

The linear stochastic order and directed inference for multivariate ordered distributions

Ori Davidov

Department of Statistics, University of Haifa, Mount Carmel, Haifa 31905 Israel, E-mail: davidov@stat.haifa.ac.il

and

Shyamal Peddada

Biostatistics Branch, National Institute of Environmental Health Sciences, Alexander Drive, RTP, NC 27709,
E-mail: peddada@niehs.nih.gov

ABSTRACT

Researchers are often interested in drawing inferences regarding the order between two experimental groups on the basis of multivariate response data. Since standard multivariate methods are designed for two sided alternatives they may not be ideal for testing for order between two groups. In this article we introduce the notion of the linear stochastic order and investigate its properties. Statistical theory and methodology are developed to both estimate the direction which best separates two arbitrary ordered distributions and to test for order between the two groups. The new methodology generalizes Roy's classical largest root test to the nonparametric setting and is applicable to random vectors with discrete and/or continuous components. The proposed methodology is illustrated using data obtained from a 90-day pre-chronic rodent cancer bioassay study conducted by the National Toxicology Program (NTP).

Key Words: nonparametric tests, order restricted statistical inference, stochastic order relations.

1. Introduction

In a variety of applications researchers are interested in comparing two treatment groups on the basis of several, potentially dependent, outcomes. For example, to evaluate if a chemical is a neuro-toxicant, toxicologists compare a treated group of animals with an untreated control group in terms of various correlated outcomes such as: Tail-pinch response, Click response, and Gait score, among many others (cf. Moser, 2000). The statistical problem of interest is to compare the multivariate distributions of the outcomes in the control and treatment groups. Moreover, the outcome distributions are expected to be ordered in some sense. The theory of stochastic order relations (Shaked and Shanthikumar, 2007) provides the theoretical foundation for such comparisons.

To fix ideas let \mathbf{X} and \mathbf{Y} be p dimensional random variables (RVs); \mathbf{X} is said to be smaller than \mathbf{Y} in the multivariate stochastic order, denoted $\mathbf{X} \preceq_{st} \mathbf{Y}$, provided $\mathbb{P}(\mathbf{X} \in U) \leq \mathbb{P}(\mathbf{Y} \in U)$ for all upper sets $U \in \mathbb{R}^p$ (Shaked and Shanthikumar, 2007). If for some upper set the above inequality is sharp we say that \mathbf{X} is strictly smaller than \mathbf{Y} (in the multivariate stochastic order) which we denote by $\mathbf{X} \prec_{st} \mathbf{Y}$. Recall that a set $U \in \mathbb{R}^p$ is called an upper set if $\mathbf{u} \in U$ implies that $\mathbf{v} \in U$ whenever $\mathbf{u} \leq \mathbf{v}$, i.e., if $u_i \leq v_i$, $i = 1, \dots, p$. Note that comparing \mathbf{X} and \mathbf{Y} with respect to the multivariate stochastic order requires comparing

their distributions over all upper sets in \mathbb{R}^p . This turns out to be a very high dimensional problem. For example if \mathbf{X} and \mathbf{Y} are multivariate binary RVs then $\mathbf{X} \preceq_{st} \mathbf{Y}$ provided $\sum_{\mathbf{t} \in U} p_{\mathbf{X}}(\mathbf{t}) \leq \sum_{\mathbf{t} \in U} p_{\mathbf{Y}}(\mathbf{t})$ where $p_{\mathbf{X}}(\mathbf{t})$ and $p_{\mathbf{Y}}(\mathbf{t})$ are the corresponding probability mass functions. Here $U \in \mathcal{U}_p$ where \mathcal{U}_p is the family of upper sets defined on the support of a p dimensional multivariate binary RV. It turns out that the cardinality of \mathcal{U}_p , denoted by N_p , grows super-exponentially with p . In fact $N_1 = 1$, $N_2 = 4$, $N_3 = 18$, $N_4 = 166$, $N_5 = 7579$ and $N_6 = 7828352$. The values of N_7 and N_8 are also known but N_9 is not. However good approximations for N_p are available for all p (cf. Davidov and Peddada 2011). Obviously the number of upper sets for general multivariate RVs is much larger. Since in many applications p is large it would seem that the analysis of high dimensional stochastically ordered data is practically hopeless. As a consequence, the methodology for analyzing multivariate ordered data is underdeveloped. It is worth mentioning that Sampson and Whitaker (1989) as well as Lucas and Wright (1991) studied stochastically ordered bivariate multinomial distributions. They noted the difficulty of extending their methodology to high dimensional data due to the large number of constraints that need to be imposed. Recently Davidov and Peddada (2011) proposed a framework for testing for order among K , p -dimensional, ordered multivariate binary distributions.

In this paper we address the dimensionality problem by considering an easy to understand stochastic order which we refer to as the linear stochastic order.

Definition 1.1. *The RV \mathbf{X} is said to be smaller than the RV \mathbf{Y} in the (multivariate) linear stochastic order, denoted $\mathbf{X} \preceq_{l-st} \mathbf{Y}$, if for all $\mathbf{s} \in \mathbb{R}_+^p = \{\mathbf{s} : \mathbf{s} \geq 0\}$*

$$\mathbf{s}^T \mathbf{X} \preceq_{st} \mathbf{s}^T \mathbf{Y} \tag{1.1}$$

where \preceq_{st} in (1.1) denotes the usual (univariate) stochastic order.

Note that it is enough to limit (1.1) to all non-negative real vectors satisfying $\|\mathbf{s}\| = 1$ and accordingly we denote by \mathcal{S}_+^{p-1} the positive part of the unit sphere in \mathbb{R}^p . We call each $\mathbf{s} \in \mathcal{S}_+^{p-1}$ a "direction". In other words the RVs \mathbf{X} and \mathbf{Y} are ordered by the linear stochastic order if every non-negative linear combination of their components is ordered by the usual (univariate) stochastic order. Thus instead of considering all upper sets in \mathbb{R}^p we need for each $\mathbf{s} \in \mathcal{S}_+^{p-1}$ to consider only upper sets in \mathbb{R} . This is a substantial reduction in dimensionality. In fact we will show that only one value of \mathbf{s} , determined by the data, need be considered. Note that the linear stochastic order, like the multivariate stochastic order, is a generalization of the usual univariate stochastic order to multivariate data. Both of these orders indicate, in different ways, that one random vector is more likely than another to take on large values. In this paper we develop the statistical theory and methodology for estimation and testing for linearly ordered multivariate distributions. For completeness we note that weaker notions of the linear stochastic order are discussed by Hu et al. (2011) and applied to various optimization problems in queuing and finance.

Comparing linear combinations has a long history in statistics. For example, in Phase I clinical trials it is common to compare dose groups using an overall measures of toxicity. Typically, this quantity is an ad hoc weighted average of individual toxicities where the weights are often known as "severity weights" (cf. Bekele and Thall, 2004, Ivanova and Murphy, 2009). This strategy of dimension reduction is not new in the statistical literature and has been used in classical multivariate analysis when comparing two or more multivariate normal populations. For example, using the union-intersection principle, the comparison of

multivariate normal populations can be reduced to the comparison of all possible linear combinations of their mean vectors. This approach is the basis of Roy’s classical largest root test (Roy 1953, Johnson and Wichern, 1998). Our proposed test may be viewed as nonparametric generalization of the classical normal theory method described above with the exception that we limit consideration only to non–negative linear combinations (rather than all possible linear combinations) since our main focus is to make comparisons in terms of stochastic order. We emphasize that the linear stochastic order will allow us to address the much broader problem of directional ordering for multivariate ordered data, i.e., to find the direction which best separates two ordered distributions. Based on our survey of the literature, we are not aware of any methodology that addresses the problems investigated here.

This paper is organized in the following way. In Section 2 some probabilistic properties of the linear stochastic order are explored and its relationships with other multivariate stochastic orders are clarified. In Section 3 we provide the background and motivation for directional inference under the linear stochastic order. Estimation and testing procedures are developed. In particular the estimator of the best separating direction is presented and its large sampling properties derived. We note that the problem of estimating the best separating direction is a non–smooth optimization problem. The limiting distribution of the best separating direction is found to be non–normal and the limit involves cube–root asymptotics. Tests for the linear stochastic order based on the best separating direction are also developed. One advantage of our approach is that it avoids the estimation of multivariate distributions subject to order restrictions. Simulation results, presented in Section 4, reveal that for large sample sizes the proposed estimator has negligible bias and mean squared error. The bias and MSE seem to depend on the true value of the best separating direction, the dependence structure and the dimension of the problem. Furthermore, the proposed test honors the nominal type I error rate and has sufficient power. In Section 5 we illustrate the methodology using data obtained from the National Toxicology Program (NTP). Concluding remarks and some open research problems are provided in Section 6. For convenience all proofs are provided in an Appendix where additional concepts are defined when needed.

2. Some properties of the linear stochastic order

We start by clarifying the relationship between the linear stochastic order and the multivariate stochastic order. First note that $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ if and only if $\mathbb{P}(\mathbf{s}^T \mathbf{X} \geq t) \leq \mathbb{P}(\mathbf{s}^T \mathbf{Y} \geq t)$ for all $(t, \mathbf{s}) \in \mathbb{R} \times \mathbb{R}_+^p$ which is equivalent to $\mathbb{P}(\mathbf{X} \in H) \leq \mathbb{P}(\mathbf{Y} \in H)$ for all $H \in \mathcal{H}$ where \mathcal{H} is the collection of all upper–half–planes, i.e., sets which are both half planes and upper sets. Thus $\mathbf{X} \preceq_{st} \mathbf{Y} \Rightarrow \mathbf{X} \preceq_{l-st} \mathbf{Y}$. The converse does not hold in general.

Example 2.1. *Let \mathbf{X} and \mathbf{Y} be bivariate RVs such that $\mathbb{P}(\mathbf{X} = (1, 1)) = \mathbb{P}(\mathbf{X} = (0, 1)) = \mathbb{P}(\mathbf{X} = (1, 0)) = 1/3$ and $\mathbb{P}(\mathbf{Y} = (3/4, 3/4)) = \mathbb{P}(\mathbf{Y} = (1, 2)) = \mathbb{P}(\mathbf{Y} = (2, 1)) = 1/3$. It is easy to show that \mathbf{X} is smaller than \mathbf{Y} in the linear stochastic order but not in the multivariate stochastic order.*

The following theorem provides some closure results for the linear stochastic order.

Theorem 2.1. *(i) If $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ then $g(\mathbf{X}) \preceq_{l-st} g(\mathbf{Y})$ for any linear increasing function; (ii) If $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ then $\mathbf{X}_I \preceq_{l-st} \mathbf{Y}_I$ for each subset $I \in \{1, \dots, p\}$; (iii) If $\mathbf{X} | \mathbf{Z} = \mathbf{z} \preceq_{l-st}$*

$\mathbf{Y}|\mathbf{Z} = \mathbf{z}$ for all \mathbf{z} in the support of \mathbf{Z} then $\mathbf{X} \preceq_{l-st} \mathbf{Y}$; (iii) If $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent RVs with dimensions p_i and similarly for $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ and if in addition $\mathbf{X}_i \preceq_{l-st} \mathbf{Y}_i$ then $(\mathbf{X}_1, \dots, \mathbf{X}_n) \preceq_{l-st} (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$; (iv) Finally, if $\mathbf{X}_n \rightarrow \mathbf{X}$ and $\mathbf{Y}_n \rightarrow \mathbf{Y}$ where convergence can be in distribution, in probability or almost surely and if $\mathbf{X}_n \preceq_{l-st} \mathbf{Y}_n$ for all n then $\mathbf{X} \preceq_{l-st} \mathbf{Y}$.

Theorem 2.1 shows that the linear stochastic order is closed under increasing linear transformations, marginalization, mixtures, conjugations and convergence. In particular parts (ii) and (iii) of Theorem 2.1 imply that if $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ then $X_i \preceq_{st} Y_i$ and $X_i + X_j \preceq_{st} Y_i + Y_j$ for all i and j , i.e., all marginals are ordered as are all convolutions. Although the multivariate stochastic order is in general stronger than the linear stochastic order there are situation in which both orders coincide.

Theorem 2.2. *Let \mathbf{X} and \mathbf{Y} be continuous elliptically distributed RVs supported on \mathbb{R}^p with the same generator. Then $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ if and only if $\mathbf{X} \preceq_{st} \mathbf{Y}$.*

Note that the elliptical family of distributions is large and includes the multivariate normal, multivariate t and the exponential power family, see Fang et al. (1987). Thus Theorem 2.2 shows that the multivariate stochastic order coincides with the linear stochastic order in the normal family. Incidentally, in the proof of Theorem 2.2 we generalize the results of Ding and Zhang (2004) on multivariate stochastic ordering of elliptical RVs. Another interesting example is:

Theorem 2.3. *Let \mathbf{X} and \mathbf{Y} be multivariate binary RVs. Then $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ is equivalent to $\mathbf{X} \preceq_{st} \mathbf{Y}$ if and only if $p \leq 3$.*

Remark 2.1. *In the proof of Theorem 2.2 distributional properties of the elliptical family play a major role. In contrast, Theorem 2.3 is a consequence of the geometry of the upper sets of multivariate binary RVs which turn out to be upper half planes if and only if $p \leq 3$.*

We now explore the role of the dependence structure.

Theorem 2.4. *Let \mathbf{X} and \mathbf{Y} have the same copula. Then $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ if and only if $\mathbf{X} \preceq_{st} \mathbf{Y}$.*

Theorem 2.4 establishes that if two RVs have the same dependence structure as quantified by their copula function (cf. Joe 1997) then the linear and multivariate stochastic orders coincide. Such situations arise when the correlation structure among outcomes is not expected to vary with dose.

The orthant orders are also of interest in statistical applications. We say that \mathbf{X} is smaller than \mathbf{Y} in the upper orthant order, denoted $\mathbf{X} \preceq_{uo} \mathbf{Y}$, if $\mathbb{P}(\mathbf{X} \in O) \leq \mathbb{P}(\mathbf{Y} \in O)$ for all $O \in \mathcal{O}$ where \mathcal{O} is the collection of upper orthants, i.e., sets of the form $\{\mathbf{z} : \mathbf{z} \geq \mathbf{x}\}$ for some fixed $\mathbf{x} \in \mathbb{R}^p$. The lower orthant order is similarly defined (cf. Shaked and Shanthikumar, 2007 or Davidov and Herman 2011). It is obvious that the orthant orders are weaker than the usual multivariate stochastic order, i.e., $\mathbf{X} \preceq_{st} \mathbf{Y} \Rightarrow \mathbf{X} \preceq_{uo} \mathbf{Y}$ and $\mathbf{X} \preceq_{lo} \mathbf{Y}$. In general the linear stochastic order does not imply the upper (or lower) orthant order, nor is the converse true. However, as stated below, under some conditions on the copula functions the linear stochastic order implies the upper (or lower) orthant order.

Theorem 2.5. *If $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ and $C_{\mathbf{X}}(\mathbf{u}) \leq C_{\mathbf{Y}}(\mathbf{u})$ for all $\mathbf{u} \in [0, 1]^p$ then $\mathbf{X} \preceq_{lo} \mathbf{Y}$. Similarly if $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ and $\overline{C}_{\mathbf{X}}(\mathbf{u}) \leq \overline{C}_{\mathbf{Y}}(\mathbf{u})$ for all $\mathbf{u} \in [0, 1]^p$ then $\mathbf{X} \preceq_{uo} \mathbf{Y}$.*

Note that $C_{\mathbf{X}}(\mathbf{u})$ and $\overline{C}_{\mathbf{X}}(\mathbf{u})$ above are the copula and tail-copula functions for the RV \mathbf{X} (cf. Joe 1994) and are defined in the Appendix. Similarly for $C_{\mathbf{Y}}(\mathbf{u})$ and $\overline{C}_{\mathbf{Y}}(\mathbf{u})$. Further note that the relations $C_{\mathbf{X}}(\mathbf{u}) \leq C_{\mathbf{Y}}(\mathbf{u})$ and/or $\overline{C}_{\mathbf{X}}(\mathbf{u}) \leq \overline{C}_{\mathbf{Y}}(\mathbf{u})$ indicate that the components of \mathbf{Y} are more strongly dependent than the components of \mathbf{X} . This particular dependence ordering is known as positive quadrant dependence. It can be further shown that strong dependence and the linear stochastic order do not in general imply stochastic ordering.

Additional properties of the linear stochastic order as they relate to estimation and testing problems are given in Section 3.

3. Directional inference

3.1. Background and motivation

There exists a long history of well developed theory for comparing two or more multivariate normal (MVN) populations. Methods for assessing whether there are any differences between the populations, which differ? in which component(s)? and by how much? have been addressed in the literature using a variety of simultaneous confidence intervals and multiple comparison methods (cf. Johnson and Wichern, 1998). Of particular interest to us is Roy's largest root test. To fix ideas consider two multivariate normal random vectors \mathbf{X} and \mathbf{Y} with means $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, respectively, and a common variance matrix $\boldsymbol{\Sigma}$. Using the Union-Intersection principle Roy (1953) expressed the problem of testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\nu}$ versus $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\nu}$ as a collection of univariate testing problems, by showing that H_0 and H_1 are equivalent to $\bigcap_{\mathbf{s} \in \mathbb{R}^p} H_{0,\mathbf{s}}$ and $\bigcup_{\mathbf{s} \in \mathbb{R}^p} H_{1,\mathbf{s}}$ where $H_{0,\mathbf{s}} : \mathbf{s}^T \boldsymbol{\mu} = \mathbf{s}^T \boldsymbol{\nu}$ and $H_{1,\mathbf{s}} : \mathbf{s}^T \boldsymbol{\mu} \neq \mathbf{s}^T \boldsymbol{\nu}$. Implicitly Roy's test identifies the linear combination $\mathbf{s}_{\max}^T (\boldsymbol{\nu} - \boldsymbol{\mu})$ that corresponds to the largest "distance" between the mean vectors, i.e., the direction which best separates their distributions. The resulting test, known as Roy's largest root test, is given by the largest eigenvalue of $\mathbf{B}\mathbf{S}^{-1}$ where \mathbf{B} is the matrix of between groups (or populations) sums of squares and cross products and \mathbf{S} is the usual unbiased estimator of $\boldsymbol{\Sigma}$. In the special case when there are only two populations this test statistic is identical to Hotelling's T^2 statistic. From the simultaneous confidence intervals point of view, the critical values derived from the null distribution of this statistic can be used for constructing Scheffe's simultaneous confidence intervals for all possible linear combinations of the difference $(\boldsymbol{\mu} - \boldsymbol{\nu})$. Further note that the estimated direction corresponding to Roy's largest root test is $\mathbf{S}^{-1}(\overline{\mathbf{Y}} - \overline{\mathbf{X}})$ where $\overline{\mathbf{X}}$ and $\overline{\mathbf{Y}}$ are the respective sample means based on n and m observations, respectively.

Our objective is to extend and generalize the classical multivariate method, described above, to non-normal multivariate ordered data. Our approach will be nonparametric. Recall that comparing MVNs is done by considering the family of statistics $T_{n,m}(\mathbf{s}) = \mathbf{s}^T (\overline{\mathbf{Y}} - \overline{\mathbf{X}})$ for all $\mathbf{s} \in \mathbb{R}^p$. In the case of non-normal populations, the population mean alone may not be enough to characterize the distribution. In such cases, it may not be sufficient to compare the means of the populations but one may have to compare entire distributions.

One possible way of doing so is by considering rank statistics. Define

$$R_k(\mathbf{s}) = \sum_{i=1}^n \mathbb{I}_{(\mathbf{s}^T \mathbf{X}_i \leq \mathbf{s}^T \mathbf{X}_k)} + \sum_{j=1}^m \mathbb{I}_{(\mathbf{s}^T \mathbf{Y}_j \leq \mathbf{s}^T \mathbf{X}_k)}$$

to be the rank of $\mathbf{s}^T \mathbf{X}_k$ in the combined sample $\mathbf{s}^T \mathbf{X}_1, \dots, \mathbf{s}^T \mathbf{X}_n, \mathbf{s}^T \mathbf{Y}_1, \dots, \mathbf{s}^T \mathbf{Y}_m$. For fixed $\mathbf{s} \in \mathbb{R}^p$ the distributions of $\mathbf{s}^T \mathbf{X}$ and $\mathbf{s}^T \mathbf{Y}$ can be compared using a rank test. For example if we use $W_{n,m}(\mathbf{s}) = \sum_{i=1}^n R_i(\mathbf{s})$ our comparison is done in terms of Wilcoxon's rank sum statistics. It is well known that rank tests are well suited for testing for univariate stochastic order (cf. Hajek et al. 1999, Davidov 2012) where the restrictions that $\mathbf{s} \in \mathcal{S}_+^{p-1}$ must be made. Although any rank test can be used, the Mann–Whitney form of Wilcoxon's (WMW) statistic is particularly attractive in this application. Therefore in the rest of this paper we develop estimation and testing procedures for the linear stochastic order based on the family of statistics

$$\Psi_{n,m}(\mathbf{s}) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{(\mathbf{s}^T \mathbf{X}_i \leq \mathbf{s}^T \mathbf{Y}_j)} \quad (3.1)$$

where \mathbf{s} varies over \mathcal{S}_+^{p-1} . Note that (3.1) unbiasedly estimates

$$\Psi(\mathbf{s}) = \mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}). \quad (3.2)$$

The following result is somewhat surprising.

Proposition 3.1. *Let \mathbf{X} and \mathbf{Y} be independent MVNs with means $\boldsymbol{\mu} \leq \boldsymbol{\nu}$ and common variance matrix $\boldsymbol{\Sigma}$. Then Roy's maximal separating direction $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu} - \boldsymbol{\mu})$ also maximizes $\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y})$.*

Proposition 3.1 shows that the direction which separates the means, in the sense of Roy, also maximizes (3.2). Thus it provides further support for choosing (3.1) as our test statistic. Note that in general $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu} - \boldsymbol{\mu})$ may not belong to \mathcal{S}_+^{p-1} . Since we focus on the linear statistical order, we restrict ourselves to $\mathbf{s} \in \mathcal{S}_+^{p-1}$. Consequently we define $\mathbf{s}_{\max} := \arg \max_{\mathbf{s} \in \mathcal{S}_+^{p-1}} \Psi(\mathbf{s})$ and refer to \mathbf{s}_{\max} as the best separating direction. Further note that if \mathbf{X} and \mathbf{Y} are independent and continuous and if $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ then $\Psi(\mathbf{s}) \geq 1/2$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$. This simply means that $\mathbf{s}^T \mathbf{X}$ tends to be smaller than $\mathbf{s}^T \mathbf{Y}$ more than 50% of the time. Note that probabilities of the type (3.2) were introduced by Pitman (1937) and further studied by Peddada (1985) for comparing estimators. Random variables satisfying such a condition are said to be ordered by the precedence order (Arcones et al. 2002).

Once \mathbf{s}_{\max} is estimated we can plug it in (3.1) to get a test statistic. Hence our test may be viewed as a natural generalization of Roy's largest root test from MVNs to arbitrary ordered distributions. However unlike Roy's method, which does not explicitly estimate \mathbf{s}_{\max} we do. On the other hand the proposed test does not require the computation of the inverse of the sample covariance matrix whereas Roy's test and Hotteling's T^2 test require such computations. Consequently, such tests cannot be used when $n < p$ whereas our test can be used in all such instances.

3.2. Estimating the best separating direction

Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_m$ be random samples from the two populations. Rewrite (3.1) as

$$\Psi_{n,m}(\mathbf{s}) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{(\mathbf{s}^T \mathbf{Z}_{ij} \geq 0)} \quad (3.3)$$

where $\mathbf{Z}_{ij} = \mathbf{Y}_j - \mathbf{X}_i$. The maximizer of (3.3) is denoted by $\widehat{\mathbf{s}}_{\max}$, i.e.,

$$\widehat{\mathbf{s}}_{\max} = \arg \max_{\mathbf{s} \in \mathcal{S}_+^{p-1}} \Psi_{n,m}(\mathbf{s}). \quad (3.4)$$

Finding (3.4) with $\mathbf{s} \in \mathcal{S}_+^{p-1}$ is a non-smooth optimization problem. Consider first the situation where $p = 2$. In this case we maximize (3.3) subject to $\mathbf{s} \in \mathcal{S}_+^1 = \{(s_1, s_2) : s_1^2 + s_2^2 = 1, (s_1, s_2) \geq 0\}$. Geometrically \mathcal{S}_+^1 is a quarter circle spanning the first quadrant. Now let $\mathbf{Z} = (Z_1, Z_2)$ and without any loss of generality assume that $\|\mathbf{Z}\| = 1$. We examine the behavior of the function $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)}$ as a function of \mathbf{s} . Clearly if $\mathbf{Z} \geq 0$, i.e., if $Z_1 \geq 0, Z_2 \geq 0$ then for all $\mathbf{s} \in \mathcal{S}_+^1$ we have $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)} = 1$. In other words any value of \mathbf{s} on the arc \mathcal{S}_+^1 maximizes $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)}$. Similarly if $\mathbf{Z} < 0$ then for all $\mathbf{s} \in \mathcal{S}_+^1$ we have $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)} = 0$ and again the entire arc \mathcal{S}_+^1 maximizes $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)}$. Now let $Z_1 \geq 0$ and $Z_2 < 0$. It follows that $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)} = 1$ provided $\cos(\mathbf{s}^T \mathbf{Z}) \geq 0$. Thus $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)} = 1$ for all \mathbf{s} on the arc $[0, \theta]$ for some θ . If $Z_1 < 0$ and let $Z_2 \geq 0$ the situation is reversed and $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)} = 1$ for all angles \mathbf{s} on the arc $[\theta, \pi/2]$. The value of θ is given by (3.5). In other words each \mathbf{Z} is mapped to an arc on \mathcal{S}_+^1 as described above. Now, the function (3.3) simply counts the number of arcs covering each $\mathbf{s} \in \mathcal{S}_+^1$. The maximizer of (3.3) lies in the region where the maximum number of arcs overlap. Clearly this implies that the maximizer of (3.3) is not unique. A quick way to find the maximizer is:

Algorithm 3.1. Let M denote the number of \mathbf{Z}_{ij} 's which belong to the second or fourth quadrant. Map

$$\mathbf{Z}_{ij} \mapsto \theta_{ij} = \begin{cases} \pi/2 - \cos^{-1}(Z_{ij,1}) & \text{if } Z_{ij,1} \geq 0, Z_{ij,2} < 0 \\ \cos^{-1}(Z_{ij,1}) - \pi/2 & \text{if } Z_{ij,1} < 0, Z_{ij,2} \geq 0 \end{cases}. \quad (3.5)$$

Relabel and order the resulting angles as $\theta_{[1]} < \dots < \theta_{[M]}$. Also define $\theta_{[0]} = 0$ and $\theta_{[M+1]} = \pi/2$. Evaluate $\Psi_{n,m}(\mathbf{s}_{[i]})$ $i = 1, \dots, M$ where $\mathbf{s}_{[i],1} = \cos(\theta_{[i]})$ and $\mathbf{s}_{[i],2} = \sin(\theta_{[i]})$. If a maximum is attained at $\theta_{[j]}$ then any value in $[\theta_{[j-1]}, \theta_{[j]}]$ or $[\theta_{[j]}, \theta_{[j+1]}]$ maximizes (3.1).

In light of the above discussion it can be easily proved that:

Proposition 3.2. For $p = 2$ Algorithm 3.1 maximizes (3.1).

In the general case, i.e., for $p \geq 3$, each observation \mathbf{Z}_{ij} is associated with a "slice" of \mathcal{S}_+^{p-1} . The boundaries of the slice are the intersection of \mathcal{S}_+^{p-1} and some half-plane. Note that when $p = 2$ the slices are arcs. The shape of the slice depends on the quadrant to which \mathbf{Z}_{ij} belongs. The maximizer of (3.1) is again the value of \mathbf{s} which belongs to the largest number of slices. Although the geometry of the resulting optimization problem is easy to understand, we have not been able to devise a simple algorithm, which scales with p , based on the ideas above. However, we have found that (3.4) can be obtained by converting the data into polar coordinates and then using the Nelder-Mead algorithm which does not

require the objective function to be differentiable. We emphasize that this maximization process results in a single maximizer of (3.1) and we do not attempt to find the entire set of maximizers. For completeness we note that there are methods for optimizing (3.1) specifically designed for non-smooth problems. For more details see Price et al. (2008) and Audet et al. (2008) and the references therein for both algorithms and convergence results.

3.2.1. Large sample behavior

We will need the following notation. Let $\mathbf{s}_{\max}^\perp = \{\mathbf{t} \in \mathbb{R}^p : \mathbf{s}_{\max}^T \mathbf{t} = 0\}$ be the subspace orthogonal to \mathbf{s}_{\max} . For $\mathbf{t} \in \mathbf{s}_{\max}^\perp$ write

$$\mathbf{s}(\mathbf{t}) = \sqrt{1 - \|\mathbf{t}\|^2} \mathbf{s}_{\max} + \mathbf{t}. \quad (3.6)$$

Clearly $\mathbf{s}(\mathbf{t}) \in \mathcal{S}^{p-1}$ for all $\|\mathbf{t}\| \leq 1$, and \mathbf{t} are referred to as the local coordinates for \mathcal{S}^{p-1} (Kim and Pollard, 1990). It is convenient to think of \mathbf{t} as the discrepancy between the true and estimated value since $\mathbf{s}(\mathbf{t}) \rightarrow \mathbf{s}_{\max}$ if and only if $\mathbf{t} \rightarrow \mathbf{0}$. In fact, if we equate $\mathbf{s}(\mathbf{t})$ with $\widehat{\mathbf{s}}_{\max}$ and \mathbf{t} with $\widehat{\mathbf{t}}_N$, where $N = n + m$ is the total sample size, we find that $\|\widehat{\mathbf{t}}_N\| = \sqrt{1 - (\mathbf{s}_{\max}^T \widehat{\mathbf{s}}_{\max})^2}$ and it will be convenient to write the standardized limit law of $\widehat{\mathbf{s}}_{\max}$ in terms of $\widehat{\mathbf{t}}_N$.

Under suitable regularity conditions the large sampling behavior of $\widehat{\mathbf{s}}_{\max}$ is given below.

Theorem 3.1. *Let \mathbf{X} and \mathbf{Y} have continuously differentiable and bounded densities. If $\Psi(\mathbf{s})$ is uniquely maximized by $\mathbf{s}_{\max} \in \text{interior}(\mathcal{S}_+^{p-1})$. Then $\widehat{\mathbf{s}}_{\max}$ is strongly consistent, i.e.,*

$$\widehat{\mathbf{s}}_{\max} \xrightarrow{a.s.} \mathbf{s}_{\max},$$

converges at a cube root rate, that is $\widehat{\mathbf{s}}_{\max} = \mathbf{s}_{\max} + O_p(N^{-1/3})$, where $N = n + m$, and

$$N^{1/3} \widehat{\mathbf{t}}_N \Rightarrow \mathbf{W}$$

where $\widehat{\mathbf{t}}_N$ are the local coordinates for $\widehat{\mathbf{s}}_{\max}$, the symbol \Rightarrow denotes weak convergence and \mathbf{W} has the distribution of the almost surely unique maximizer of the process $\mathbf{t} \mapsto -[Q(\mathbf{t}) + \mathbb{W}(\mathbf{t})]$ where $Q(\mathbf{t})$ is a quadratic function in \mathbf{t} and $\mathbb{W}(\mathbf{t})$ is a zero mean Gaussian process described in the body of the proof.

Theorem 3.1 shows that under regularity conditions $\widehat{\mathbf{s}}_{\max}$ is strongly consistent and converges at a cube-root rate to a non-normal limit. The cube root rate is due to the discontinuous nature of the objective function (3.1). General results dealing with this kind of asymptotics for independent observations are given by Kim and Pollard (1990).

Remark 3.1. *Note that if either \mathbf{X} or \mathbf{Y} are continuous RVs then $\Psi(\mathbf{s})$ is continuous.*

We have not been able to find general necessary condition(s) for a unique maximizer of $\Psi(\mathbf{s})$ although important sufficient conditions can be found. For example,

Proposition 3.3. *If $\mathbf{Z} = \mathbf{Y} - \mathbf{X}$ and there exist $\boldsymbol{\delta} = \boldsymbol{\nu} - \boldsymbol{\mu} \geq \mathbf{0}$ and $\boldsymbol{\Sigma}$ so the distribution of*

$$\frac{\mathbf{s}^T \mathbf{Z} - \mathbf{s}^T \boldsymbol{\delta}}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}}$$

is independent of \mathbf{s} then the maximizer of $\Psi(\mathbf{s})$ is unique.

The condition above is satisfied by location scale families and it may be convenient to think of $\boldsymbol{\delta}$ and $\boldsymbol{\Sigma}$ as the location and scale parameters for \mathbf{Z} .

Remark 3.2. In general, however, $\Psi(\mathbf{s})$ may not have a unique maximum nor be continuous. For example, if both \mathbf{X} and \mathbf{Y} are discrete RVs then $\Psi(\mathbf{s})$ is a step function. In such situations \mathbf{s}_{\max} is set valued. As we have seen earlier $\widehat{\mathbf{s}}_{\max}$ is always set valued. Consider for example the case where $\mathbb{P}(\mathbf{X} = (-1, -1)) = \mathbb{P}(\mathbf{X} = (1, 1)) = 1/2$ and let $\mathbb{P}(\mathbf{Y} = (-1, -1)) = 1/2 - \varepsilon$ and $\mathbb{P}(\mathbf{Y} = (1, 1)) = 1/2 + \varepsilon$ for some small $\varepsilon > 0$. It is clear that $\mathbf{X} \prec_{st} \mathbf{Y}$. Further note that $\mathbf{Z} \in \{(2, 2), (0, 0), (-2, -2)\}$ and it follows that $\Psi_{n,m}(\mathbf{s})$ is constant on \mathcal{S}_+^{p-1} which implies that \mathbf{s}_{\max} and $\widehat{\mathbf{s}}_{\max}$ coincide with \mathcal{S}_+^{p-1} . With this convention $\widehat{\mathbf{s}}_{\max} = \mathbf{s}_{\max}$ for all N , i.e., consistency is guaranteed and the limiting distribution is degenerate.

Remark 3.3. More generally, in order to deal with the discrete case we may define $\widehat{\mathbf{s}}_{\max}$ as the set of values that nearly maximizes (3.1). The proof of Theorem 3.1 can be modified to show that $\widehat{\mathbf{s}}_{\max}$ is consistent as a set for \mathbf{s}_{\max} . Such results require a careful definition of set convergence which we avoid here. It is clear, though, that convergence for discrete problems occurs at much quicker rate. Thus, interestingly, $\widehat{\mathbf{s}}_{\max}$ exhibits very different asymptotic behavior for discrete and continuous problems. It follows that the limiting distribution given in Theorem 3.1 describes a "worst-case" scenario.

Analyzing the distribution function of the RV \mathbf{W} given in Theorem 3.1 is analytically difficult, see Groeneboom and Wellner (2001) for the one dimensional case. Hence we investigate the distribution of $\widehat{\mathbf{s}}_{\max}$ by simulation. For simplicity we choose $p = 3$ and generated \mathbf{X}_i ($i = 1, \dots, n$) distributed as $N_3(\mathbf{0}, \boldsymbol{\Sigma})$ and \mathbf{Y}_j ($j = 1, \dots, m$) distributed as $N_3(\boldsymbol{\delta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (1 - \rho)\mathbf{I} + \rho\mathbf{J}$, \mathbf{I} is the identity matrix and \mathbf{J} is a square matrix of 1's. We simulated 1000 realizations of $\widehat{\mathbf{s}}_{\max}$ for various sample sizes and correlation coefficients. To get a visual description of the density of $\widehat{\mathbf{s}}_{\max}$, we provide a pair of plots for each configuration of ρ and sample size n . In Figure 1 we provide the joint density of the 2-dimensional polar angles (θ, ϕ) of $\widehat{\mathbf{s}}_{\max}$. There are four panels in Figure 1, corresponding to all combinations of $\rho = 0, 0.9$ and $n = 10, 100$. The mean vector $\boldsymbol{\delta}$ in this plot was taken to be $\boldsymbol{\delta} = (2, 2, 2)^T$. In Figure 2 we provide the density of the polar residual defined by $1 - \widehat{\mathbf{s}}_{\max}^T \mathbf{s}_{\max}$. The four panels of Figure 2 correspond to all combinations of $\rho = 0, 0.9$ and $n = 10, 100$ and two patterns of $\boldsymbol{\delta}$, namely, $(2, 2, 2)^T$ and $(3, 2, 1)^T$. We see from Figure 1 that $\widehat{\mathbf{s}}_{\max}$ converges to a unimodal distribution as the sample size increases. Interestingly, from Figure 2 we see that the concentration of the distribution around the true parameter depends upon the values of $\boldsymbol{\delta}$ and ρ (which together determine \mathbf{s}_{\max}). If the components of the underlying random vector are exchangeable (e.g., $\boldsymbol{\delta} = (2, 2, 2)^T$) the residuals tend to concentrate more closely around zero (Figure 2(a) and 2(c)) compared to the case when they are not exchangeable (Figure 2(b) and 2(d)).

Figures 1 & 2 Come Here

3.2.2. A confidence set for $\widehat{\mathbf{s}}_{\max}$

Since the parameter space is the surface of a unit sphere it is natural to define the $(1 - \alpha) \times 100\%$ confidence set for \mathbf{s}_{\max} centered at $\widehat{\mathbf{s}}_{\max}$ by

$$\{\mathbf{s} \in \mathcal{S}_+^{p-1} : \widehat{\mathbf{s}}_{\max}^T \mathbf{s} \leq C_{\alpha, N}\}$$

where $C_{\alpha, N}$ satisfies $\mathbb{P}(\widehat{\mathbf{s}}_{\max}^T \mathbf{s} \leq C_{\alpha, N}) = 1 - \alpha$. For more details see Fisher and Hall (1989) or Peddada and Chang (1996). Hence the confidence set is the set of all $\mathbf{s} \in \mathcal{S}_+^{p-1}$ which have a small angle with $\widehat{\mathbf{s}}_{\max}$. In theory one may appeal to Theorem 3.1 to derive the critical value for any $\alpha \in (0, 1)$. However the limit law in Theorem 3.1 requires knowledge of unknown parameters and the maximization of a stochastic process. In addition owing to the slow convergence rate and the small sample sizes observed in practice, the asymptotic confidence regions may not be accurate. For this reason, we explore the bootstrap for estimating $C_{\alpha, N}$. It is known that for cube root problems the standard bootstrap may yield confidence intervals which have coverage probabilities that are different from the nominal ones (Abrevaya and Huang 2005, Sen et al. 2010). However, the situation can be corrected by using the "M out of N bootstrap" methodology (cf. Lee, 1999, Delgado et al., 2001). Here N refers to the total sample size, i.e., $N = n + m$, and M is the size of the bootstrap subsample. Furthermore in our simulation study the group sizes are equal, i.e., $n = m$ and therefore we subsample equally from both groups. For large N , there are formal methods available for optimum choice of M (Bickel and Sackov, 2008). However, in our application N is not large and we experimented with a variety of ratios of M/N .

3.3. Testing for order

Consider testing the hypothesis

$$H_0 : \mathbf{X} =_{st} \mathbf{Y} \text{ versus } H_1 : \mathbf{X} \prec_{st} \mathbf{Y}. \quad (3.7)$$

Thus (3.7) tests whether the distributions of \mathbf{X} and \mathbf{Y} are equal or ordered (later on we briefly discuss testing $H_0 : \mathbf{X} \preceq_{st} \mathbf{Y}$ versus $H_1 : \mathbf{X} \not\prec_{st} \mathbf{Y}$). In this section we propose a new test for detecting an ordering among two multivariate distributions based on the maximal separating direction. The test is based on the following observation:

Theorem 3.2. *Let \mathbf{X} and \mathbf{Y} be independent and continuous RVs. If $\mathbf{X} =_{st} \mathbf{Y}$ then $\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}) = 1/2$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$ and if both (i) $\mathbf{X} \preceq_{st} \mathbf{Y}$, and (ii) $\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}) > 1/2$ for some $\mathbf{s} \in \mathcal{S}_+^{p-1}$, hold, then $\mathbf{X} \prec_{st} \mathbf{Y}$.*

Theorem 3.2 says that if it is known a priori that $\{\mathbf{X} \preceq_{st} \mathbf{Y}\} = \{\mathbf{X} =_{st} \mathbf{Y}\} \cup \{\mathbf{X} \prec_{st} \mathbf{Y}\}$, i.e., the RVs are either equal or ordered (which is exactly what (3.7) implies) then a strict linear statistical ordering implies a strict ordering by the usual multivariate stochastic order. In particular under the alternative there must exist a direction $\mathbf{s} \in \mathcal{S}_+^{p-1}$ for which $\mathbf{s}^T \mathbf{X} \prec_{l-st} \mathbf{s}^T \mathbf{Y}$.

Remark 3.4. *The assumption that $\mathbf{X} \preceq_{st} \mathbf{Y}$ is natural in applications such as environmental sciences where high exposures are associated with increased risk. Nevertheless if the assumption that $\mathbf{X} \preceq_{st} \mathbf{Y}$ is not warranted then the alternative hypothesis formulated in terms of the linear stochastic order actually tests whether there exists a $\mathbf{s} \in \mathcal{S}_+^{p-1}$ for which $\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}) > 1/2$. This amounts to a precedence (or Pitman) ordering between $\mathbf{s}^T \mathbf{X}$ and $\mathbf{s}^T \mathbf{Y}$.*

Remark 3.5. *In the proof of Theorem 3.2 we use the fact that given that $\mathbf{X} \preceq_{st} \mathbf{Y}$ we have $\mathbf{X} \prec_{st} \mathbf{Y}$ provided $X_i \prec_{st} Y_i$ for some $1 \leq i \leq p$. Note that if $X_i \prec_{st} Y_i$ then $\mathbb{E}(X_i) < \mathbb{E}(Y_i)$. Thus it is possible to test (3.7) by comparing means (or any other monotone function of the*

data). Although such a test will be consistent it may lack power because tests based on means are often far from optimal when the data is not normally distributed. The WMW procedure, however, is known to have high power for a broad collection of underlying distributions.

Hence (3.7) can be reformulated in terms of the linear stochastic. In particular it justifies using the statistic

$$S_{n,m} = N^{1/2}(\Psi_{n,m}(\widehat{\mathbf{s}}_{\max}) - 1/2). \quad (3.8)$$

To the best of our knowledge this is the first general test for multivariate ordered distributions. In practice we first estimate $\widehat{\mathbf{s}}_{\max}$ and then define $\widehat{U}_i = \widehat{\mathbf{s}}_{\max}^T \mathbf{X}_i$ and $\widehat{V}_j = \widehat{\mathbf{s}}_{\max}^T \mathbf{Y}_j$ where $i = 1, \dots, n$ and $j = 1, \dots, m$. Hence (3.8) is nothing but a WMW test based on the \widehat{U} 's and \widehat{V} 's. It is also a Kolmogorov-Smirnov type test.

The large sample distribution of (3.8) is given in the following.

Theorem 3.3. *Suppose the null (3.7) holds. Let $n, m \rightarrow \infty$ and $n/(n+m) \rightarrow c \in (0, 1)$. Then,*

$$S_{n,m} \Rightarrow S = \sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} \mathbb{G}(\mathbf{s})$$

where $\mathbb{G}(\mathbf{s})$ is a zero mean Gaussian process with covariance function given by (7.19).

Remark 3.6. *The evaluation of probabilities of the type $\mathbb{P}(S > u)$ is discussed in detail by Adler and Taylor (2010). However, computing $\mathbb{P}(S > u)$ requires knowledge of unknown parameters and the computation of a complicated covariance function. Therefore we implement (3.8) using bootstrap methodology.*

Remark 3.7. *Since $\widehat{\mathbf{s}}_{\max} \xrightarrow{a.s.} \mathbf{s}_{\max}$ a Slutsky type argument shows that the power of the test (3.8) converges to the power of a WMW test comparing the samples $(\mathbf{s}_{\max}^T \mathbf{X}_1, \dots, \mathbf{s}_{\max}^T \mathbf{X}_n)$ and $(\mathbf{s}_{\max}^T \mathbf{Y}_1, \dots, \mathbf{s}_{\max}^T \mathbf{Y}_m)$. The "synthetic" test assuming that \mathbf{s}_{\max} is known serves as a gold standard as verified by our simulation study.*

Remark 3.8. *Furthermore, the power of the test under local alternatives, i.e., when $\mathbf{Y} =_{st} \mathbf{X} + N^{-1/2} \boldsymbol{\delta}$ and $N \rightarrow \infty$ is bounded by the power of the WMW test comparing the distributions of $\mathbf{s}_{\max}^T \mathbf{X}$ and $\mathbf{s}_{\max}^T \mathbf{Y} = \mathbf{s}_{\max}^T \mathbf{X} + N^{-1/2} \mathbf{s}_{\max}^T \boldsymbol{\delta}$.*

Alternatives to the "sup" statistic (3.8) are the "integrated" statistics

$$I_{n,m} = \int_{\mathbf{s} \in \mathcal{S}_+^{p-1}} [N^{1/2}(\Psi_{n,m}(\mathbf{s}) - 1/2)] d\mathbf{s} \quad \text{and} \quad I_{n,m}^+ = \int_{\mathbf{s} \in \mathcal{S}_+^{p-1}} [N^{1/2}(\Psi_{n,m}(\mathbf{s}) - 1/2)]_+ d\mathbf{s} \quad (3.9)$$

where $[x]_+ = \max(0, x)$. It is clear that $I_{n,m} \Rightarrow N(0, \sigma^2)$ where $\sigma^2 = \int_{\mathbf{u} \in \mathcal{S}_+^{p-1}} \int_{\mathbf{v} \in \mathcal{S}_+^{p-1}} C(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v}$ and $C(\mathbf{u}, \mathbf{v})$, the covariance function of \mathbb{G} , is given by (7.19). Also

$$I_{n,m}^+ \Rightarrow \int_{\mathbf{s} \in \mathcal{S}_+^{p-1}} [\mathbb{G}(\mathbf{s})]_+ d\mathbf{s}.$$

This distribution does not have a closed form. The statistics $I_{n,m}$ and $I_{n,m}^+$ have univariate analogues, cf. Davidov and Herman (2012). Finally,

Theorem 3.4. *The tests (3.8) and (3.9) are consistent. Furthermore if $\mathbf{X} \preceq_{l-st} \mathbf{Y} \preceq_{l-st} \mathbf{Z}$ then all three tests for $H_0 : \mathbf{X} =_{st} \mathbf{Z}$ versus $H_1 : \mathbf{X} \prec_{st} \mathbf{Z}$ are more powerful than the respective tests for $H_0 : \mathbf{X} =_{st} \mathbf{Y}$ versus $H_1 : \mathbf{X} \prec_{st} \mathbf{Y}$.*

Theorem 3.4 shows that the tests are consistent and that their power function is monotone in the linear stochastic order.

4. Simulations

4.1. Study design

The simulation study consists of three parts. In the first part we evaluate the accuracy and precision of $\widehat{\mathbf{s}}_{\max}$ by estimating its bias and mean squared error (MSE). In the second part we investigate the coverage probability of M out of N bootstrap confidence intervals. In the third part we estimate type I errors and powers of the proposed test $S_{n,m}$ as well as the integral tests $I_{n,m}$ and $I_{n,m}^+$.

To evaluate the bias and MSEs we generated $\mathbf{X}_1, \dots, \mathbf{X}_n \sim N_3(\mathbf{0}, \Sigma)$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_m \sim N_3(\boldsymbol{\delta}, \Sigma)$ where $n = m = 20$ or 100 observations. The common variance matrix is assumed to have intra-class correlation structure, i.e., $\Sigma = (1 - \rho)\mathbf{I} + \rho\mathbf{J}$ where \mathbf{I} is the identity matrix and \mathbf{J} is a matrix of ones. Various patterns of the mean vectors $\boldsymbol{\delta}$ and correlation coefficient ρ were considered as described in Table 1.

We conducted extensive simulation studies to evaluate the performance of M out of N bootstrap confidence intervals in the present context. Specifically, motivated by our applications, we were interested in studying the performance of M out of N bootstrap when the sample sizes m and n were small. This simulation study is therefore unlike those conducted in Lee (1999), Delgado et al. (2001) and Bickel and Sakov (2008) where the sample size N was large, i.e., 100 and even 10000 in some cases. In this paper we present a small sample of our study. Our goal was to evaluate the coverage probability of M out of N bootstrap confidence intervals for different values of M/N . We generated data from two 5-dimensional normal populations with means $\mathbf{0}$ and $\boldsymbol{\delta}$, respectively and a common covariance $\Sigma = (1 - \rho)\mathbf{I} + \rho\mathbf{J}$. We considered 5 patterns of ρ , 2 patterns of sample sizes ($n = m = 20$ and 40) and different patterns of M/N . In the case of $n = m = 20$ we considered two patterns of M/N , namely, 1/2 and 1. In the case of $n = m = 40$ we considered three patterns of M/N , namely, 1/4, 1/2 and 1. The nominal coverage probability was 0.95. Results are summarized in Table 2.

The goal of the third part of our simulation study was to evaluate the type I error and the power of the test (3.8). To evaluate the type I error three different baseline distributions for the two populations \mathbf{X} and \mathbf{Y} were employed as follows: (1) both distributed as $N(\mathbf{0}, \Sigma)$; (2) both distributed as $\pi N(\mathbf{0}, \Sigma) + (1 - \pi)N(\boldsymbol{\delta}, \Sigma)$ with $\pi = 0.2$ or $\pi = 0.8$; and (3) both distributed as $\exp(\mathbf{Z}) = (\exp(Z_1), \dots, \exp(Z_p))$ where \mathbf{Z} follows a $N(\boldsymbol{\delta}, \Sigma)$. We refer to this distribution as the multivariate lognormal distribution. Throughout the variance matrix is assumed to have the intra-class structure described above. Various patterns of the mean vectors $\boldsymbol{\delta}$ and correlation coefficient ρ the dimension p were considered as described in Table 3. Sample sizes of $n = m = 15$ or 25 are reported.

Power comparisons were carried out for data generated from $\mathbf{X}_1, \dots, \mathbf{X}_n \sim N_p(\mathbf{0}, \Sigma)$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_m \sim N_p(\boldsymbol{\delta}, \Sigma)$ where $p = 3$ or 5 and a variety of patterns for $\boldsymbol{\delta}$ as described in Table 4. If Roy's maximal separating direction (cf. Proposition 3.1) was known then

a “natural gold standard” would be the test based on $\Psi_{n,m}(\mathbf{s}_{\max})$. We shall refer to this test as the true maximal direction (TMD) test. Clearly the TMD test cannot be used in practice since it involves the unknown direction \mathbf{s}_{\max} . Nevertheless the TMD test provides an upper bound for the power of the proposed test which uses the estimated direction. Hence we compute the efficiency of the proposed test relative to TMD test. An additional test, referred to as the RMD test is also compared. The RMD test has the same form but uses Roy’s maximal direction given by $\mathbf{S}^{-1}(\bar{\mathbf{Y}} - \bar{\mathbf{X}})$. As suggested by a reviewer we also evaluated the power of the two integral based tests, described in Remark ??, which do not require the determination of the best separating direction.

Additionally, in Table 5 we evaluate the type I error and power of our test when $\mathbf{X}_1, \dots, \mathbf{X}_n \sim N_p(\mathbf{0}, \Sigma)$ and $\mathbf{Y}_1, \dots, \mathbf{Y}_m \sim N_p(\boldsymbol{\delta}, \Sigma)$ and $n = m = p = 10$ and $n = m = 10$ and $p = 20$ (i.e., $p < n$ set up). Note that in neither of these cases the standard Hotteling’s T^2 (or Roy’s largest root test) can be computed whereas the proposed test can be calculated.

Simulation results reported in this paper are based on 1000 simulation runs. Confidence sets are calculated using 1000 bootstrap samples, The bootstrap critical values for estimating type I error were based on 500 bootstrap samples. Since the results between 100 bootstrap samples and 500 bootstrap samples did not differ by much, all powers were estimated using 100 bootstrap samples.

4.2. Simulation results

The Bias and MSEs for the patterns considered are summarized in Table 1.

Table 1 Comes Here

It is clear that the bias decreases with the sample size as do the MSEs. We observe that the bias tends to be smaller under independence and negative dependence compared with positive dependence. It also tends to be smaller when the data are exchangeable. It is interesting to note that some patterns exhibit cube root decreases in the MSEs while other patterns seem to behave more like square root problems. This is likely due to the non-standard asymptotics behavior of $\hat{\mathbf{s}}_{\max}$ which exhibits different rates of convergence for discrete and continuous problems. Although results are not presented, we evaluated squared bias and MSE for larger values of p (e.g., $p = 5, 10$ and 20) and as expected the total squared bias and total MSE increased with the dimension p .

In Table 2 we summarize the estimated coverage probabilities of the M out of N bootstrap methodology for different choices of M and N . Although we have considered a variety of patterns of m, n and p ; in Table 2 we report only those corresponding to $p = 5$. Observe that the coverage probabilities were closest to the nominal level of when $M = N$. The results improved as N increased. It is well known that even in the context parametric constrained inference confidence interval estimation is a challenging problem (cf. Peddada 1997, Andrews, 2001). At the moment there is no satisfactory bootstrap solution even in the (constrained) parametric inference setting. The problem is exacerbated in the present nonparametric setting due to the unusual asymptotics. Thus for small sample sizes it appears to us that the existing bootstrap methodology is not very satisfactory. The estimation of confidence intervals under order restrictions continues to be an elusive, but important problem, especially when the sample sizes are small.

Table 2 Comes Here

Type I errors for different patterns considered in our simulation study are summarized in Table 3.

Table 3 Comes Here

It is important to recognize that although the standard bootstrap may not be valid for constructing confidence intervals, it is valid for testing the null hypothesis (3.7). Our simulation studies suggest that in every case the proposed bootstrap based test maintains the nominal level of 0.05. In general it is slightly conservative. The performance of the test is not affected by the shape of the underlying distribution. That is not surprising owing to the nonparametric nature of the test. Furthermore, we evaluated the type I error of the proposed bootstrap test for testing the null hypothesis (3.7) for p as large as 20 with $n = m = 10$ and discovered that the proposed test attains the nominal level of 0.05 even $n \leq p$. See Table 4. As commented earlier in the paper Hotelling's T^2 statistic can not be applied here since the Wishart matrix is singular in this case. However, the proposed method is still applicable since the estimation of the best direction does not require the inversion of a matrix.

Tables 4 Comes Here

The power of the tests (3.8) and (3.9) for various patterns considered in our simulation study are summarized in Table 5a and 5b.

Tables 5a and 5b Come Here

As expected, in every case the power of the TMD test is higher than that of $S_{n,m}$ test and the RMD test. The $S_{n,m}$ test is almost always more powerful than the RMD test. The relative efficiency of $S_{n,m}$ compared to the TMD test is quite high in most cases. When $n = m = 15$ the relative efficiency ranges between 65–95%. It is almost always above 90% when the sample size increases to 25 per group. In general the two integral tests had very similar power. They had larger power than $S_{n,m}$ when $\rho < 0$. As ρ increased the power of $S_{n,m}$ improved relative to the two integral tests. The test (3.8) seems to perform better when the components of $\boldsymbol{\delta}$ were unequal. We also note that when the integral tests outperform $S_{n,m}$ the difference is usually small whereas the $S_{n,m}$ test can outperform the integral tests substantially. For example, observe pattern 2 where the powers of $S_{n,m}$ and $I_{n,m}$ are 0.93 and 0.97 respectively when $\rho = -0.25$ and 0.63 versus 0.40 when $\rho = .90$.

5. Illustration

Prior to conducting a 2-year rodent cancer bioassay to evaluate the toxicity/carcinogenicity of a chemical, the National Toxicology Program (NTP) routinely conducts a 90-day pre-chronic dose finding study. One of the goals of the 90-day study is to determine the maximum tolerated dose (MTD) that can be used in the 2-year chronic exposure study. Accurate determination of the MTD is critical for the success of the 2-year cancer bioassay. Cancer bioassays are typically very expensive and time consuming. Therefore their proper design, i.e., choosing the correct dosing levels, is very important. When the highest dose used in the 2-year study exceeds the MTD, a large proportion of animals in the high dose group(s)

may die well before the end of the study and the data from such group(s) cannot be used reliably. This results in inefficiency and wasted resources.

Typically the NTP uses the 90-day study to determine the MTD on the basis of a large number of correlated endpoints that provide information regarding toxicity. These include body weight, organ weights, clinical chemistry (red blood cell counts, cell volume, hemoglobin, hematocrit, lymphocytes, etc.), histopathology (lesions in various target organs), number of deaths and so forth. The dose response data is analyzed for each variable separately using Dunnett's or the Williams' test (or their nonparametric versions, Dunn's test and Shirley's test, respectively). NTP combines results from all such analyses qualitatively and uses other biological and toxicological information when making decisions regarding the highest dose for the 2-year cancer bioassay. Analyzing correlated variables one at a time may result in loss of information. The proposed methodology provides a convenient method to combine information from several outcome variables to make comparisons between groups.

We now illustrate our methodology by re-analyzing data obtained from a recent NTP study of the chemical Citral (NTP, 2003). Citral is a flavoring agent that is widely used in a variety of food items. The NTP assigned a random sample of 10 male rats to the control group and 10 to the 1785 mg/Kg dose group. Hematological and clinical chemistry measurements such as the number of Platelets (in 1000 per L), Urea Nitrogen (UN) (in mg/dL), Alkaline Phosphatase (AP) (in IU/L), and Bile Acids (BA) (in Mol/L) were recorded on each animal at the end of the study. The NTP performed univariate analysis on each of these variables and found no significant difference between the control and dose group except for the concentration of Urea Nitrogen which was increased in the high dose group. This increase was marginally significant at the 5% level and not at all after correcting for multiplicity. We applied the proposed methodology to compare the control with the high dose group (1785 mg/Kg) in terms of all non-negative linear combinations of the above mentioned four variables. We test the null hypothesis of no difference between the control and the high dose group against the alternative that the high dose group is stochastically larger (in the above four variables) than the control group. The resulting p-value based on 10,000 bootstrap samples was 0.025, which is significant at a 5% level of significance. The estimated value of \mathbf{s}_{\max} was $(0.074, 0.986, 0.012, 0.150)^T$ and the estimated 95% confidence region is given by $\{\mathbf{s} \in \mathcal{S}_+^{p-1} : \widehat{\mathbf{s}}_{\max}^T \mathbf{s} \leq 0.93\}$. Hence the confidence set includes any \mathbf{s} which is within 21.5° degrees of $\widehat{\mathbf{s}}_{\max}$. This is a relatively large set due to the small sample sizes. Clearly our methodology appears to be sensitive to detect statistical differences which were not noted by NTP. Furthermore, our methodology allows us to infer that indeed 1785 mg/Kg dose group is larger in the multivariate stochastic order than the control group. This is a much stronger conclusion than the simple ordering of their means. Thus we believe that the proposed framework and methodology for studying ordered distributions can serve as a useful tool in toxicology and is also applicable to a wide range of other problems as alluded to in this paper.

6. Concluding remarks and some open problems

In many applications, researchers are interested in comparing two experimental conditions, e.g., a treatment and a control group, in terms of a multivariate response. In classical multivariate analysis one addresses such problems by comparing the mean vectors using Hotelling's T^2 statistic. The assumption of MVN, underlying Hotelling's T^2 test, may not hold in prac-

tice. Moreover if the data is not MVN then the comparison of population means may not always provide complete information regarding the differences between the two experimental groups. Secondly, Hotelling's T^2 statistics is designed for two-sided alternatives and may not be ideal if a researcher is interested in one sided, i.e., ordered, alternatives. Addressing such problems requires one to compare the two experimental groups nonparametrically in terms of the multivariate stochastic order. Such comparisons, however, are very high dimensional and not easy to perform.

In this article we circumvent this challenge by considering the notion of the linear stochastic order between two random vectors. The linear stochastic order is a "weak" generalization of the univariate stochastic order. The linear stochastic order is simple to interpret and has an intuitive appeal. Using this notion of ordering, we developed nonparametric directional inference procedures. Intuitively, the proposed methodology seeks to determine the direction that best separates two multivariate populations. Asymptotic properties of the estimated direction are derived. Our test based on the best separating direction may be viewed as a generalization of Roy's classical largest root test for comparing several MVN populations. To the best of our knowledge this is the first general test for multivariate ordered distributions. Since in practice sample sizes are small, we use the bootstrap methodology for drawing inferences.

We illustrated the proposed methodology using a data obtained from a recent toxicity/carcinogenicity study conducted by the US National Toxicology Program (NTP) on the chemical Citral. A re-analysis of their 90-day data using our proposed methodology revealed a linear stochastic increase in Platelets, Urea Nitrogen, Alkaline Phosphatase, and Bile Acids in the high dose group relative to the control group, which was not seen in the original univariate analysis conducted by the NTP. These findings suggest that the proposed methodology may have greater sensitivity than the commonly used univariate statistical procedures. Our methodology is sufficiently general since it is nonparametric and can be applied to discrete and/or continuous outcome variables. Furthermore, our methodology exploits the underlying dependence structure in the data, rather than analyzing one variable at a time.

We note that our example and some of our results pertain to continuous RVs. However the methodology may be used, with appropriate modification (e.g., methods for dealing with ties) with discrete (or mixed) data with no problem. Although the focus of this paper has been the comparison of two multivariate vectors, in many applications, especially in dose response studies, researchers may be interested in determining trends (order) among several groups. Similar to classical parametric order restricted inference literature one could generalize the methodology developed in this paper to test for order restrictions among multiple populations. For example one could extend the results to $K \geq 2$ RVs ordered by the simple ordering, i.e., $\mathbf{X}_1 \prec_{l-st} \mathbf{X}_2 \prec_{l-st} \cdots \prec_{l-st} \mathbf{X}_K$ or to RVs ordered by the tree ordering, i.e., $\mathbf{X}_1 \prec_{l-st} \mathbf{X}_j$ where $j = 2, \dots, K$. As pointed out by a referee the hypotheses $H_0 : \mathbf{X} \preceq_{st} \mathbf{Y}$ versus $H_1 : \mathbf{X} \not\prec_{st} \mathbf{Y}$ can also be formulated and tested using the approach described. First note that the null hypothesis implies $\Psi(\mathbf{s}) \geq 1/2$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$. On the other hand under the alternative there is an $\mathbf{s} \in \mathcal{S}_+^{p-1}$ for which $\Psi(\mathbf{s}) < 1/2$. Thus a test may be based on the statistic

$$N^{1/2}(\Psi_{n,m}(\widehat{\mathbf{s}}_{\min}) - 1/2)$$

where $\widehat{\mathbf{s}}_{\min}$ is the value which minimizes $\Psi_{n,m}(\mathbf{s})$. It is also clear that the least favorable configuration occurs when $\Psi(\mathbf{s}) = 1/2$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$ which is equivalent to $\mathbf{X} =_{st} \mathbf{Y}$.

We believe that the result obtained here may be useful beyond order restricted inference.

Our simulation study suggests that our estimator of the best separating direction, i.e., (3.4), may be useful even in the context of classical multivariate analysis where it may be viewed as a robust alternative to Roy's classical estimate. Finally we note that the linear stochastic order may be useful in a variety of other statistical problems. For example, we believe that it provide a useful framework for linearly combining the results of several diagnostic markers. This is a well known problem in the context of ROC curve analysis in diagnostic medicine.

ACKNOWLEDGMENTS

The research of Ori Davidov was partially supported by the Israeli Science Foundation Grant No 875/09 and was conducted in part when visiting Shyamal Das Peddada at the NIEHS. This research [in part] was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES101744-04). We thank Grace Kissling (NIEHS), Yair Goldberg and Danny Segev (University of Haifa), for their useful comments and suggestions.

References

- [1] Abrevaya J, Haung J (2005). On the Bootstrap of the maximum score estimator. *Econometrica*, 73: 1175-1204.
- [2] Adler RJ, Taylor JE (2010). *Random Fields and Geometry*. Springer.
- [3] Andrews DWK (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68: 399-406.
- [4] Arcones MA, Kvam PH, Samaniego FJ (2002). Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *Journal of the American Statistical Association*, 97: 170-182.
- [5] Audet C, Becharad V, Le Digabel S (2008). Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search. *Journal of Global Optimization*, 41:299–318.
- [6] Bekele BN, Thall PF (2004). Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *Journal of the American Statistical Association*, 99: 26-35.
- [7] Bickel P, Sackov A (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, 18: 967-985.
- [8] DasGupta A (2008). *Asymptotic Theory of Statistics and Probability*. Springer.
- [9] Davey BA, Priestley HA (2002). *Introduction to Lattices and Order*. Cambridge University Press.
- [10] Davidov O, Herman A (2011). Multivariate stochastic orders induced by case-control sampling. *Methodology and Computing in Applied Probability*, 13: 139-154.

- [11] Davidov O, Peddada SD (2011). Order restricted inference for multivariate binary data with application to toxicology. *Journal of the American Statistical Association*, 106: 1394-1404.
- [12] Davidov O (2012). Ordered inference, rank statistics and combining p-values: a new perspective. *Statistical Methodology*, 9: 456-465.
- [13] Davidov O, Herman A (2012). Ordinal dominance curve based inference for stochastically ordered distributions. *Journal of the Royal Statistical Society, Series B*, In Press, DOI: 10.1111/j.1467-9868.2012.01031.x.
- [14] Delagdo MA, Rodriguez-Poo JM, Wolf M (2001). Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator. *Economics Letters*, 73: 241-250.
- [15] Ding Y, Zhang X (2004). Some stochastic orders for Kotz type distributions. *Statistics and Probability Letters*, 69: 389-396.
- [16] Fang KT, Kots S, Ng KW (1989). *Symmetric Multivariate and Related Distributions*. Chapman and Hall.
- [17] Fisher NI, Hall P (1989). Bootstrap confidence regions for directional data. *Journal of the American Statistical Association*, 84: 996-1002.
- [18] Groeneboom P, Wellner JA (2001). Computing Chernoff's distribution. *Journal of Computational and Graphical Statistics*, 10: 388-400.
- [19] Hajek J, Sidak Z, Sen PK (1999). *Theory of Rank Tests*. Academic Press.
- [20] Hu J, Homem-de-Mello T, Mehrotra S (2011). Concepts and applications of stochastically weighted stochastic dominance. Unpublished manuscript, see http://www.optimization-online.org/DB_FILE/2011/04/2981.pdf
- [21] Ivanova A, Murphy M (2009). An adaptive first in man dose-escalation study of NGX267: statistical, clinical, and operational considerations. *Journal of Biopharmaceutical Statistics*, 19: 247-255.
- [22] Joe H (1997). *Multivariate Models and Dependence Concepts*. Chapman and Hall.
- [23] Johnson R, Wichern D (1998). *Applied Multivariate Statistical Analysis*. Prentice Hall.
- [24] Kim J, Pollard D (1990). Cube root asymptotics. *The Annals of Statistics*, 18: 191-219.
- [25] Kosorok MR (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer.
- [26] Lee S (1999). On a class of m out of n bootstrap confidence intervals. *Journal of the Royal Statistical Society, Series B*, 61: 901-911.
- [27] Lucas LA, Wright FT (1991). Testing for and against stochastic ordering between multivariate multinomial populations. *Journal of Multivariate Analysis*, 38: 167-186.

- [28] Moser V (2000). Observational batteries in neurotoxicity testing. *International Journal of Toxicology*, 19: 407-411.
- [29] Neumayer N (2004). A central limit theorem for two sample U-processes. *Statistics and Probability Letters*, 67: 73-85.
- [30] NTP (2003). NTP toxicology and carcinogenesis studies of citral (microencapsulated) (CAS No. 5392-40-5) in F344/N rats and B6C3F1 mice (feed studies).
- [31] Peddada SD (1985). A short note on the Pitman measure of nearness. *American Statistician*, 39: 298-299.
- [32] Peddada SD, Chang T (1996). Bootstrap confidence region estimation of the motion of rigid bodies. *Journal of the American Statistical Association*, 91: 231-241.
- [33] Peddada SD (1997). Confidence interval estimation of population means Subject to order restrictions using resampling procedures. *Statistics and Probability Letters*, 31: 255-265.
- [34] Pitman EJM (1937). The closest estimates of statistical parameters. *Proceedings of the Cambridge Philosophical Society*, 33: 212-222.
- [35] Price CJ, Reale M, Robertson BL (2008). A direct search method for smooth and non-smooth unconstrained optimization problems. *Australian & New Zealand Industrial and Applied Mathematics Journal*, 48: 927-948.
- [36] Roy SN (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, 24: 220-238.
- [37] Sampson AR, Whitaker LR (1989). Estimation of multivariate distributions under stochastic ordering. *Journal of the American Statistical Association*, 84: 541-548.
- [38] Serfling RJ (1980). *Approximation Theorem of Mathematical Statistics*. Wiley.
- [39] Sen B, Banerjee M, Woodroffe M (2010). Inconsistency of bootstrap: The Grenander estimator. *The Annals of Statistics*, 38: 1953-1977.
- [40] Shaked M, Shanthikumar JG (2007). *Stochastic Orders*. Springer.
- [41] Silvapulle MJ, Sen PK (2005). *Constrained Statistical Inference*. Wiley & Sons.
- [42] van der Vaart AW (2000). *Asymptotic Statistics*. Cambridge University Press.

7. Proofs

Proof of Theorem 2.1:

Proof. (i) Let $g : \mathbb{R}^p \rightarrow \mathbb{R}^n$ be a linear increasing function. Clearly $g(\mathbf{x}) = \mathbf{v} + \mathbf{M}\mathbf{x}$ for some n vector \mathbf{v} and $n \times p$ matrix \mathbf{M} with non-negative elements. Thus for any $\mathbf{u} \in \mathbb{R}_+^p$ we have $\mathbf{s} = \mathbf{M}^T \mathbf{u} \in \mathbb{R}_+^n$. Hence

$$\mathbf{u}^T g(\mathbf{X}) = \mathbf{u}^T (\mathbf{v} + \mathbf{M}\mathbf{X}) = \mathbf{u}^T \mathbf{v} + \mathbf{s}^T \mathbf{X} \preceq_{st} \mathbf{u}^T \mathbf{v} + \mathbf{s}^T \mathbf{Y} = \mathbf{u}^T (\mathbf{v} + \mathbf{M}\mathbf{Y}) = \mathbf{u}^T g(\mathbf{Y})$$

as required where the inequality holds because $\mathbf{X} \preceq_{l-st} \mathbf{Y}$. (ii) Fix $I \in \{1, \dots, p\}$. Let $\mathbf{X} = (\mathbf{X}_I, \mathbf{X}_{\bar{I}})$, $\mathbf{Y} = (\mathbf{Y}_I, \mathbf{Y}_{\bar{I}})$ where \bar{I} is the complement of I in $\{1, \dots, p\}$. Further define, $\mathbf{s}^T = (\mathbf{s}_I^T, \mathbf{s}_{\bar{I}}^T)$ where $\mathbf{s} \in \mathbb{R}_+^p$ and set $\mathbf{s}_{\bar{I}}^T = 0$. It follows that for all $\mathbf{s}_I \in \mathbb{R}^{\dim(I)}$ we have

$$\mathbf{s}_I^T \mathbf{X}_I = \mathbf{s}^T \mathbf{X} \preceq_{st} \mathbf{s}^T \mathbf{Y} = \mathbf{s}_I^T \mathbf{Y}_I$$

as required. (iii) Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$ be any increasing function. Note that

$$\mathbb{E}(\phi(\mathbf{s}^T \mathbf{X})) = \mathbb{E}(\mathbb{E}(\phi(\mathbf{s}^T \mathbf{X}) | \mathbf{Z})) \leq \mathbb{E}(\mathbb{E}(\phi(\mathbf{s}^T \mathbf{Y}) | \mathbf{Z})) = \mathbb{E}(\phi(\mathbf{s}^T \mathbf{Y})).$$

The inequality is a consequence of $\mathbf{X} | \mathbf{Z} = \mathbf{z} \preceq_{l-st} \mathbf{Y} | \mathbf{Z} = \mathbf{z}$. Since ϕ is arbitrary it follows that $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ as required. (iv) Let $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ and define \mathbf{Y} similarly. Let $\mathbf{s} \in \mathbb{R}_+^p$ where $p = p_1 + \dots + p_n$. Now

$$\mathbf{s}^T \mathbf{X} = \mathbf{s}_1^T \mathbf{X}_1 + \dots + \mathbf{s}_n^T \mathbf{X}_n \text{ and } \mathbf{s}^T \mathbf{Y} = \mathbf{s}_1^T \mathbf{Y}_1 + \dots + \mathbf{s}_n^T \mathbf{Y}_n$$

by assumption $\mathbf{s}_i^T \mathbf{X}_i \preceq_{st} \mathbf{s}_i^T \mathbf{Y}_i$ for $i = 1, \dots, n$. In addition $\mathbf{s}_i^T \mathbf{X}_i$ and $\mathbf{s}_j^T \mathbf{X}_j$ are independent for $i \neq j$. It follows from Theorem 1.A.3 in Shaked and Shanthikumar (2007) that $\mathbf{s}_1^T \mathbf{X}_1 + \dots + \mathbf{s}_n^T \mathbf{X}_n \preceq_{st} \mathbf{s}_1^T \mathbf{Y}_1 + \dots + \mathbf{s}_n^T \mathbf{Y}_n$, i.e., $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ as required. (v) By assumption $\mathbf{X}_n \Rightarrow \mathbf{X}$ and $\mathbf{Y}_n \Rightarrow \mathbf{Y}$ where the symbol \Rightarrow denotes convergence in distribution. By the continuous mapping theorem $\mathbf{s}^T \mathbf{X}_n \Rightarrow \mathbf{s}^T \mathbf{X}$ and $\mathbf{s}^T \mathbf{Y}_n \Rightarrow \mathbf{s}^T \mathbf{Y}$. It follows that

$$\mathbb{P}(\mathbf{s}^T \mathbf{X}_n \geq t) \rightarrow \mathbb{P}(\mathbf{s}^T \mathbf{X} \geq t) \text{ and } \mathbb{P}(\mathbf{s}^T \mathbf{Y}_n \geq t) \rightarrow \mathbb{P}(\mathbf{s}^T \mathbf{Y} \geq t). \quad (7.1)$$

Moreover since $\mathbf{X}_n \preceq_{l-st} \mathbf{Y}_n$ we have

$$\mathbb{P}(\mathbf{s}^T \mathbf{X}_n \geq t) \leq \mathbb{P}(\mathbf{s}^T \mathbf{Y}_n \geq t) \text{ for all } n \in \mathbb{N}. \quad (7.2)$$

Combining (7.1) and (7.2) we have $\mathbb{P}(\mathbf{s}^T \mathbf{X} \geq t) \leq \mathbb{P}(\mathbf{s}^T \mathbf{Y} \geq t)$, i.e., $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ as required. ■

Before proving Theorem 2.2 we provide a definition and a preliminary Lemma.

Definition 7.1. We say that the RV \mathbf{X} has an elliptical distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and generator $\phi(\cdot)$, denoted $\mathbf{X} \sim E_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$, if its characteristic function is given by $\exp(i\mathbf{t}^T \boldsymbol{\mu}) \phi(\mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t})$.

For this and other facts about elliptical distributions which we use in the proofs below see Fang et al. (1987).

Lemma 7.1. Let $X \sim E_1(\mu, \sigma, \phi)$ and $Y \sim E_1(\mu', \sigma', \phi)$ be univariate RVs supported on \mathbb{R} . Then $X \preceq_{st} Y$ if and only if $\mu \leq \mu'$ and $\sigma = \sigma'$.

Proof. Since X and Y have the same generator they have the stochastic representation, i.e.,

$$X =_{st} \mu + \sigma RU \text{ and } Y =_{st} \mu' + \sigma' RU \quad (7.3)$$

where R is a non-negative RV independent of the RV U satisfying $\mathbb{P}(U = \pm 1) = 1/2$ (cf., Fang et al. 1987). It follows that RU is a symmetric RV supported on \mathbb{R} with a strictly increasing DF which we denoted by F_0 . Let F_X and F_Y denote the DFs of X and Y

respectively. Note that $X \preceq_{st} Y$ if and only if $F_X(t) \geq F_Y(t)$ for all $t \in \mathbb{R}$, or equivalently by (7.3), if and only if

$$F_0\left(\frac{t-\mu}{\sigma}\right) \geq F_0\left(\frac{t-\mu'}{\sigma'}\right) \quad (7.4)$$

for all $t \in \mathbb{R}$. It is obvious that (7.4) holds when $\mu \leq \mu'$ and $\sigma = \sigma'$ establishing sufficiency. Now assume that $X \preceq_{st} Y$. Put $t = \mu$ in (7.4) and use the strict monotonicity of F_0 to get $0 \geq (\mu - \mu')/\sigma'$, i.e., $\mu' \geq \mu$. Suppose now that $\sigma' > \sigma$. It follows from (7.4) and the the strict monotonicity of F_0 that $(t - \mu)/\sigma \geq (t - \mu')/\sigma'$ which is equivalent to $t \geq (\mu\sigma' - \mu'\sigma) / (\sigma' - \sigma)$. The latter, however, contradicts the fact that (7.4) holds for all $t \in \mathbb{R}$. A similar argument shows that $\sigma' < \sigma$ can not hold, hence we must have $\sigma = \sigma'$ as required. ■

Remark 7.1. *Note that Lemma 7.1 may not hold for distributions with a finite support. For example if $R \sim U(0, 1)$ then by (7.3) $X \sim U(\mu - \sigma, \mu + \sigma)$ and $Y \sim U(\mu' - \sigma', \mu' + \sigma')$. It is easily verified that in this case $X \preceq_{st} Y$ if and only if $\Delta = \mu' - \mu \geq 0$ and $-\Delta \leq \sigma' - \sigma \leq \Delta$, i.e., it is not required that $\sigma = \sigma'$. Hence the assumption that X and Y are supported on \mathbb{R} is necessary.*

We continue with the proof of Theorem 2.2:

Proof. Let \mathbf{X} and \mathbf{Y} be be $E_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ and $E_p(\boldsymbol{\mu}', \boldsymbol{\Sigma}', \phi)$ supported on \mathbb{R}^p . Suppose that $\mathbf{X} \preceq_{l-st} \mathbf{Y}$. Choose $\mathbf{s} = \mathbf{e}_i$ where $\mathbf{e}_{ik} = 1$ if $i = k$ and 0 otherwise. It now follows from the definition 1.1 that $X_i \preceq_{st} Y_i$. Since X_i and Y_i are marginally elliptically distributed RVs with the same generator and supported on \mathbb{R} then by Lemma 7.1 we must have

$$\mu_i \leq \mu'_i \text{ and } \sigma_{ii} = \sigma'_{ii}. \quad (7.5)$$

The latter holds, of course, for all $1 \leq i \leq p$. Choosing $\mathbf{s} = \mathbf{e}_i + \mathbf{e}_j$ we have $X_i + X_j \preceq_{st} Y_i + Y_j$. Note that $X_i + X_j$ and $Y_i + Y_j$ are supported on \mathbb{R} and follow a univariate elliptical distribution with the same generator (Fang et al. 1987). Applying Lemma 7.1 again we find that

$$\mu_i + \mu_j \leq \mu'_i + \mu'_j \text{ and } \sigma_{ii} + \sigma_{jj} + 2\sigma_{ij} = \sigma'_{ii} + \sigma'_{jj} + 2\sigma'_{ij}. \quad (7.6)$$

The latter holds, of course, for all $1 \leq i \neq j \leq p$. It is easy to see that equations (7.5) and (7.6) imply that $\boldsymbol{\mu} \leq \boldsymbol{\mu}'$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}'$. Recall (cf. Fang et al. 1987) that we may write $\mathbf{X} =_{st} \boldsymbol{\mu} + R\mathbf{S}\mathbf{U}$ and $\mathbf{Y} =_{st} \boldsymbol{\mu}' + R\mathbf{S}\mathbf{U}$ where $\boldsymbol{\Sigma} = \mathbf{S}^T \mathbf{S}$, \mathbf{U} is a uniform RV on \mathcal{S}_+^{p-1} and R is a non-negative RV. Let S be an upper set in \mathbb{R}^p . Clearly the set $[S - \boldsymbol{\mu}] := \{\mathbf{x} - \boldsymbol{\mu} : \mathbf{x} \in S\}$ is also an upper set and $[S - \boldsymbol{\mu}] \subseteq [S - \boldsymbol{\mu}']$ since $\boldsymbol{\mu} \leq \boldsymbol{\mu}'$. Now,

$$\mathbb{P}(\mathbf{X} \in S) = \mathbb{P}(\mathbf{X}_0 \in [S - \boldsymbol{\mu}]) \leq \mathbb{P}(\mathbf{X}_0 \in [S - \boldsymbol{\mu}']) = \mathbb{P}(\mathbf{Y} \in S)$$

where $\mathbf{X}_0 = R\mathbf{S}\mathbf{U}$, hence $\mathbf{X} \preceq_{st} \mathbf{Y}$. This proves the if part. The only if part follows immediately. ■

Proof of Theorem 2.3:

Proof. Let $\mathcal{X}_p = \{\mathbf{x} : (x_1, \dots, x_p) \in \{0, 1\}^p\}$ denote the support of a p dimensional multivariate binary (MVB) RV. By definition the relationship $\mathbf{X} \preceq_{l-st} \mathbf{Y}$ implies that for all $(t, \mathbf{s}) \in \mathbb{R}_+ \times \mathbb{R}_+^p$

$$\mathbb{P}(\mathbf{s}^T \mathbf{X} > t) \leq \mathbb{P}(\mathbf{s}^T \mathbf{Y} > t). \quad (7.7)$$

Now note that

$$\mathbb{P}(\mathbf{s}^T \mathbf{X} > t) = \sum_{\mathbf{x} \in \mathcal{X}_p} f(\mathbf{x}) \mathbb{I}_{(\mathbf{s}^T \mathbf{x} > t)} \text{ and } \mathbb{P}(\mathbf{s}^T \mathbf{Y} > t) = \sum_{\mathbf{x} \in \mathcal{X}_p} g(\mathbf{x}) \mathbb{I}_{(\mathbf{s}^T \mathbf{x} > t)} \quad (7.8)$$

where f and g are the probability mass functions of \mathbf{X} and \mathbf{Y} , respectively. Let U be an upper set on \mathcal{X}_p . It is well known (cf. Davey and Priestley 2002) that U can be written as

$$U = \cup_{j \in J} U(\mathbf{x}_j) \quad (7.9)$$

where \mathbf{x}_j are the distinct minimal elements of U and $U(\mathbf{x}_j) = \{\mathbf{x} : \mathbf{x} \geq \mathbf{x}_j\}$ are themselves upper sets (in fact $U(\mathbf{x}_j)$ is an upper orthant). The set $\{\mathbf{x}_j : j \in J\}$ is often referred to as an anti-chain. Now observe that for any $\mathbf{s} \in \mathbb{R}_+^p$ the set $\{\mathbf{x} : \mathbf{s}^T \mathbf{x} > t\}$ is an upper set. Hence it must be of the form (7.9) for some anti-chain $\{\mathbf{x}_j : j \in J\}$. Suppose now, that for some $U \in \mathcal{X}_p$ there is a vector $\mathbf{s}_U \in \mathbb{R}_+^p$ such that $U = \{\mathbf{x} : \mathbf{s}_U^T \mathbf{x} > t\}$ for some fixed $t > 0$. Then using (7.7) and (7.8) we have

$$\mathbb{P}(\mathbf{X} \in U) = \sum_{\mathbf{x} \in \{\mathbf{x} : \mathbf{s}_U^T \mathbf{x} > t\}} f(\mathbf{x}) = \mathbb{P}(\mathbf{s}_U^T \mathbf{X} > t) \leq \mathbb{P}(\mathbf{s}_U^T \mathbf{Y} > t) = \sum_{\mathbf{x} \in \{\mathbf{x} : \mathbf{s}_U^T \mathbf{x} > t\}} g(\mathbf{x}) = \mathbb{P}(\mathbf{Y} \in U).$$

We will complete the proof by showing that for each upper set $U \in \mathcal{X}_p$ we can find a vector \mathbf{s}_U for which $\mathbf{s}_U^T \mathbf{x} > t$ for $\mathbf{x} \in U$ and $\mathbf{s}_U^T \mathbf{x} \leq t$ for $\mathbf{x} \in U^c = \mathcal{X} \setminus U$ if and only if $p \leq 3$. To do so we will first solve the system of equations $\mathbf{s}^T \mathbf{x}_j = t$ for $j \in J$. This system can also be written as $\mathbf{X} \mathbf{s} = \mathbf{t}$ where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_J \end{pmatrix}$$

is a $J \times p$ matrix whose rows are the member of the anti-chain defining U and $\mathbf{t} = (t, \dots, t)$ has dimension J . Clearly the elements of \mathbf{X} are ones and zeros. If $J \leq p$ the matrix \mathbf{X} is of full rank since its rows are linearly independent by the fact that they are an anti-chain. Hence a solution for \mathbf{s} exists. With a bit of algebra we can further show that a solution $\mathbf{s} \geq 0$ exists. This, of course, is trivially verified when $p \leq 3$. Now set $\mathbf{s}_U = \mathbf{s} + \boldsymbol{\varepsilon}$ for some $\boldsymbol{\varepsilon} \geq 0$. It is clear that we can choose $\boldsymbol{\varepsilon}$ small enough to guarantee that $\mathbf{s}_U^T \mathbf{x} > t$ if and only if $\mathbf{x} \in U$. Hence if $J \leq p$ the upper set (7.9) can be mapped to a vector \mathbf{s}_U . However the inequality $J \leq p$ for *all* upper sets $U \subset \mathcal{X}_p$ holds if and only if $p \leq 3$. This can be easily shown by enumerating all 18 upper sets belonging to \mathcal{X}_3 (cf. Davidov and Peddada 2011) and noting that they have at most three minimal elements. Hence if $p \leq 3$ then $\mathbf{X} \preceq_{t-st} \mathbf{Y} \iff \mathbf{X} \preceq_{st} \mathbf{Y}$ as required.

Now let $p = 4$ and consider the upper set U generated by the antichain \mathbf{x}_j , $j = 1, \dots, J$ where \mathbf{x}_j are all the distinct permutations of the vector $(1, 1, 0, 0)$. Clearly $J = 6$. Note that although $J > p$ the system of equations $\mathbf{X} \mathbf{s} = \mathbf{t}$ is uniquely solved by $\mathbf{s}_*^T = (t/2, t/2, t/2, t/2)$. However, this solution coincides with the solution of the system $\mathbf{X}' \mathbf{s} = \mathbf{t}$ where \mathbf{X}' is any matrix obtained from \mathbf{X} by deleting any two (or just one) of its rows. Note that the rows of \mathbf{X}' corresponds to an upper set $U' \subset U$. This, in turn, implies that for any such U' one can not find a vector $\mathbf{s}_{U'}$ satisfying $\mathbf{s}_{U'}^T \mathbf{x} > t$ if and only if $\mathbf{x} \in U'$ because the inequality will hold for all $\mathbf{x} \in U$. Thus U' does not define a upper half plane. This shows that the linear stochastic order and the multivariate stochastic order do not coincide when $p = 4$. A similar

argument may be used for any $p \geq 5$. This completes the proof. ■

We first define the term copula.

Definition 7.2. Let F be the DF of a p dimensional RV with marginal DFs F_1, \dots, F_p . The copula C associated with F is a DF such that

$$F(\mathbf{x}) = C(\mathbf{x}) = C(F_1(x_1), \dots, F_p(x_p)).$$

It follows that the tail-copula $\bar{C}(\cdot)$ is nothing but the tail of the DF $C(\cdot)$.

Proof of Theorem 2.4:

Proof. Suppose that \mathbf{X} and \mathbf{Y} have the same copula. Let $\mathbf{X} \preceq_{l-st} \mathbf{Y}$. Choosing $\mathbf{s} = \mathbf{e}_i$ where $\mathbf{e}_{ik} = 1$ if $i = k$ and 0 otherwise we find using the definition that $X_i \preceq_{st} Y_i$. The latter holds, of course, for all $1 \leq i \leq p$. Applying Theorem 6.B.14 in Shaked and Shanthikumar (2007) we find that $\mathbf{X} \preceq_{st} \mathbf{Y}$. The reverse direction is immediate. ■

Proof of Theorem 2.5:

Proof. Note that for any $\mathbf{x} \in \mathbb{R}^p$ we have

$$\begin{aligned} F(\mathbf{x}) &= C_{\mathbf{X}}(F_1(x_1), \dots, F_p(x_p)) \geq C_{\mathbf{X}}(G_1(x_1), \dots, G_p(x_p)) \\ &\geq C_{\mathbf{Y}}(G_1(x_1), \dots, G_p(x_p)) = G(\mathbf{x}). \end{aligned}$$

This means that $\mathbf{X} \preceq_{lo} \mathbf{Y}$. The other part of the Theorem is proved similarly. ■

Proof of Proposition 3.1:

Proof. Let \mathbf{X} and \mathbf{Y} be independent MVNs with means $\boldsymbol{\mu} \leq \boldsymbol{\nu}$ and common variance matrix $\boldsymbol{\Sigma}$. Clearly

$$\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}) = \Phi\left(-\frac{\mathbf{s}^T(\boldsymbol{\mu} - \boldsymbol{\nu})}{\sqrt{2\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}}\right).$$

where Φ is the DF of a standard normal RV. It follows that $\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y})$ is maximized when the ratio $\mathbf{s}^T(\boldsymbol{\nu} - \boldsymbol{\mu})/\sqrt{\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}$ is maximized. From the Cauchy-Schwartz inequality we have

$$\frac{\mathbf{s}^T(\boldsymbol{\nu} - \boldsymbol{\mu})}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}} \leq \sqrt{(\boldsymbol{\nu} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\nu} - \boldsymbol{\mu})} \quad (7.10)$$

for all \mathbf{s} . It is now easily verified that $\mathbf{s} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\nu} - \boldsymbol{\mu})$ maximizes the left hand side of (7.10). ■

Proof of Proposition 3.2:

Proof. Let Q_q , $q = 1, \dots, 4$ be the four quadrants. It is clear that maximizing (3.1) is equivalent to maximizing

$$\Psi'_{n,m}(\mathbf{s}) = \sum_{\mathbf{Z}_{ij} \in Q_2} \mathbb{I}_{(\mathbf{s}^T \mathbf{Z}_{ij} \geq 0)} + \sum_{\mathbf{Z}_{ij} \in Q_4} \mathbb{I}_{(\mathbf{s}^T \mathbf{Z}_{ij} \geq 0)} \quad (7.11)$$

It is also clear that for any \mathbf{s} the indicators $\mathbb{I}_{(\mathbf{s}^T \mathbf{Z}_{ij} \geq 0)}$ are independent of the length of \mathbf{Z}_{ij} which we therefore take to have length unity. Observe that the value of (7.11) is constant in the intervals $(\theta_{[i]}, \theta_{[i+1]})$ where $\theta_{[i]}$ are defined in Algorithm 3.1. At each point $\theta_{[i]}$, $i =$

$0, \dots, M+1$ the value of (7.11) may increase or decrease. It follows that for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$ $\Psi'_{n,m}(\mathbf{s}) \in \{\Psi'_{n,m}(\mathbf{s}_{[0]}), \dots, \Psi'_{n,m}(\mathbf{s}_{[M+1]})\}$ where $\mathbf{s}_{[i]}$ are defined in Algorithm 3.1. Therefore the maximum value of (3.1) is an element of the above list. Now suppose that $\mathbf{s}_{[i]}$ is a global maximizer of (7.11). Clearly either $\Psi'_{n,m}(\mathbf{s}_{[i]}) = \Psi'_{n,m}(\mathbf{s}_{[i-1]})$ or $\Psi'_{n,m}(\mathbf{s}_{[i]}) = \Psi'_{n,m}(\mathbf{s}_{[i+1]})$ must hold. In which case any value in $[\theta_{[i-1]}, \theta_{[i]}]$ or $[\theta_{[i]}, \theta_{[i+1]}]$ is a global maximizer. This concludes the proof. ■

Proof of Theorem 3.1:

Proof. Fix \mathbf{s} . Using Hajek's projection we may write

$$\Psi_{n,m}(\mathbf{s}) = \Psi(\mathbf{s}) + n^{-1} \sum_{i=1}^n \psi_1(\mathbf{X}_i, \mathbf{s}) + m^{-1} \sum_{j=1}^m \psi_2(\mathbf{Y}_j, \mathbf{s}) + R_{n,m}(\mathbf{s}) \quad (7.12)$$

Since $\Psi_{n,m}(\mathbf{s})$ is bounded $R_{n,m}(\mathbf{s}) = o(\log(N)/N^k)$ with probability one (cf. Theorem 5.3.3 in Serfling 1980) for all $k \in \mathbb{N}$. Moreover it is clear that the latter holds uniformly for all \mathbf{s} . Also

$$\begin{aligned} \psi_1(\mathbf{X}_i, \mathbf{s}) &= \bar{G}(\mathbf{s}^T \mathbf{X}_i) - \mathbb{E}(\bar{G}(\mathbf{s}^T \mathbf{X}_i)), \\ \psi_2(\mathbf{Y}_j, \mathbf{s}) &= F(\mathbf{s}^T \mathbf{Y}_j) - \mathbb{E}(F(\mathbf{s}^T \mathbf{Y}_j)), \end{aligned}$$

where $\bar{G}(\mathbf{s}^T \mathbf{x}) = \mathbb{P}(\mathbf{s}^T \mathbf{Y} \geq \mathbf{s}^T \mathbf{x})$ and $F(\mathbf{s}^T \mathbf{y}) = \mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{y})$. Clearly $\mathbb{E}[\psi_1(\mathbf{X}_i, \mathbf{s})] = \mathbb{E}[\psi_2(\mathbf{Y}_j, \mathbf{s})] = 0$ for all i and j so by the strong law of large numbers $n^{-1} \sum_{i=1}^n \psi_1(\mathbf{X}_i, \mathbf{s})$ and $m^{-1} \sum_{j=1}^m \psi_2(\mathbf{Y}_j, \mathbf{s})$ both converge to zero with probability one. Hence for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$ we have $\Psi_{n,m}(\mathbf{s}) \xrightarrow{a.s.} \Psi(\mathbf{s})$ as $n, m \rightarrow \infty$. Now,

$$\sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} |\Psi_{n,m}(\mathbf{s}) - \Psi(\mathbf{s})| \leq \sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} \left| n^{-1} \sum_{i=1}^n \psi_1(\mathbf{X}_i, \mathbf{s}) \right| + \sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} \left| m^{-1} \sum_{j=1}^m \psi_2(\mathbf{Y}_j, \mathbf{s}) \right| + \sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} |R_{n,m}(\mathbf{s})|.$$

The set \mathcal{S}_+^{p-1} is compact and the function $\psi_1(\mathbf{x}, \mathbf{s})$ is continuous in $\mathbf{s} \in \mathcal{S}_+^{p-1}$ for all values of \mathbf{x} and bounded, i.e., $|\psi_1(\mathbf{x}, \mathbf{s})| \leq 2$. Thus the conditions in Theorem 3.1 part (iv) in DasGupta (2008) are satisfied and it follows that $\sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} |n^{-1} \sum_{i=1}^n \psi_1(\mathbf{X}_i, \mathbf{s})| \xrightarrow{a.s.} 0$ as $n \rightarrow \infty$. Similarly $\sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} |m^{-1} \sum_{j=1}^m \psi_2(\mathbf{Y}_j, \mathbf{s})| \xrightarrow{a.s.} 0$ as $m \rightarrow \infty$. Thus,

$$\sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} |\Psi_{n,m}(\mathbf{s}) - \Psi(\mathbf{s})| \xrightarrow{a.s.} 0$$

as $n, m \rightarrow \infty$. The latter is a uniform LLN for a family of U-statistics. By assumption $\Psi(\mathbf{s}_{\max}) > \Psi(\mathbf{s})$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1} \setminus \mathbf{s}_{\max}$ so we can apply Theorem 5.7 in van der Vaart (2000) to conclude that

$$\widehat{\mathbf{s}}_{\max} \xrightarrow{a.s.} \mathbf{s}_{\max}$$

i.e., $\widehat{\mathbf{s}}_{\max}$ is strongly consistent. This completes the first part of the proof.

Since the densities of \mathbf{X} and \mathbf{Y} are differentiable it follows that $\Psi(\mathbf{s})$ is continuous and twice differentiable on the manifold \mathcal{S}_+^{p-1} . In particular at $\mathbf{s}_{\max} \in \mathcal{S}_+^{p-1}$ the matrix $-\nabla^2 \Psi(\mathbf{s}_{\max})$ exists and is positive definite (we will compute its derivative in a moment). A

Taylor expansion implies that

$$\sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} \Psi(\mathbf{s}) - \Psi(\mathbf{s}_{\max}) \leq -C\delta^2 \quad (7.13)$$

holds. Furthermore by (7.12),

$$\begin{aligned} & \mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} N^{1/2} |(\Psi_{n,m}(\mathbf{s}) - \Psi(\mathbf{s})) - (\Psi_{n,m}(\mathbf{s}_{\max}) - \Psi(\mathbf{s}_{\max}))| \quad (7.14) \\ & \leq \sqrt{\frac{N}{n}} \mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} |(\mathbb{P}_n(\psi_1(\mathbf{s}) - \psi_1(\mathbf{s}_{\max})))| + \sqrt{\frac{N}{m}} \mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} |(\mathbb{P}_m(\psi_2(\mathbf{s}) - \psi_2(\mathbf{s}_{\max})))| \\ & \quad + \mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} \sqrt{N} |R_{n,m}(\mathbf{s})| + \mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} \sqrt{N} |R_{n,m}(\mathbf{s}_{\max})|. \end{aligned}$$

Here \mathbb{E}^* denotes the outer expectation, $\mathbb{P}_n \psi_1(\mathbf{s}) = n^{-1} \sum_{i=1}^n \psi_1(\mathbf{X}_i, \mathbf{s})$ and $\mathbb{P}_m \psi_2(\mathbf{s})$ is similarly defined. Recall that $\psi_1(\mathbf{s}) = \mathbb{P}(\mathbf{s}^T \mathbf{Y} \geq \mathbf{s}^T \mathbf{x}) - \Psi(\mathbf{s})$. It follows by Fatou's lemma that

$$\mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} |(\mathbb{P}_n(\psi_1(\mathbf{s}) - \psi_1(\mathbf{s}_{\max})))| \leq \mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} |(\mathbb{P}_n(\bar{\psi}(\mathbf{s}) - \bar{\psi}(\mathbf{s}_{\max})))| \quad (7.15)$$

where $\bar{\psi}(\mathbf{s}) = \mathbb{I}_{(\mathbf{s}^T \mathbf{Z} \geq 0)} - \Psi(\mathbf{s})$, $\mathbf{Z} =_{st} \mathbf{Y} - \mathbf{X}$ and the RHS of (7.15) is computed with respect to an IID sequence $\mathbf{Z}_1, \mathbf{Z}_2, \dots$. Equation (7.15) implies that a bound for (7.14) can be found using standard empirical process theory. We first note that the bracketing entropy of the upper half-planes is of the order δ/ε^2 . The envelope function of the class $\mathbb{I}_{(\mathbf{s}^T \mathbf{z} \geq 0)} - \mathbb{I}_{(\mathbf{s}_{\max}^T \mathbf{z} \geq 0)}$ where $\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta$ is bounded by $\mathbb{I}_{(\mathbf{s}^T \mathbf{z} \geq 0 > \mathbf{s}_{\max}^T \mathbf{z})} + \mathbb{I}_{(\mathbf{s}_{\max}^T \mathbf{z} \geq 0 > \mathbf{s}^T \mathbf{z})}$ whose squared L_2 norm is

$$\mathbb{P}(\mathbf{s}^T \mathbf{Z} \geq 0 > \mathbf{s}_{\max}^T \mathbf{Z}) + \mathbb{P}(\mathbf{s}_{\max}^T \mathbf{Z} \geq 0 > \mathbf{s}^T \mathbf{Z}). \quad (7.16)$$

Note that we may replace the RV \mathbf{Z} in (7.16) with the RV $\mathbf{Z}' = \mathbf{Z} / \|\mathbf{Z}\|$ whose mass is concentrated on the unit sphere. The condition that $\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta$ implies that the angle between \mathbf{s} and \mathbf{s}_{\max} is of the order $O(\delta)$ and therefore $\mathbb{P}(\mathbf{s}^T \mathbf{Z}' \geq 0 > \mathbf{s}_{\max}^T \mathbf{Z}')$ is computed as surface integral on a spherical wedge with maximum width δ . It follows that (7.16) is bounded by $2A_{p-1} \delta \|h'\|_\infty$ where A_{p-1} is the area of \mathcal{S}_+^{p-1} and $\|h'\|_\infty$ is the supremum of the density of \mathbf{Z}' . Clearly $\|h'\|_\infty < \infty$ since the densities of \mathbf{X} and \mathbf{Y} are bounded by assumption. Thus by Corollary 19.35 in van der Vaart (2000) we have

$$\sqrt{\frac{N}{n}} \mathbb{E}^* \sup_{\|\mathbf{s} - \mathbf{s}_{\max}\| < \delta} |(\mathbb{P}_n(\bar{\psi}(\mathbf{s}) - \bar{\psi}(\mathbf{s}_{\max})))| \leq C\delta^{1/2}$$

The same bound holds for the second term in (7.14). Recall that $\sqrt{N} |R_{n,m}(\mathbf{s})| \leq o(1)$ for all \mathbf{s} as $n, m \rightarrow \infty$ taking care of the third and fourth terms of (7.14). Hence (7.14) is bounded by $C\delta^{1/2}$. It now follows that

$$\Psi_{n,m}(\hat{\mathbf{s}}_{\max}) \geq \sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} \Psi_{n,m}(\mathbf{s}) - o_p(N^{-2/3}) \quad (7.17)$$

which implies by Theorem 5.52 in van der Vaart (2000) and Theorem 14.4 of Kosorok (2008)

that

$$\widehat{\mathbf{s}}_{\max} = \mathbf{s}_{\max} + O_p(N^{-1/3})$$

i.e., $\widehat{\mathbf{s}}_{\max}$ converges to \mathbf{s}_{\max} at a cube root rate. This completes the second part of the proof.

The limit distribution is derived by verifying the conditions in Theorem 1.1 of Kim and Pollard (1990), denoted henceforth by KP. It is important to note that the framework of KP may be applied to $\Psi_{n,m}$ using the decomposition (7.12). Hence standard empirical processes theory is used. Our proof closely follows example 6.4 in KP where local coordinates (3.6) were used. Although the model motivating that example is very different from ours the asymptotic analysis is similar. First note that (7.17) is condition (i) in KP. Since $\widehat{\mathbf{s}}_{\max}$ is consistent condition (ii) also holds and condition (iii) holds by assumption. Recall that $\Psi(\mathbf{s}) = \int \mathbb{I}_{(\mathbf{s}^T \mathbf{z} \geq 0)} h(\mathbf{z}) d\mathbf{z}$ where h is the density of \mathbf{Z} . The differentiability of the densities of \mathbf{X} and \mathbf{Y} implies that $\Psi(\mathbf{s})$ is twice differentiable. In fact a bit of differential geometry shows that

$$\nabla^2 \Psi(\mathbf{s}_{\max}) = \int \mathbb{I}_{(\mathbf{s}_{\max}^T \mathbf{z} = 0)} \mathbf{z} \mathbf{z}^T h(\mathbf{z}) d\sigma$$

where σ is the surface measure on the set $\{\mathbf{s}^T \mathbf{z} = 0\}$. This matrix is nonsingular on \mathcal{S}_+^{p-1} , see KP for more details, establishing condition (iv). Using (7.12) we see that condition (v) in KP is equivalent to the existence of the limit

$$\begin{aligned} C(\mathbf{u}, \mathbf{v}) &= \frac{1}{\lambda} \lim_{\alpha \rightarrow \infty} \alpha \mathbb{E} [\psi_1(\mathbf{X}, \mathbf{s}(\mathbf{u}/\alpha)) \psi_1(\mathbf{X}, \mathbf{s}(\mathbf{v}/\alpha))] \\ &\quad + \frac{1}{1-\lambda} \lim_{\alpha \rightarrow \infty} \alpha \mathbb{E} (\psi_2(\mathbf{Y}, \mathbf{s}(\mathbf{u}/\alpha)) \psi_2(\mathbf{Y}, \mathbf{s}(\mathbf{v}/\alpha))) \end{aligned}$$

where $\lambda = \lim((n+m)/n)$ and $\mathbf{s}(\mathbf{u}/\alpha)$ is given by the local coordinates (3.6). As in KP this limit exists provided

$$\lim_{\alpha \rightarrow \infty} \alpha \mathbb{E} |\psi_j(\cdot, \mathbf{s}(\mathbf{u}/\alpha)) - \psi_j(\cdot, \mathbf{s}(\mathbf{v}/\alpha))|^2$$

exists for $j = 1, 2$. By the modulus inequality the latter is bounded by

$$\lim_{\alpha \rightarrow \infty} \alpha \mathbb{E} \left| \mathbb{I}_{(\mathbf{s}(\mathbf{u}/\alpha)^T \mathbf{Z} \geq 0)} - \mathbb{I}_{(\mathbf{s}(\mathbf{v}/\alpha)^T \mathbf{Z} \geq 0)} \right|^2.$$

KP showed that the above limit exists, hence condition (v) holds. Conditions (vi) – (vii) were verified in the second part of the proof. Thus we may apply Theorem 1.1 in KP to get

$$N^{1/3} \widehat{\mathbf{t}}_N \Rightarrow \arg \max \{-Q(\mathbf{t}) + \mathbb{W}(\mathbf{t}) : \mathbf{t} \in \mathbf{s}_{\max}^\perp\}$$

where by KP $Q(\mathbf{t}) = \int \mathbb{I}(\mathbf{s}_{\max}^T \mathbf{z} = 0) (\mathbf{t}^T \mathbf{z})^2 h(\mathbf{z}) d\sigma$ and $\mathbb{W}(\mathbf{t})$ is a zero mean Gaussian process with covariance function $C(\mathbf{u}, \mathbf{v})$. This completes the proof. ■

Proof of Proposition 3.3:

Proof. Note that

$$\Psi(\mathbf{s}) = \mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}) = \mathbb{P}(\mathbf{s}^T \mathbf{Z} \geq 0) = \mathbb{P}\left(\frac{\mathbf{s}^T \mathbf{Z} - \mathbf{s}^T \boldsymbol{\delta}}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}} \geq -\frac{\mathbf{s}^T \boldsymbol{\delta}}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}}\right) = 1 - F\left(-\frac{\mathbf{s}^T \boldsymbol{\delta}}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}}\right).$$

Now, by assumption the DF F is independent of \mathbf{s} . Therefore $\Psi(\mathbf{s})$ is uniquely maximized

on \mathcal{S}_+^{p-1} if and only if the function

$$\varkappa(\mathbf{s}) = \frac{\mathbf{s}^T \boldsymbol{\delta}}{\sqrt{\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s}}}$$

is uniquely maximized on \mathcal{S}_+^{p-1} . If $\boldsymbol{\Sigma} = \mathbf{I}$ then $\varkappa(\mathbf{s}) = \mathbf{s}^T \boldsymbol{\delta}$ and we wish to maximize a linear function on \mathcal{S}_+^{p-1} . It is easily verified (by using ideas from linear programming) that the maximizer is unique if $\boldsymbol{\delta} \geq \mathbf{0}$ which is true by assumption. Incidentally, it is easy to show directly that $\varkappa(\mathbf{s})$ is maximized at $\mathbf{s}^*/\|\mathbf{s}^*\|$ where

$$\mathbf{s}^* = (\delta_1 \mathbb{I}_{(\delta_1 \geq 0)}, \dots, \delta_p \mathbb{I}_{(\delta_p \geq 0)}).$$

Now let $\boldsymbol{\Sigma} \neq \mathbf{I}$ and assume that a unique maximizer does not exist, i.e., suppose that $\varkappa(\mathbf{s})$ is maximized by both \mathbf{s}_1 and \mathbf{s}_2 . It is clear that $\varkappa(\lambda_1 \mathbf{s}_1) = \varkappa(\lambda_2 \mathbf{s}_2)$ for all $\lambda_1, \lambda_2 > 0$, i.e., the value of $\varkappa(\cdot)$ is constant along rays through the origin. The rays passing through \mathbf{s}_1 and \mathbf{s}_2 , respectively, intersect the ellipsoid $\mathbf{s}^T \boldsymbol{\Sigma} \mathbf{s} = 1$ at the points \mathbf{p}_1 and \mathbf{p}_2 . It follows that $\varkappa(\mathbf{p}_1) = \varkappa(\mathbf{p}_2)$, moreover \mathbf{p}_1 and \mathbf{p}_2 maximize $\varkappa(\cdot)$ on the ellipsoid. Now since $\mathbf{p}_1^T \boldsymbol{\Sigma} \mathbf{p}_1 = 1 = \mathbf{p}_2^T \boldsymbol{\Sigma} \mathbf{p}_2$ we must have $\mathbf{p}_1^T \boldsymbol{\delta} = \mathbf{p}_2^T \boldsymbol{\delta}$. Recall that a linear function on ellipsoid is uniquely maximized (just like on a sphere, see the comment above). Therefore we must have $\mathbf{p}_1 = \mathbf{p}_2$ which implies that $\mathbf{s}_1 = \mathbf{s}_2$ as required. ■

Proof of Theorem 3.2:

Proof. If $\mathbf{X} =_{st} \mathbf{Y}$ then for all \mathbf{s} we have $\mathbf{s}^T \mathbf{X} =_{st} \mathbf{s}^T \mathbf{Y}$. By assumption both $\mathbf{s}^T \mathbf{X}$ and $\mathbf{s}^T \mathbf{Y}$ are continuous RVs so $\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}) = 1/2$. Suppose now that both $\mathbf{X} \preceq_{st} \mathbf{Y}$ and $\mathbb{P}(\mathbf{s}^T \mathbf{X} \leq \mathbf{s}^T \mathbf{Y}) > 1/2$ for some $\mathbf{s} \in \mathcal{S}_+^{p-1}$, hold. Then we must have $\mathbf{X} \prec_{l-st} \mathbf{Y}$. Since $\mathbf{X} \preceq_{st} \mathbf{Y}$ we have $X_j \preceq_{st} Y_j$ for $1 \leq j \leq p$. One of these inequalities must be strict, otherwise $\mathbf{X} =_{st} \mathbf{Y}$ contradicting the fact that $\mathbf{X} \prec_{l-st} \mathbf{Y}$. Now use Theorem 1 in Davidov and Peddada (2011) to complete the proof. ■

Proof of Theorem 3.3:

Proof. The functions ψ_1 and ψ_2 defined in the proof of Theorem 3.1 are Donsker (cf. example 19.7 in van der Vaart 2000). Hence by the theory of empirical processes applied to (7.12) we find that

$$N^{1/2}(\Psi_{n,m}(\mathbf{s}) - \Psi(\mathbf{s})) \Rightarrow \mathbb{G}(\mathbf{s}) \quad (7.18)$$

where $\mathbb{G}(\mathbf{s})$ is a zero mean Gaussian process and convergence holds for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$. We also note that (7.18) is a two sample U-Processes. A central limit theorem for such processes is described by Neumeyer (2004). Hence by the continuous mapping theorem, and under H_0 , we have $N^{1/2}(\Psi_{n,m}(\widehat{\mathbf{s}}_{\max}) - 1/2) \Rightarrow \sup_{\mathbf{s} \in \mathcal{S}_+^{p-1}} \mathbb{G}(\mathbf{s})$ where the covariance function of $\mathbb{G}(\mathbf{s})$, denoted by $C(\mathbf{u}, \mathbf{v})$, is given by

$$\frac{1}{\lambda} \mathbb{P}(\mathbf{u}^T \mathbf{X}_1 \leq \mathbf{u}^T \mathbf{X}_2, \mathbf{v}^T \mathbf{X}_1 \leq \mathbf{v}^T \mathbf{X}_3) + \frac{1}{1-\lambda} \mathbb{P}(\mathbf{u}^T \mathbf{X}_1 \leq \mathbf{u}^T \mathbf{X}_2, \mathbf{v}^T \mathbf{X}_3 \leq \mathbf{v}^T \mathbf{X}_2) - \frac{1}{4\lambda(1-\lambda)} \quad (7.19)$$

where $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are IID from the common DF. ■

Proof of Theorem 3.4:

Proof. Suppose that $\mathbf{X} \prec_{l-st} \mathbf{Y}$. Then for some $\mathbf{s}_* \in \mathcal{S}_+^{p-1}$ we have $\mathbf{s}_*^T \mathbf{X} \prec_{st} \mathbf{s}_*^T \mathbf{Y}$ which implies that $\mathbb{P}(\mathbf{s}_*^T \mathbf{X} \leq \mathbf{s}_*^T \mathbf{Y}) > 1/2$. By definition $\mathbb{P}(\mathbf{s}_{\max}^T \mathbf{X} \leq \mathbf{s}_{\max}^T \mathbf{Y}) \geq \mathbb{P}(\mathbf{s}_*^T \mathbf{X} \leq \mathbf{s}_*^T \mathbf{Y})$

so $\Psi(\mathbf{s}_{\max}) > 1/2$. It follows from the proof of Theorem 3.1 that $\Psi_{n,m}(\widehat{\mathbf{s}}_{\max}) \rightarrow \Psi(\mathbf{s}_{\max})$ with probability one. Thus,

$$S_{n,m} = N^{1/2} (\Psi_{n,m}(\widehat{\mathbf{s}}_{\max}) - 1/2) \xrightarrow{a.s.} \infty \text{ as } n, m \rightarrow \infty.$$

Therefore by Slutsky's theorem

$$\mathbb{P}(S_{n,m} > q_{n,m,1-\alpha}; H_1) \rightarrow 1 \text{ as } n, m \rightarrow \infty,$$

where $q_{n,m,1-\alpha}$ is the critical value for an α level test based on samples of size n and m and $q_{n,m,1-\alpha} \rightarrow q_{1-\alpha}$. Hence the test based on $S_{n,m}$ is consistent. Consistency for $I_{n,m}$ and $I_{n,m}^+$ is established in a similar manner.

Now assume that $\mathbf{X} \preceq_{l-st} \mathbf{Y} \preceq_{l-st} \mathbf{Z}$ so that $\mathbf{s}^T \mathbf{Y} \preceq_{st} \mathbf{s}^T \mathbf{Z}$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$. Fix \mathbf{x}_i , $i = 1, \dots, n$ and choose $\mathbf{s} \in \mathcal{S}_+^{p-1}$. Without any loss of generality assume that $\mathbf{s}^T \mathbf{x}_1 \leq \mathbf{s}^T \mathbf{x}_2 \leq \dots \leq \mathbf{s}^T \mathbf{x}_n$. Define $U_j = \sum_{i=1}^n \mathbb{I}_{(\mathbf{s}^T \mathbf{x}_i \leq \mathbf{s}^T \mathbf{Y}_j)}$ and $V_j = \sum_{i=1}^n \mathbb{I}_{(\mathbf{s}^T \mathbf{x}_i \leq \mathbf{s}^T \mathbf{Z}_j)}$. Clearly U_j and V_j take values in $J = \{0, \dots, n\}$. Now, for $k \in J$ we have

$$\mathbb{P}(U_j \geq k) = \mathbb{P}(\mathbf{s}^T \mathbf{Y}_j \geq \mathbf{s}^T \mathbf{x}_k) \leq \mathbb{P}(\mathbf{s}^T \mathbf{Z}_j \geq \mathbf{s}^T \mathbf{x}_k) = \mathbb{P}(V_j \geq k)$$

where we use the fact that $\mathbf{s}^T \mathbf{Y} \preceq_{st} \mathbf{s}^T \mathbf{Z}$. It follows that $U_j \preceq_{st} V_j$ for $j = 1, \dots, m$. Moreover $\{U_j\}$ and $\{V_j\}$ are all independent and it follows from Theorem 1.A.3 in Shaked and Shanthikumar (2007) that $\sum_{j=1}^m U_j \preceq_{st} \sum_{j=1}^m V_j$. Thus $\sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{(\mathbf{s}^T \mathbf{x}_i \leq \mathbf{s}^T \mathbf{Y}_j)} \preceq_{st} \sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{(\mathbf{s}^T \mathbf{x}_i \leq \mathbf{s}^T \mathbf{Z}_j)}$. The latter holds for every value of $\mathbf{x}_1, \dots, \mathbf{x}_n$ and therefore it holds unconditionally as well, i.e.,

$$\sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{(\mathbf{s}^T \mathbf{X}_i \leq \mathbf{s}^T \mathbf{Y}_j)} \preceq_{st} \sum_{i=1}^n \sum_{j=1}^m \mathbb{I}_{(\mathbf{s}^T \mathbf{X}_i \leq \mathbf{s}^T \mathbf{Z}_j)}.$$

It follows that $\Psi_{n,m}^{\mathbf{X},\mathbf{Y}}(\mathbf{s}) \preceq_{st} \Psi_{n,m}^{\mathbf{X},\mathbf{Z}}(\mathbf{s})$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$ where $\Psi_{n,m}^{\mathbf{X},\mathbf{Y}}(\mathbf{s})$ and $\Psi_{n,m}^{\mathbf{X},\mathbf{Z}}(\mathbf{s})$ are defined in (3.1) and the superscripts emphasize the different arguments used to evaluate them. Thus,

$$\Psi_{n,m}^{\mathbf{X},\mathbf{Y}}(\widehat{\mathbf{s}}_{\max}^{\mathbf{X},\mathbf{Y}}) \preceq_{st} \Psi_{n,m}^{\mathbf{X},\mathbf{Z}}(\widehat{\mathbf{s}}_{\max}^{\mathbf{X},\mathbf{Y}}) \preceq_{st} \Psi_{n,m}^{\mathbf{X},\mathbf{Z}}(\widehat{\mathbf{s}}_{\max}^{\mathbf{X},\mathbf{Z}})$$

and as a consequence $\mathbb{P}(S_{n,m}^{\mathbf{X},\mathbf{Y}} > q_{n,m,1-\alpha}) \leq \mathbb{P}(S_{n,m}^{\mathbf{X},\mathbf{Z}} > q_{n,m,1-\alpha})$ as required. The monotonicity of the power function of $I_{n,m}$ and $I_{n,m}^+$ follows immediately from the fact that $\Psi_{n,m}^{\mathbf{X},\mathbf{Y}}(\mathbf{s}) \preceq_{st} \Psi_{n,m}^{\mathbf{X},\mathbf{Z}}(\mathbf{s})$ for all $\mathbf{s} \in \mathcal{S}_+^{p-1}$. ■

$n = m = 20$			
δ	ρ	Bias	MSE
(1, 1, 1)	-0.25	0.001	0.072
(1, 1, 1)	0	0.004	0.129
(1, 1, 1)	0.25	0.009	0.187
(1, 1, 1)	0.50	0.012	0.216
(1, 1, 1)	0.90	0.010	0.203
(3, 2, 1)	-0.25	0.018	0.090
(3, 2, 1)	0	0.001	0.066
(3, 2, 1)	0.25	0.053	0.114
(3, 2, 1)	0.50	0.060	0.113
(3, 2, 1)	0.90	0.112	0.170
$n = m = 100$			
δ	ρ	Bias	MSE
(1, 1, 1)	-0.25	0.00009	0.014
(1, 1, 1)	0	0.00021	0.027
(1, 1, 1)	0.25	0.00045	0.041
(1, 1, 1)	0.50	0.00079	0.056
(1, 1, 1)	0.90	0.00050	0.044
(3, 2, 1)	-0.25	0.02400	0.039
(3, 2, 1)	0	0.00004	0.012
(3, 2, 1)	0.25	0.05200	0.065
(3, 2, 1)	0.50	0.06400	0.077
(3, 2, 1)	0.90	0.14100	0.158

Table 1: Bias and MSE of $\widehat{\mathfrak{s}}_{\max}$.

Set Up		Coverage for various values of (N, M)				
Pattern	ρ	(20, 10)	(20, 20)	(40, 10)	(40, 20)	(40, 40)
1	-0.25	0.945	0.981	0.624	0.875	0.971
1	0	0.859	0.913	0.564	0.800	0.918
1	0.25	0.931	0.916	0.754	0.888	0.933
1	0.50	0.992	0.971	0.982	0.981	0.969
1	0.90	1.000	0.993	1.000	0.997	0.989
2	-0.25	0.921	0.982	0.702	0.880	0.967
2	0	0.945	0.984	0.798	0.905	0.972
2	0.25	0.953	0.986	0.852	0.943	0.978
2	0.50	0.947	0.968	0.869	0.937	0.968
2	0.90	0.932	0.950	0.818	0.929	0.954

Table 2: Coverage probabilities for the M out of N bootstrap for $p = 5$ normal data. Pattern $i = 1, 2$ corresponds to $\delta_1 = (.1, .25, .5, .75, .9)$ and $\delta_2 = (.5, .5, .5, .5, .5)$.

Set Up			Type I error	
Distribution	p	ρ	$n = m = 15$	$n = m = 25$
MVNs	3	-0.25	0.041	0.037
MVNs	3	0.00	0.023	0.044
MVNs	3	0.25	0.037	0.033
MVNs	3	0.50	0.027	0.032
MVNs	3	0.90	0.031	0.036
MVNs	5	-0.25	0.035	0.035
MVNs	5	0.00	0.040	0.041
MVNs	5	0.25	0.045	0.032
MVNs	5	0.50	0.038	0.043
MVNs	5	0.90	0.044	0.031
MV-LogN	3	-0.25	0.025	0.040
MV-LogN	3	0.00	0.038	0.049
MV-LogN	3	0.25	0.025	0.027
MV-LogN	3	0.50	0.028	0.037
MV-LogN	3	0.90	0.026	0.034
MV-LogN	5	-0.25	0.026	0.039
MV-LogN	5	0.00	0.035	0.018
MV-LogN	5	0.25	0.039	0.039
MV-LogN	5	0.50	0.036	0.046
MV-LogN	5	0.90	0.034	0.042
Mix-MVNs	3	-0.25	0.032	0.040
Mix-MVNs	3	0.00	0.038	0.028
Mix-MVNs	3	0.25	0.039	0.032
Mix-MVNs	3	0.50	0.036	0.035
Mix-MVNs	3	0.90	0.041	0.028
Mix-MVNs	5	-0.25	0.042	0.035
Mix-MVNs	5	0.00	0.040	0.031
Mix-MVNs	5	-0.25	0.041	0.028
Mix-MVNs	5	0.50	0.034	0.040
Mix-MVNs	5	0.90	0.042	0.036

Table 3: Type I errors for the proposed procedure with nominal level $\alpha = .05$. Three types of distributions are considered: MVNs, MV-LogN (multivariate lognormal) and Mix-MVN (mixtures of MVNs).

Type I error and power $p = 10, n = m = 10$			Type I error and power $p = 20, n = m = 10$		
δ	ρ	Type I error	δ	ρ	Type I error
$\mathbf{0}$	0.00	0.054	$\mathbf{0}$	0.00	0.081
$\mathbf{0}$	0.25	0.051	$\mathbf{0}$	0.25	0.050
$\mathbf{0}$	0.50	0.028	$\mathbf{0}$	0.50	0.046
$\mathbf{0}$	0.90	0.038	$\mathbf{0}$	0.90	0.048
Power			Power		
δ_1	0.00	0.83	δ_1	0.00	0.97
δ_1	0.25	0.48	δ_1	0.25	0.53
δ_1	0.50	0.26	δ_1	0.50	0.42
δ_1	0.90	0.20	δ_1	0.90	0.22
δ_2	0.00	0.98	δ_2	0.00	0.98
δ_2	0.25	0.80	δ_2	0.25	0.59
δ_2	0.50	0.67	δ_2	0.50	0.43
δ_2	0.90	0.71	δ_2	0.90	0.40

Table 4: Type I errors and power for some settings with $p \geq n$. Here $n = m = 10$ and δ_1 has components $1/2$ and δ_2 has components i/p

Power & RE % ($n = m = 15$)							
Set Up			Directional Tests			Integral Tests	
p	δ	ρ	$S_{n,m}$	RMD test	TMD test	$I_{n,m}$	$I_{n,m}^+$
3	δ_1	-0.25	0.79 (90%)	0.62 (71%)	0.88	0.89 (100%)	0.89 (100%)
3	δ_1	0.00	0.64 (82%)	0.45 (57%)	0.78	0.68 (87%)	0.68 (87%)
3	δ_1	0.25	0.53 (78%)	0.38 (56%)	0.68	0.54 (79%)	0.54 (79%)
3	δ_1	0.50	0.51 (73%)	0.41 (59%)	0.70	0.47 (67%)	0.47 (67%)
3	δ_1	0.90	0.62 (64%)	0.85 (99%)	0.97	0.40 (41%)	0.41 (42%)
5	δ_2	-0.25	0.93 (95%)	0.74 (76%)	0.98	0.97 (99%)	0.97 (99%)
5	δ_2	0.00	0.80 (87%)	0.56 (60%)	0.92	0.86 (93%)	0.86 (93%)
5	δ_2	0.25	0.59 (73%)	0.39 (47%)	0.81	0.66 (81%)	0.66 (81%)
5	δ_2	0.50	0.56 (67%)	0.42 (50%)	0.84	0.48 (57%)	0.48 (57%)
5	δ_2	0.90	0.63 (64%)	0.88 (89%)	0.99	0.40 (40%)	0.40 (40%)
3	δ_3	-0.25	0.74 (89%)	0.54 (64%)	0.83	0.83 (100%)	0.83 (100%)
3	δ_3	0.00	0.56 (87%)	0.34 (53%)	0.64	0.59 (92%)	0.59 (92%)
3	δ_3	0.25	0.42 (87%)	0.23 (48%)	0.49	0.46 (93%)	0.46 (93%)
3	δ_3	0.50	0.33 (86%)	0.15 (40%)	0.38	0.37 (97%)	0.37 (97%)
3	δ_3	0.90	0.27 (83%)	0.12 (38%)	0.32	0.27 (83%)	0.27 (83%)
5	δ_4	-0.25	0.92 (95%)	0.65 (68%)	0.96	0.95 (99%)	0.95 (99%)
5	δ_4	0.00	0.75 (90%)	0.43 (51%)	0.83	0.82 (99%)	0.82 (99%)
5	δ_4	0.25	0.49 (87%)	0.20 (35%)	0.57	0.60 (100%)	0.60 (100%)
5	δ_4	0.50	0.41 (90%)	0.16 (34%)	0.45	0.43 (100%)	0.43 (100%)
5	δ_4	0.90	0.29 (92%)	0.10 (32%)	0.31	0.33 (100%)	0.33 (100%)

Table 5a: Power comparisons of the two proposed test procedures with type I error of .050. Here $\delta_1 = (0.1, 0.5, .9)$, $\delta_2 = (0.1, 0.25, 0.5, 0.75, 0.9)$, $\delta_3 = (0.5, 0.5, 0.5)$ and $\delta_4 = (0.5, 0.5, 0.5, 0.5, 0.5)$ and $n = m = 15$.

Power & RE ($n = m = 25$)							
Set Up			Directional Tests			Integral Tests	
p	δ	ρ	$S_{n,m}$	RMD test	TMD test	$I_{n,m}$	$I_{n,m}^+$
3	δ_1	-0.25	0.96 (98%)	0.90 (91%)	0.98	0.98 (100%)	0.98 (100%)
3	δ_1	0.00	0.85 (92%)	0.72 (78%)	0.92	0.85 (92%)	0.86 (92%)
3	δ_1	0.25	0.80 (88%)	0.69 (76%)	0.90	0.75 (83%)	0.75 (83%)
3	δ_1	0.50	0.75 (84%)	0.67 (75%)	0.89	0.66 (74%)	0.66 (74%)
3	δ_1	0.90	0.89 (89%)	0.98 (99%)	1.00	0.59 (59%)	0.61 (61%)
5	δ_2	-0.25	1.00 (100%)	0.98 (98%)	1.00	1.00 (100%)	1.00 (100%)
5	δ_2	0.00	0.96 (97%)	0.85 (86%)	0.99	0.98 (99%)	0.98 (99%)
5	δ_2	0.25	0.85 (88%)	0.74 (76%)	0.97	0.83 (86%)	0.83 (86%)
5	δ_2	0.50	0.81 (84%)	0.74 (77%)	0.96	0.70 (73%)	0.70 (73%)
5	δ_2	0.90	0.90 (90%)	.99 (100%)	1.00	0.57 (57%)	0.58 (58%)
3	δ_3	-0.25	0.94 (96%)	0.85 (87%)	0.98	0.96 (98%)	0.96 (98%)
3	δ_3	0.00	0.75 (92%)	0.57 (69%)	0.82	0.79 (96%)	0.79 (96%)
3	δ_3	0.25	0.62 (89%)	0.39 (56%)	0.70	0.66 (94%)	0.66 (94%)
3	δ_3	0.50	0.54 (90%)	0.31 (52%)	0.60	0.55 (92%)	0.55 (92%)
3	δ_3	0.90	0.44 (90%)	0.20 (42%)	0.49	0.42 (86%)	0.42 (86%)
5	δ_4	-0.25	0.99 (99%)	0.94 (94%)	1.00	1.00 (100%)	1.00 (100%)
5	δ_4	0.00	0.94 (96%)	0.72 (74%)	0.97	0.97 (100%)	0.97 (100%)
5	δ_4	0.25	0.71 (90%)	0.41 (52%)	0.79	0.79 (100%)	0.79 (100%)
5	δ_4	0.50	0.58 (91%)	0.25 (39%)	0.63	0.64 (100%)	0.64 (100%)
5	δ_4	0.90	0.42 (87%)	0.18 (36%)	0.49	0.46 (94%)	0.46 (94%)

Table 5b: Power comparisons of the two proposed test procedures with type I error of .050. Here $\delta_1 = (0.1, 0.5, .9)$, $\delta_2 = (0.1, 0.25, 0.5, 0.75, 0.9)$, $\delta_3 = (0.5, 0.5, 0.5)$ and $\delta_4 = (0.5, 0.5, 0.5, 0.5, 0.5)$ and $n = m = 25$.