DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 1/98

Checking for orthant orderings between discrete multivariate distributions: An algorithm

by

Rainer Dyckerhoff

Hartmut Holz

Karl Mosler

DISKUSSIONSBEITRÄGE ZUR STATISTIK UND ÖKONOMETRIE

SEMINAR FÜR WIRTSCHAFTS- UND SOZIALSTATISTIK
UNIVERSITÄT ZU KÖLN

Albertus-Magnus-Platz, D-50923 Köln, Deutschland

DISCUSSION PAPERS IN STATISTICS AND ECONOMETRICS

SEMINAR OF ECONOMIC AND SOCIAL STATISTICS
UNIVERSITY OF COLOGNE

No. 1/98

Checking for orthant orderings between discrete multivariate distributions: An algorithm

by

Rainer Dyckerhoff[1]
Hartmut Holz[2]
Karl Mosler[3]

Abstract

We consider four orthant stochastic orderings between random vectors X and Y that have finitely discrete probability distributions in $\mathbb{R}^k$. For each of the orderings conditions have been developed that are necessary and sufficient for dominance of Y over X. We present an algorithm that checks these conditions in an efficient way by operating on a semilattice generated by the support of the two distributions. In particular, the algorithm can be used to compute multivariate Smirnov statistics.

Keywords: Multivariate stochastic orders, decision under risk, comparison of empirical distribution functions.

AMS Subject Classification: Primary 60E15, Secondary 90A43.

[1]Seminar für Wirtschafts– und Sozialstatistik, Universität zu Köln, D-50923 Köln

[2]Institut für Statistik und Quantitative Ökonomik, Universität der Bundeswehr Hamburg, D-22039 Hamburg

[3]Seminar für Wirtschafts– und Sozialstatistik, Universität zu Köln, D-50923 Köln

1 Description and purpose of the algorithm

For random vectors X, Y in $\mathbb{R}^k$, the four orthant orderings $\preceq_{lo}$, $\preceq_{uo}$, $\preceq_{locc}$ and $\preceq_{uocx}$ are defined in the following way:

$$
\begin{array}{lclclcll}
X & \preceq_{lo} & Y & \Longleftrightarrow & P[X \le a] & \ge & P[Y \le a] & \text{for all } a \in \mathbb{R}^k, \\
X & \preceq_{uo} & Y & \Longleftrightarrow & P[X \ge a] & \le & P[Y \ge a] & \text{for all } a \in \mathbb{R}^k, \\
X & \preceq_{locc} & Y & \Longleftrightarrow & \int\limits_{]-\infty,a]} P[X \le z]dz & \ge & \int\limits_{]-\infty,a]} P[Y \le z]dz & \text{for all } a \in \mathbb{R}^k, \\
X & \preceq_{uocx} & Y & \Longleftrightarrow & \int\limits_{[a,\infty[} P[X \ge z]dz & \le & \int\limits_{[a,\infty[} P[Y \ge z]dz & \text{for all } a \in \mathbb{R}^k.
\end{array}
$$

The orders are called *lower orthant order*, *upper orthant order*, *lower orthant concave order* and *upper orthant convex order*, respectively.

We provide an algorithm to check whether one of these four stochastic orderings holds, when X and Y have finitely discrete probability distributions.

This algorithm can also be used to compute four multivariate Smirnov statistics. Let

$$
\begin{aligned}
D_1^+(X,Y) &= \max_{x\in\mathbb{R}^k}(F(x) - G(x)), && (1)\\
D_1^-(X,Y) &= \min_{x\in\mathbb{R}^k}(F(x) - G(x)), && (2)\\
D_2^+(X,Y) &= \max_{x\in\mathbb{R}^k}(\overline{G}(x) - \overline{F}(x)), && (3)\\
D_2^-(X,Y) &= \min_{x\in\mathbb{R}^k}(\overline{G}(x) - \overline{F}(x)), && (4)
\end{aligned}
$$

where

$$F(x) = P[X \le x], \quad \overline{F}(x) = P[X \ge x], \quad G(x) = P[Y \le x], \quad \overline{G}(x) = P[Y \ge x].$$

When $k = 1$, both $\preceq_{lo}$ and $\preceq_{uo}$ become the usual stochastic order (= first degree stochastic dominance), while $\preceq_{locc}$ and $\preceq_{uocx}$ become the univariate concave, respectively convex, order (= second degree stochastic dominance). Then our problem reduces to checking two finite discrete distributions for first and second degree stochastic dominance. A number of computational approaches has been proposed in the literature to solve this problem when $k = 1$. See Porter et al. (1973), Markowitz (1977), Bawa et al. (1979), Levy and Sarnat (1984), Levy (1992), Aboudi and Thon (1994).

For general $k \ge 1$, Dyckerhoff and Mosler (1997) present the background, theory and applications of the four orthant stochastic orderings and the theoretical justification of the present algorithm. Let S denote the joint support of the distributions of X and Y, and $J(S)$ the join-semilattice (Birkhoff 1940) generated by S. For each $s \in S$ let $\delta_s = P[X = s] - P[Y = s]$. We introduce Δ^{lo} by

$$\Delta^{lo}(z) = \sum_{s \in S, s \leq z} \delta_s \ , \quad z \in \mathbb{R}^k. \tag{5}$$

For a nonempty $I \subset \{1, 2, \dots, k\}$ and $x \in \mathbb{R}^k$, let $x_I = (x_i)_{i \in I} \in \mathbb{R}^I$. Further, let X_I and Y_I be the marginals with respect to I, and S_I their joint support. Δ_I^{locc} is defined as follows:

$$\Delta_I^{locc}(z) = \sum_{s \in S, s_I \leq z} \delta_s \cdot \prod_{i \in I} (z_i - s_i) \ , \quad z \in \mathbb{R}^I. \tag{6}$$

Dyckerhoff and Mosler (1997) prove the following results.

Result 1. $X \preceq_{lo} Y$ if and only if

$$\Delta^{lo}(z) \geq 0 \quad \text{for all } z \in J(S). \tag{7}$$

Result 2. $X \preceq_{locc} Y$ if and only if

$$\Delta_I^{locc}(z) \geq 0 \quad \text{for all } z \in J(S_I) \text{ and all nonempty subsets } I \text{ of } \{1, \dots, k\}. \tag{8}$$

Result 3. $X \preceq_{lo} Y \Rightarrow X \preceq_{locc} Y \Rightarrow$

$$E[X] \leq E[Y]. \tag{9}$$

Result 4.

$$\begin{aligned} X \leq_{uo} Y \quad &\Longleftrightarrow \quad -Y \leq_{lo} -X, \\ X \leq_{uocx} Y \quad &\Longleftrightarrow \quad -Y \leq_{locc} -X. \end{aligned}$$

The main subroutine in our algorithm is the procedure *CheckJoinSemilattice*. This procedure proceeds as follows:

Step 1 For a given set I of components the joint support S_I of the two distributions w.r.t. I is constructed. Further, the δ_s, $s \in S_I$, are computed.

Step 2 S_I is put into lexicographical order.

Step 3 In a recursive way, all joins z of at most k points of S which are not comparable in the usual componentwise ordering of $\mathbb{R}^k$ are determined. This generates all points of the join-semilattice $J(S_I)$. As soon as a point z is generated, one of the following steps is done.

- **Step 3a** If the algorithm was called to check for $\preceq_{lo}$, inequality (7) is checked. Once a violation is detected the calculations are stopped.
- **Step 3b** If the algorithm was called to check for $\preceq_{locc}$, inequality (8) is checked. Once a violation is detected the calculations are stopped.

Step 3c If the algorithm was called to compute the Smirnov statistics $D_1^+(X,Y)$ and $D_1^-(X,Y)$ the maximum and minimum of $\Delta^{lo}(z)$ over all z in the join-semilattice is computed.

To decide whether the *lower orthant order* holds between X and Y, our algorithm proceeds as follows: First, it is checked whether the necessary condition (9) holds. If not, the algorithm stops. If the inequality (9) holds, the procedure *CheckJoinSemilattice* is called with $I = \{1, \ldots, k\}$.

To compute the Smirnov statistics $D_1^+(X,Y)$ and $D_1^-(X,Y)$ *CheckJoinSemilattice* is again called with $I = \{1, \ldots, k\}$. Since $\Delta^{lo}(z) = F(z) - G(z)$, this procedure yields the desired result.

To check for the *lower orthant concave order* between X and Y, more extensive calculations have to be performed. Again, first we check the inequality (9) which is a necessary condition for the lower orthant concave order, too. Then the procedure *CheckAllMargins* is called. It constructs all subsets I of $\{1, \ldots, k\}$. This is again done by a recursive procedure. Whenever such a subset is constructed *CheckJoinSemilattice* is called with the I that was just constructed.

Upper orthant order and *upper orthant convex order* between X and Y are checked using Result 4, i.e., by applying the previous procedures to the transformed random Vectors $-Y$ and $-X$. The same holds for the Smirnov statistics $D_2^+(X,Y)$ and $D_2^-(X,Y)$. The transformation of the random vectors is done in Step 1 of *CheckJoinSemilattice*.

The algorithm has been used (Holz and Mosler 1994) to determine a nondominated (with respect to one of the orderings) set of distributions from a given finite set of distributions. This is a standard problem in multiattribute decision making under risk. The algorithm has also been employed to construct statistical tests on $F \succeq_{lo} G$ and $F \succeq_{uo} G$ which are based on resampling.

2 Structure

Language

ISO-Pascal (Level 0)

Procedures

PROCEDURE CheckForOrthantOrdering(PX, PY: PDistribution; Dim: Integer; VAR Info: TInfo; VAR IFault: Integer);

Global constant

MaxDim is the maximum dimension.

Global types

```
TPoint        = ARRAY[1..MaxDim] OF Real;

PDistribution = ^TDistribution;
TDistribution = RECORD
                  Point : TPoint;
                  Prob : Real;
                  Link : PDistribution;
                END;
```

Remark. Distributions are represented by simply linked lists of records of the type *TDistribution*. Every such record contains a point of the support, its probability, and a pointer at the next support point. If there is no further point, the value of the *Link*-field is *nil*. A distribution is identified with a variable of the type *PDistribution* which points at the first support point of the distribution.

```
TCheck = (lo,locc,uo,uocx,Smirnov1,Smirnov2);

TInfo  = RECORD
           CASE Check:TCheck OF
             lo,uo,locc,uocx: ( Dominance: Boolean; Index: -1..1);
             Smirnov1,Smirnov2: ( DeltaMin, DeltaMax: Real );
           END;
```

Remark. A *TInfo*-record is used as an input-output parameter. The tag field *Check* specifies which check shall be done or which statistics shall be computed. It serves as an input parameter to the algorithm. Depending on the value of the tag field the algorithm returns the result in the following way.

If *Check* is *lo, uo, locc* or *uocx*, then *Dominance* is *true* if X dominates Y in the respective order or vice versa. In this case

$Index = 1$	if X dominates Y,
$Index = -1$	if Y dominates X,
$Index = 0$	if X and Y are equivalent.

If neither X dominates Y nor vice versa, then *Dominance* is *false* and *Index* is undefined.

If *Check* is *Smirnov1*, then *DeltaMax* is the value of $D_1^+(F,G)$ and *DeltaMin* is the value of $D_1^-(F,G)$.

If *Check* is *Smirnov2*, then *DeltaMax* is the value of $D_2^+(F,G)$ and *DeltaMin* is the value of $D_2^-(F,G)$.

Formal parameters

PX	PDistribution	value:	the distribution of X
PY	PDistribution	value:	the distribution of Y
Dim	Integer	value:	the dimension k
Info	TInfo	input/output:	the tag field *Check* specifies the operation to be made, the result is returned in the variant part of the record, see above
IFault	Integer	output:	the error indicator

Local constants

MaxSupport	is the maximum size of the joint support of *PX* and *PY*.
Eps	is the precision. Numbers whose absolute values are smaller than or equal to *Eps* will be treated as zero. This ensures that small rounding errors do not lead to an erroneous violation of (7) or (8). Thus, *Eps* should be set to at least the accuracy of the data.

Failure indications

IFault = *0*	no error occurred.
IFault = *1*	the constraint $1 \leq Dim \leq MaxDim$ is not satisfied.
IFault = *2*	the joint support of *PX* and *PY* exceeds *MaxSupport*.
IFault = *3*	*PX* or *PY* is not a probability distribution.

3 Accuracy and time

The accuracy of the results depends on the compiler. All constants are set in a declaration part and can be adapted to the machine and the compiler used. Since in calculating the Smirnov statistics only additions are involved, the accuracy of the Smirnov statistics is the same as the accuracy of the data.

Because the algorithm is very efficient it has been used in statistical resampling procedures and in building nondominated sets of distributions.

The CPU-times depend strongly on the given data. Consider two k-variate probability distributions with n points in the joint support. In the case of $\preceq_{lo}$ and $\preceq_{uo}$ to prove (7), at most $\sum_{i=1}^{k} \binom{n}{i}$ points have to be checked. This stems from the fact that every point in the join-semilattice $J(S)$ is the join of at most k points in S. To prove lower orthant concave order (or upper orthant convex order), all marginal distributions are examined. Thus in the worst case $\sum_{i=1}^{k} \sum_{j=1}^{i} \binom{k}{i} \binom{n}{j}$ points are constructed and checked.

In practical applications the parameters should stay within certain limits. For $k = 1$ a number of 100000 points in the joint support seems to be feasible, whereas for $k = 5$ the number of points in the joint support should not exceed 100.

It should be emphasized that the above bounds are no equalities. In general, the join-semilattice has cardinality much smaller than $\sum_{i=1}^{k} \binom{n}{i}$. Apart from n and k, the size of the semilattice depends strongly on how the points of S are dispersed in k-space. Further, if the algorithm just checks for one of the orthant orders, it stops as soon as (7) or (8) is violated at some $z \in J(S)$. This reduces the computation time when the distributions are not ordered and no Smirnov statistic is calculated.

Table 1 summarizes some computation times. Two samples of size $n/2$ were drawn from a k-variate normal distribution having expectations $\mu_1 = \mu_2 = \ldots = \mu_k = 0$ and covariances $\sigma_{ij} = \rho^{|i-j|}$, $i, j = 1, \ldots, k$. Then the one-sided Smirnov statistics (1) and (2) were computed. For every triple (n, k, ρ) this procedure was carried out ten times. The table shows the average computation times in seconds on a 60 MHz Pentium™.

ρ	n	k						
		2	3	4	5	6	7	8
	20	0.00	0.03	0.20	0.93	2.68	9.24	22.63
0.6	40	0.02	0.36	4.90	55.32			
	100	0.23	13.88	482.70				
	20	0.01	0.07	0.42	1.94	6.08	16.94	36.57
0.0	40	0.04	0.91	12.33	123.91	821.99	5336.62	
	100	0.53	34.38	1167.45				
	20	0.01	0.09	0.53	2.31	7.01	18.43	37.82
–0.6	40	0.07	1.38	16.66	150.94			
	100	0.93	51.27	1571.41				

Table 1. Computation times [in seconds] of the algorithm.

As can be seen from Table 1 the computation times depend not only on n and k. The greater the correlation of the distributions, as measured by ρ, the faster is the algorithm.

4 Additional comments

Although the algorithm works for every $k \geq 1$, for $k = 1$ a special approach is advisable. In the unidimensional case we have $J(S) = S$, since the support S is linearly ordered. Thus, the differences Δ^{lo} and Δ_I^{locc} can be easily calculated in a recursive way, which is simpler and more efficient than the above algorithm.

The algorithm as presented here is capable of checking for four different orders and of computing four different statistics. If one is interested only in some of these issues, its structure can be simplified in an obvious way.

However, it should be noted, that these modifications will only simplify the structure. They will not result in a significant reduction of computation times compared to our algorithm.

5 Availability

The algorithm is available by request from the authors or can be downloaded from our website `http://www.uni-koeln.de/wiso-fak/wisostatsem/algorithms`.

References

Aboudi, R., and Thon, D. (1994) Efficient algorithms for stochastic dominance tests based on financial market data. *Management Science* **40**, 508–515.

Bawa, V., Lindenberg, E., and Rafsky, L. (1979) An algorithm to determine stochastic dominance admissible sets. *Management Science* **25**, 609–622.

Birkhoff, G. (1940) *Lattice Theory*. Providence, RI: American Mathematical Society.

Dyckerhoff, R., and Mosler, K. (1997) Orthant orderings of discrete random vectors. *Journal of Statistical Planning and Inference* **62**, 193–205.

Holz, H., and Mosler, K. (1994) An interactive decision procedure with multiple attributes under risk. *Annals of Operations Research* **52**, 151–170.

Levy, H. (1992) Stochastic dominance and expected utility: survey and analysis. *Management Science* **38**, 555–593.

Levy, H., and Sarnat, M. (1984) *Portfolio and Investment Selection: Theory and Practice*. Englewood Cliffs, NJ: Prentice-Hall.

Markowitz, H.M. (1977) An algorithm for finding undominated portfolios. In: *Financial Decision Making under Uncertainty* (eds. H. Levy and M. Sarnat), pp. 3–10. New York: Academic Press.

Porter, R., Wart, I., and Ferguson, D. (1973) Efficient algorithms for conducting stochastic dominance tests on large numbers of portfolios. *Journal of Financial and Quantitative Analysis* **8**, 71–81.

Discussion Papers in Statistics and Econometrics

Seminar of Economic and Social Statistics
University of Cologne

No.	Author	Title
	Trede, M.	Statistical Inference in Mobility Measurement: Sex Differences in Earnings Mobility
	Bomsdorf, E.	Allgemeine Sterbetafel 1986/88 für die Bundesrepublik Deutschland und Allgemeine Sterbetafel 1986/87 für die DDR – ein Vergleich
	Heer, B.; Trede, M.; Wahrenburg, M.	The Effect of Option Trading at the DTB on the Underlying Stocks' Return Variance
	Schmid, F.; Trede, M.	Testing for First Order Stochastic Dominance: A New Distribution Free Test
1/95	Trede, M.M.	The Age-Profile of Earnings Mobility: Statistical Inference of Conditional Kernel Density Estimates
2/95	Stich, A.	Die axiomatische Herleitung einer Klasse von dynamischen Ungleichheitsmaßen
3/95	Bomsdorf, E.	Ein alternatives Modell zur Reform des Einkommensteuertarifs
4/95	Schmid, F.; Trede, M.	Testing for First Order Stochastic Dominance in Either Direction
5/95	Brachmann, K.	Nichtparametrische Analyse parametrischer Wachstumsfunktionen — Eine Anwendung auf das Wachstum des globalen Netzwerks Internet
6/95	Brachmann, K.; Stich, A.; Trede, M.	Evaluating Parametric Income Distribution Models
7/95	Koshevoy, G.A.; Mosler, K.	Multivariate Gini indices
8/95	Brachmann, K.	Choosing the optimal bandwidth in case of correlated data
9/95	Heer, B.; Trede, M.	Taxation of Labor and Capital Income in an OLG Model with Home Production and Endogenous Fertility
10/95	Stich, A.	Insurance and Concentration: The Change of Concentration in the Swedish and Finnish Insurance Market 1989 – 1993
1/96	Barth, W.; Bomsdorf, E.	Besteht Long-Memory in Devisenkursen? — Eine semiparametrische Analyse
2/96	Schmid, F.; Trede, M.	Nonparametric Inference for Second Order Stochastic Dominance
3/96	Schmid, F.; Trede, M.	A Kolmogorov-Type Test for Second Order Stochastic Dominance

Discussion Papers in Statistics and Econometrics

Seminar of Economic and Social Statistics
University of Cologne

No.	Author	Title
4/96	Stich, A.	Inequality and Negative Income
5/96	Stich, A.	Poverty and Life Cycle Effects. A Nonparametric Analysis for Germany
6/96	Eurich, A.; Stich, A.; Weidenfeld, G.	Die Entwicklung der Anbieterkonzentration auf dem deutschen Erstversicherungsmarkt von 1991 bis 1994
1/97	Stich, A.	Simultaneous Inference for Proportions in Arbitrary Sampling Designs
2/97	Trede, M.	Making Mobility Visible: A Graphical Device
3/97	Mosler, K.; Seidel, W.; Jaschinger, C.	A Power Comparison of Homogeneity Tests in Mixtures of Exponentials
1/98	Dyckerhoff, R.; Holz, H.; Mosler, K.	Checking for orthant orderings between discrete multivariate distributions: An algorithm