}^D$ almost surely as m tends to infinity.

PROOF. Note that $D_m(\beta_0, \boldsymbol{\beta})$ can also be expressed as $D_m(\beta_0, \boldsymbol{\beta}) = \frac{m_1}{m}\frac{1}{m_1}\sum_{i=1}^{m_1} V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x}_i) + \frac{m_2}{m}\frac{1}{m_2}\sum_{i=1}^{m_2} V(-\beta_0 - \boldsymbol{\beta}^{'}\mathbf{y}_i)$. For any fixed β_0 and $\boldsymbol{\beta}$, $V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x}_i)$ $(i = 1, 2, \ldots, m_1)$ are *i.i.d.* bounded random variables. So, using Hoeffding inequality, we can find a constant A_0 such that for every $\epsilon > 0$, $P\{|\frac{1}{m_1}\sum_{i=1}^{m_1} V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x}_i) - E[V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x})]| > \epsilon\} < 2e^{-A_0 m_1 \epsilon^2}$. Since V is Lipschitz continuous, for any $\boldsymbol{\beta}_{+1} = \begin{pmatrix}\beta_{01}\\ \boldsymbol{\beta}_1\end{pmatrix}$ and $\boldsymbol{\beta}_{+2} = \begin{pmatrix}\beta_{02}\\ \boldsymbol{\beta}_2\end{pmatrix}$, we have $|V(\beta_{01} + \boldsymbol{\beta}_1^{'}\mathbf{x}) - V(\beta_{02} + \boldsymbol{\beta}_2^{'}\mathbf{x})| \leq \|\mathbf{x}\| \|\boldsymbol{\beta}_{+1} - \boldsymbol{\beta}_{+2}\|$. So, under the assumption of the existence of the second moments of F and G, following Theorem 19.4 and Example 19.7 of van der Vaart (2000, pp. 270-71), one can show that the class of functions $\{V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x}) : (\beta_0, \boldsymbol{\beta}) \in R^d\}$ has finite VC dimension ν (say). So, using the results on probability inequalities (see e.g., Devroye, Gyorfi and Lugosi, 1996; van der Vaart, 2000), we get

$$P\{\sup_{\beta_0, \boldsymbol{\beta}} \Big|\frac{1}{m_1}\sum_{i=1}^{m_1} V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x}_i) - E[V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x})]\Big| > \epsilon\} < 2m_1^{\nu} e^{-A_0 m \epsilon^2}.$$

Since $\sum_{m_1} m_1^{\nu} e^{-A_0 m_1 \epsilon^2} < \infty$, using the Borel-Cantelli Lemma, one gets $\sup_{\beta_0, \boldsymbol{\beta}} |\frac{1}{m_1}\sum_{i=1}^{m_1} V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x}_i) - E[V(\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x})]| \xrightarrow{a.s} 0$ as $m_1 \to \infty$. Similarly, we have $\sup_{\beta_0, \boldsymbol{\beta}} |\frac{1}{m_2}\sum_{i=1}^{m_2} V(-\beta_0 - \boldsymbol{\beta}^{'}\mathbf{y}_i) - E[V(-\beta_0 - \boldsymbol{\beta}^{'}\mathbf{y})]| \xrightarrow{a.s} 0$ as $m_2 \to \infty$. Now, m_1/m and m_2/m both converges to $1/2$ as $m \to \infty$. So, we have $\sup_{\beta_0, \boldsymbol{\beta}} |D_m(\beta_0, \boldsymbol{\beta}) - D(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s} 0$ as $m \to \infty$.

Now, from the definition of $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ and $(\beta_0^D, \boldsymbol{\beta}^D)$, it is clear that $|D_m(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) - D((\beta_0^D, \boldsymbol{\beta}^D)| \leq \sup_{\beta_0, \boldsymbol{\beta}} |D_m(\beta_0, \boldsymbol{\beta}) - D(\beta_0, \boldsymbol{\beta})| \xrightarrow{a.s} 0$ as $m \to \infty$. Again, we have $D_m(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) \leq D_m((\beta_0^D, \boldsymbol{\beta}^D)$ and $D(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) \geq D((\beta_0^D, \boldsymbol{\beta}^D)$ for all m. This implies that $|D(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D) - D((\beta_0^D, \boldsymbol{\beta}^D)|$ converges to 0 on a set of probability one. Now, on the same set, if $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ converges, it has to converge to $(\beta_0^D, \boldsymbol{\beta}^D)$ in view of uniqueness of $(\beta_0^D, \boldsymbol{\beta}^D)$ and the continuity of the function $D(\beta_0, \boldsymbol{\beta})$. Here without loss of generality, we can assume that for all m, $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ lies in the compact surface of the unit ball in R^{d+1}. So, any subsequence of the sequence of these estimates has a convergent subsequence converging to $(\beta_0^D, \boldsymbol{\beta}^D)$ on that set of probability one. Hence, $(\widehat{\beta}_{0m}^D, \widehat{\boldsymbol{\beta}}_m^D)$ also converges to $(\beta_0^D, \boldsymbol{\beta}^D)$ almost surely.

LEMMA 3. *If F and G are elliptically symmetric, and they differ only in their location, $\widehat{\boldsymbol{\beta}}_m^M$ converges almost surely to a constant multiple of $\boldsymbol{\beta}_*$ (defined in Section 5) as m tends to infinity. If F and G have finite second moments, we also have this almost sure convergence for $\widehat{\boldsymbol{\beta}}_m^F$ and $\widehat{\boldsymbol{\beta}}_m^D$ and probability convergence for $\widehat{\boldsymbol{\beta}}_m^S$ when λ, the regularization parameter in SVM, is of the order $o(m^{-1/2})$.*

PROOF. If F and G are elliptically symmetric and they differ in their location, the Bayes discriminant function is linear with direction vector proportional to $\boldsymbol{\beta}_*$. Since $\boldsymbol{\beta}_*/\|\boldsymbol{\beta}_*\|$ is unique the maximizer of $U(\boldsymbol{\beta})$ and $KS(\boldsymbol{\beta})$ (see proposition 2.1 and note that we maximize $U(\boldsymbol{\beta})$ and $KS(\boldsymbol{\beta})$

over $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$), from Lemma 1, we have $\widehat{\boldsymbol{\beta}}_m^M \xrightarrow{a.s.} \boldsymbol{\beta}_*/\|\boldsymbol{\beta}_*\|$ both for WMW and KS statistics.

Now, from Fisher consistency (see Qiao et. al., 2010) of the DWD classifier, we have $\boldsymbol{\beta}^D \propto \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^D$ is as defined in Lemma 2. So, when F and G have finite second moments, from Lemma 2, we get the almost sure convergence of $\widehat{\boldsymbol{\beta}}_m^D$ to a constant multiple of $\boldsymbol{\beta}_*$.

The Firsher discriminant function computed from the data is given by $\widehat{\boldsymbol{\beta}}_m^F = \hat{\boldsymbol{\Sigma}}^{-1}(M_1 - M_2)$, where M_1 and M_2 are sample means for $\mathbf{x}$ and $\mathbf{y}$, and $\hat{\boldsymbol{\Sigma}}$ is the moment based estimate of the pooled dispersion matrix. Now, under the assumption of existence of second order moments of F and G, we have $\widehat{\boldsymbol{\beta}}_m^F \xrightarrow{a.s.} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{\beta}_*$.

Now, consider the case of SVM. First note that F and G have disjoint support, there is nothing to prove. So, we assume that they have some overlapping regions. Now, if F and G have finite second moments, they satisfy the assumptions (A1)-(A4) of Koo *et. al.* (2008). So, if λ is of the order $o(m^{-1/2})$, $\widehat{\boldsymbol{\beta}}_m^S$, the minimizer of $S_m(\beta_0, \boldsymbol{\beta})$ converges (in probability) to $\boldsymbol{\beta}^S$, the minimizer of $S(\beta_0, \boldsymbol{\beta}) = 0.5(E[1 - (\beta_0 + \boldsymbol{\beta}^{'}\mathbf{x})_+] + E[1 - (-\beta_0 - \boldsymbol{\beta}^{'}\mathbf{y})_+])$ (follows from Theorem 1 of Koo *et. al.*, 2008). Now, due to Fisher consistency of SVM (see e.g., Lin, 2002), we have $\boldsymbol{\beta}^S$ proportional to $\boldsymbol{\beta}^*$.

COMMENTS. In the case of matched pair data, we have $m_1 = m_2$. Though $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_j$ are independent for $i \neq j$, we have dependency between $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_i$. Note that the convergence of $\widehat{\boldsymbol{\beta}}_m^F$ does not require independence of observations on $\mathbf{x}$ and $\mathbf{y}$. So, similar result holds for the matched pair data. In the proof of Lemma 1 and Lemma 2, we did not use independence between obsevations on $\mathbf{x}$ and $\mathbf{y}$. So, analogous convergence can be proved for matched pair data as well. The convergence result similar to that $|S_m(\beta_0, \boldsymbol{\beta}) - S(\beta_0, \boldsymbol{\beta})|$ for the matched pair data can be proved by writing $S_m(\beta_0, \boldsymbol{\beta})$ as a sum of the functions of the $\mathbf{x}_i$'s and that of $\mathbf{y}_i$'s (like the alterative expression for $D_m(\beta_0, \boldsymbol{\beta})$ used in Lemma 2) and then repeating the arguments used in the proof of Theorem 1 of Koo *et. al.* (2008). So, convergence result analogous to that of $\widehat{\boldsymbol{\beta}}_m^S$ can also be proved.

References

[1] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Mack, D. and Leine, A. J. (1999) Broad pattern of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci., USA*, **96**, 6745-6750.

[2] Andrews, D. W. K. (1988) Laws of large numbers for dependnt non-identially distributted random variables. *Economic Theory*, **4**, 458-467.

[4] Baringhaus, L. and Franz, C. (2004) On a new multivariate two-sample test. *J. Multivariate Anal.*, **88**, 190-206.

[4] Bickel, P. and Levina, E. (2004) Some theory for Fisher's linear discriminant function, naive Bayes and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989-1010.

[5] Billingsley, P. (1995) *Probability and Measure*, Wiley, New York.

[6] Brown, B. M. and Hettmansperger, T. P. (1987) Affine invariant rank methods in the bivariate location model. *J. Royal Statist. Soc., Ser. B*, **49**, 301-310.

[7] Brown, B. M. and Hettmansperger, T. P. (1989) Affine invariant bivariate version of the sign test. *J. Royal Statist. Soc., Ser. B*, **51**, 117-125.

[8] Chakraborty, B. and Chaudhuri, P. (1999) On affine invariant sign and rank tests in one and two-sample multivariate problems. In *Multivariate Analysis, Design of Experiments and Survey Sampling*, (Ed. S. Ghosh), Marcel Dekker, New York, pp. 499-522.

[9] 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.

[10] 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.

[11] Choi, M. and Marden, J. (1997) An approach to multivariate rank tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.*, **92**, 1581-1590.

[12] Cuesta-Albertos, J. A., Fraiman, R. and Ransford, T. (2007) A sharp form of the Cramer-Wold Theorem. *J. Theo. Prob.*, **20**, 201-209.

[13] Devroye, L., Gyorfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*, Springer, New York.

[14] Friedman, J. and Rafsky, C. (1979) Multivariate generalizations of the Wald-Wolfwitz and Smirnov two-sample tests. *Ann. Statist.*, **7**, 697-717.

[15] Ghosh, A. K. and Chaudhuri, P. (2005) On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, **11**, 1-27.

[16] Hajek, J., Sidak, Z. and Sen, P. K. (1999) *Theory of Rank Tests*. Academic Press, New York.

[17] 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.

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

[19] Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004) The entire regularization path for the support vector machine. *J. Mach. Learn. Res.*, **5**, 1391-1415.

[20] Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning : Data Mining, Inference and Prediction*. Springer Verlag, New York.

[21] Henze, N. (1988) A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist*, **16**, 772-783.

[22] Hettmansperger, T. P. and Oja, H. (1994) Affine invariant multivariate multisample sign tests. *J. Royal Statist. Soc., Ser. B*, **56**, 235-249.

[23] Hettmansperger, T. P., Mottonen, J. and Oja, H. (1998) On affine invariant multivariate rank tests for several samples. *Statistica Sinica*, **8**, 785-800.

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

[25] Koo, J.-Y., Lee, Y., Kim, Y. and Park, C. (2008) A Bahadur representation of the linear support vector machine. *J. Mach. Learn. Res.*, **9**, 1343-1368.

[26] Lin, Y. (2002) A note on margin-based loss function in classification. *Statist. Probab. Lett.*, **68**, 73-82.

[27] Liu, R. and Singh, K. (1993) A quality index based on data depth and multivariate rank tests. *J. Amer. Statist. Assoc.*, **88**, 252-260.

[28] Marden, J. (1999) Multivariate rank tests. In *Multivariate Analysis, Design of Experiments and Survey Sampling*, (Ed. S. Ghosh), Marcel Dekker, New York, pp. 401-432.

[29] Marron, J. S., Todd, M. J. and Ahn, J. (2007) Distance weighted discrimination. *J. Amer. Statist. Assoc.*, **102**, 1267-1271.

[30] Mottonen, J. and Oja, H. (1995) Multivariate spatial sign and rank methods. *J. Nonparametric Statist.*, **5**, 201-213.

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

[32] Oja, H. and Randles, R. (2004) Multivariate nonparametric tests. *Statistical Science*, **19**, 598-605.

[33] Oja, H. (2010) *Multivariate Nonparametric Methods with R*, Springer, New York.

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

[35] Qiao, X., Zhang, H. H., Liu, Y., Todd, M. J. and Marron, J. S. (2010) Weighted distance weighted discrimination and its asymptotic properties. *J. Amer. Statist. Assoc.*, **105**, 401-414.

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

[37] Randles, R. and Peters, D. (1990) Multivariate rank tests for the two-sample location problem. *Comm. Statist., Theory and Methods*, **19**, 4225-4238.

[38] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. Royal Statist. Soc. Ser. B*, **67**, 515-530.

[39] Schilling, M. F. (1986) Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.*,**81**, 799-806.

[40] van der Vaart, A. W. (2000) *Asymptotic Statistics*, Cambridge University Press, Cambridge.

[41] Vapnik. V. N. (1998) *Statistical Learning Theory.* Wiley, New York.

[42] Yata, K. and Aoshima, M. (2010) Effective PCA for high dimension, low sample size data with singular value decomposition of cross data matrix. *J. Multivariate Anal.*, **101**, 2060-2077.