R package GeneralCorr functions for Portfolio Choice

H. D. Vinod *

November 9, 2021

Abstract

We explain the usage of following new R functions in my package called 'generalCorr.' `sudoCoefParcor()` for pseudo regression coefficients for kernel regressions. `decileVote()`, `momentVote()`, `exactSdMtx()` for exact computation of stochastic dominance from ECDF areas. `dif4mtx()` computes growth, change in growth etc. up-to order 4 differencing of time series. We illustrate all these functions with a toy example of only seven observations. The last section has the code which produced various tables in this document.

1 Introduction

We assume that a portfolio manager allocates the available capital (=1 million dollars, say) among p stocks (assets, prospects) with relative weights,

$$w_1 \geq w_2 \geq \ldots \geq w_p, \quad \Sigma_j w_j = 1. \tag{1}$$

We choose w_j by studying the inequalities among probability distributions of best predictions of future returns,

$$f(x_1) \geq f(x_2) \geq \ldots \geq f(x_p). \tag{2}$$

Since future returns are never known, one assumes that the researcher has stable estimates of future densities $f(x_j)$ using past data.

Remark 1: A fundamental problem in portfolio analysis is as follows. While comparing individual stock returns on particular dates is straightforward, comparing their densities $f(x_j)$ is challenging because the inequalities among overlapping densities (2) are intrinsically fuzzy.

*address: H. D. Vinod, Professor of Economics, Fordham University, Bronx, New York, USA 10458. E-mail: vinod@fordham.edu. JEL codes C30, C51. Keywords: portfolio choice, poverty ranking, cumulative density, bootstrap, step-function

Remark 2: The 'compensation principle' developed by economists Pareto, Kaldor, Hicks, among others (Blaug, 1962, p.393), allows quantification of (2). If $p = 2$, $f(x_1) > f(x_2)$ means that the portfolio manager who invests everything in x_1 (choosing $w_1 = 1, w_2 = 0$) can more than compensate the manager who invests everything in x_2 (choosing $w_1 = 0, w_2 = 1$).

Economists have long avoided interpersonal utility comparisons since utility experience is too personal, rarely identical across individuals, and exhibits marked change for the same individual over time. Moreover, psychologists have documented that human utility experience is asymmetric with respect to rewards versus losses and sensitive to reward sizes.

Remark 3: Since mathematical formulas for utility functions cannot handle real-world complexities, there is a need to de-link portfolio choice from explicit risk aversion (utility) functions.

The parametric tools to assess (2) include the first four moments (mean, variance, skewness, kurtosis), and deciles. This paper provides single index summaries based on four moments and nine deciles associated with each of the p densities, more directly helping in the choice of w_j. Davidson and Duclos (2000) also avoid utility theory, their reliance on formal statistical tests is problematic. Kopa and Petrova (2018) cite thirty references to applications of SD methods in portfolio choice. Many papers attempt to identify which stock (asset) belongs to SDk-efficient (order k=1,2,3,4) set and which does not. Limitations of these methods include the unrealistic assumption that the risk profile of the investor is known and that weights w_j for stocks (1) within the efficient set should be equal.

In portfolio selection applications, we must work with data on returns $x_{j,t}$ for j-th stock. Denote the sorted magnitudes as $x_{j,(1)}$ to $x_{j,(Nj)}$ using parentheses to denote order statistics. Note that sorting loses all information about the date when a particular stock return was observed (the time subscript). Empirical cumulative distribution functions (ECDFs) mentioned before represent CDFs. An ECDF is always a step function. The widths of ECDF steps equal "differences" between consecutive values between sorted magnitudes.

A numerical measure for SD1 needs computation of the difference between two step functions measuring two ECDFs requiring trapezoidal approximation. We shall see that a fixed (x.ref) with simple ECDFs parallel to the two axes allows exact computation of areas without any approximation. Quantitative measures for higher-order dominance SDk (k=2,3,4) requires (k-1) times integration of $F_{12}(.)$ measured for SD1. They involve sorting and differencing of widths of the previous step. Since differencing reveals new aspects related to the widths of the previous step, SDk of larger k are attempted in the literature.

2 Toy Example with only seven observations, Nj=7.

We use a toy example to explain the intuitive meaning of *stochastic* dominance. Our toy example has only seven (Nj=7) data values, x1=c(2, 5, 6, 9, 13, 18, 21), and x2=c(3, 6, 9, 12, 14, 19, 27). They are already sorted from the smallest to the largest. It is convenient not to use the subscript notation in most numerical examples, since data names in software do not have subscripts.

The idea of one density $f(x_2)$ dominating another $f(x_1)$ in elementary statistics in-

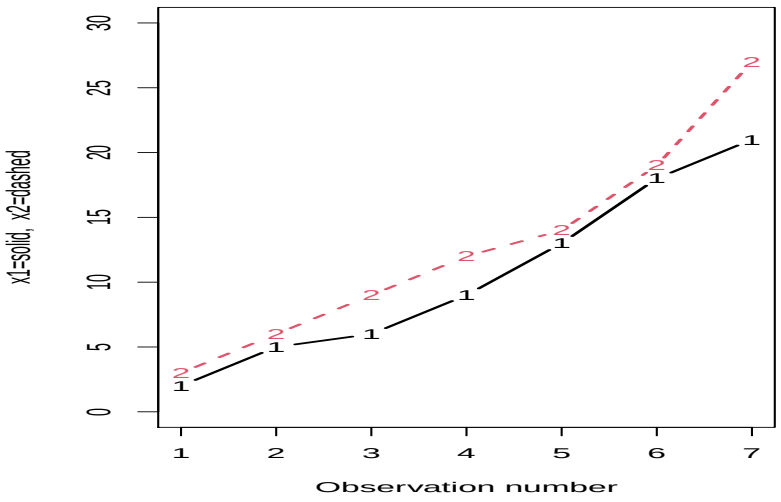

Figure 1: Toy example data for x1 (solid line), x2 (dashed line)

volves comparing their quantiles (deciles), means, standard deviations (sd), and Pearson's measures of skewness (skew) and excess kurtosis (kurt). Recalling the first objective of this paper, the reader can see that the following two subsections develop suitable summary measures for deciles and moments, respectively.

2.1 Decile Comparisons for the toy example:

If the data vectors are incomes, then if deciles of $f(x_2)$ are larger than the corresponding deciles of $f(x_1)$, we can conclude that x2 is richer than x1, or x2 dominates x1 at those deciles. If data vectors refer to stock returns, then if deciles of $f(x_2)$ are larger than corresponding deciles of $f(x_1)$, we can conclude that x2 is a superior investment opportunity than x1, or again x2 dominates x1. Our toy example had (x2.t > x1.t) for all t=1,2,.., 7. Hence, we expect all x2 deciles to exceed those of x1 in Table 1. The vote count in favor of declaring x2 to be dominant over x1 in toy data is also 9/9.

The above output is created by the R command `decileVote(cbind(x1,x2))`

We conclude this subsection by noting that nine deciles summarized by the column sum of votes can directly help in choosing portfolio weights w_j.

2.2 Moment comparisons for the toy example:

The columns entitled 'x1' and 'x2' in Table 2 report the moment-based parametric statistics from x1 and x2 data vectors, respectively. Let the sampling reliability weights (SRW) on the first five rows be (1,1,0.5,0.5,1). They suggest that since skewness and kurtosis are subject to higher sampling variability (involving third and fourth powers of deviations from the mean), they are less reliably estimated. The row 'wtedSumRanks' contains the weighted

Table 1: Decile comparison: toy example x1 and x2 with Nj=7

	x1	x2	x2-x1	vote
10%	3.8	4.8	1.0	1
20%	5.2	6.6	1.4	1
30%	5.8	8.4	2.6	1
40%	7.2	10.2	3.0	1
50%	9.0	12.0	3.0	1
60%	11.4	13.2	1.8	1
70%	14.0	15.0	1.0	1
80%	17.0	18.0	1.0	1
90%	19.2	22.2	3.0	1
colsum	92.6	110.4	17.8	9

sum of ranks measuring choice of the relative desirability of the asset. The R command for this work is `momentVote(mtx)`

Table 2: Comparison of moments and Sharpe Ratio. Lower panel has reliability ranks. Positive (resp. negative) weights in the column 'sampRelWt' suggest larger (smaller) values are more desirable. The last row has the implied choice of the asset.

	x1	x2	sampRelWt
mean	10.5714	12.8571	1
sd	7.0441	8.1533	-1
skewness	0.3380	0.5726	0.5
kurtosis-3	-1.2853	-0.6544	-0.5
Sharpe Ratio	1.5008	1.5769	1
Rank.mean	2	1	1
Rank.sd	1	2	-1
Rank.skewness	2	1	0.5
Rnk.kurtosis-3	1	2	-0.5
Rank.ShRatio	2	1	1
wtedSumRanks	6.5	5.5	
choice	2	1	

The row labeled 'Sharpe Ratio' reports the well-known average risk-adjusted return (ratio of mean to standard deviation). Our toy example x2 dominates x1 by construction. We conclude this subsection by noting that moments confirm the dominance of x2 over x1.

2.3 Stochastic Dominance Computation and Toy Data

This section describes our improvements to Anderson's computation of SDk. We use the toy data for illustration and include graphs of ECDFs to explain computation of SDk. Anderson's algorithm compares only two densities $f(x_j)$ at a time. If one wants to select

from p=1000 stocks, Anderson's algorithm will need to compare each SDk for (p!/[(p-2)!2!]) or 499,500 pairs. No wonder his algorithm is not popular on Wall Street.

Our algorithm includes a unique fictitious reference stock (x.ref), assumed to yield the lowest return m0 during all (N.ref) time periods. Thus every available stock will dominate (x.ref) by construction. Hence, we compute a unique set of n comparable numbers measuring the extent to which each stock beats (x.ref) for each SDk. Even if one has p=1000 stocks to choose from, there are only n sets of SDk numbers. A portfolio manager can rank n stocks by their SDk numbers to help determine its weight w_j in the portfolio.

Construction of a fictional stock (x.ref) with the lowest return m0

We need additional notation to create fictional stock (x.ref) to make sure that its return is "slightly lower" than the lowest in the set of n stocks in the data at hand. Our data values are ($x_{j.1}, x_{j.2}$, to $x_{j.Nj}$). The sorted data values from the smallest to the largest are order statistics denoted by $(x_{j.(1)}, x_{j.(2)}, .., x_{j.(Nj)})$ with added parentheses.

Let mj=min(x_j) denote the smallest return, and note that it is also the first value $x_{j.(1)}$ in the sorted set. Also, let Mj=max(x_j) denote the largest return and the last value $x_{j.(Nj)}$ in the sorted set. Denote by σ_j the standard deviation of x_j. We use a fixed multiple ξ (default ξ=0.10) of $max(\sigma_j)$ in defining "slightly lower."

Recall that we want (x.ref) stock to yield m0 as slightly lower than the lowest return. We make m0 as the lower limit of the common support range of all n stock densities $f(x_j)$ under consideration here. The number of data points in the fictional stock are (N.ref)=max(Nj). Thus, fictional stock returns are (x.ref)=(m0, m0, ..., m0), with (N.ref) repetitions.

Common support range [m0, m*] for n densities being compared

We want our algorithm to be applicable to many diverse data sets, without having adjust end points. Hence, we define our common support range for all n densities $f(x_j)$ as follows:

$$m0 = min(m1, m2, , mn) - \xi * max(\sigma_j) \quad (3)$$

$$m* = max(M1, M2, , Mn) + \xi * max(\sigma_j) \quad (4)$$

For example, the toy data with seven observations has m0=1.18 and m*=27.82 rounded to two places.

Three ECDFs for (x.ref), toy data x1, and x2

Figure 2 depicts ECDF(x1) as a solid line step function and ECDF(x2) as a dashed line. Since (x.ref)=(1.18, 1.18, ..., 1.18) repeated seven times, the imaginary stock reaches its return m0 immediately at the start and stays there for all seven time periods. The vertical dotted line along with a horizontal line at $F_n(x) = 1$ together depict (x.ref)'s ECDF. The simplicity and accuracy of our algorithm comes from the fictional ECDF(x.ref) being parallel to the two axes.

Recall that the difference between two CDFs denoted by $F_{12}(x) = F(x_1) - F(x_2)$ appears in the definition of SD1 dominance in the literature. Our algorithm needs to

consider its empirical equivalent. By construction, each x_j will dominate (x.ref). We can readily compute the area above each step of the step function representing each ECDF(x_j) subject to endpoint adjustments. Now, these n areas are comparable and quantify the dominance of each x_j over (x.ref). Since the toy example has p=2, we have two areas to compare. Since x2 is farthest to the right than x1 in Figure 2, we expect that x2 will have a larger SD1 area number than x1. A larger SD1 number suggests larger portfolio allocation, $w_2 > w_1$, implied by the first-order dominance of x2 over x1.

Remark 4: The ECDF for (x.ref) stock is parallel to the two axes. It involves the vertical axis at m0 and a horizontal line at unit probability, parallel to the horizontal axis. We quantify SD1 for n stocks by the area between the ECDF(x.ref) and ECDF(x_j). It equals the area above the pillars representing the step function depicting ECDF(x_j). Since all pillar widths and relevant heights are known, our algorithm quantifies SD1 areas without approximation. By contrast, area between two step functions with many irregular step widths and heights required by Anderson's formulation is impractical without a trapezoidal approximation.

2.4 Computing endpoint-adjusted areas above ECDFs

The ECDF(x.ref) is parallel to the two axes. After including the left-hand endpoint at m0, the ECDF(x_j) is a step-function having Nj steps. It is customary to associate the following set of probabilities with each step. The toy example has $t = 0, 1, \ldots Nj$, Nj=7, with associated probabilities (0, 1/7, 2/7, .., 1)$\in [0, 1]$, where j=1,2. Figure 2 depicts three ECDF step-functions. A dotted vertical line and horizontal dashed line shows ECDF(x.ref), solid line depicts ECDF(x1) and a dashed line represents ECDF(x2).

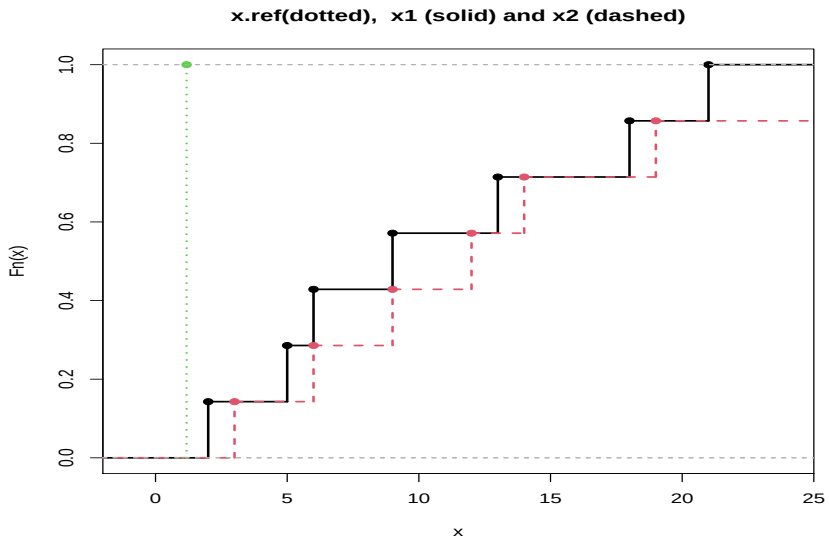

Figure 2: Toy example ECDFs for x.ref (dotted line), x1 (solid line), x2 (dashed line), and areas above x1 and x2

Important properties of ECDFs relevant here: The empirical cumulative distribution function for x_j, ECDF(x_j) is always a well-defined 'sufficient' statistic, in the sense that it uses all available information in the sample. It is important to remember that the ECDF(x_j) is a sequence of (Nj+1) steps after including the first step with zero height. The toy example (x1, x2) has seven steps (with nonzero heights) in Figure 2.

Computing heights above steps representing dominance over (x.ref): Starting with zero height of the first step, the height of each step is always (1/Nj) higher than that of the previous step with all heights in the range [0,1], depicted on the vertical axis. Since Nj=7 for the toy example, the step heights equal (0, 1/7, 2/7,.., 1), the cumulative probabilities are illustrated in Figure 2. The area representing the dominance of each x_j over (x.ref) is *above* the ECDF(x_j) for each x_j. Denote the heights of 'pillars' above the t-th step for x_j by hj.t. For the toy example, hj.t = (1, 6/7, 5/7, .., 1/7, 0), for (t=1, 2, . . . , Nj+1).

Computing Step Widths: First, we insert the lower limit m0 of the support range of the common support [m0, m*] of all x_j in eq. (3). Now, individual elements ($x_{j.t}$), for $t = 0, 1, 2, \ldots Nj$ of data x_j become (m0, $x_{j.1}, x_{j.2}, \ldots, x_{j.Nj}$). The step widths are computed from differences between sorted order statistics $x_{j.(t)}$ as: $\Delta x_{j.t} = [x_{j.(t)} - x_{j.(t-1)}]$, defined for $t \in \{1, 2, 3, \ldots Nj\}$. Since the area above (Nj+1)-th width, [hj.(Nj+1)], is always zero, the right-hand endpoint adjustment amounts to setting the last area =0. That is, $R_{1j.t} = 0$ when $t = Nj + 1$.

What are the step widths for the toy example? Verify that consecutive x1 widths are: $\Delta(x_{j.(t)})$=((m1-m0), 3, 1, 3, 4, 5, 3), for j=1 and t=1,2,3,..,7. Since x2.(t)=(m0, 3, 6, 9, 12, 14, 19, 27), the x2 widths are $w_{2.t}$=((m2-m0), 3, 3, 3, 2, 5, 8) for j=2 and t=1,2,3,..,7.

Recall that SD1(x_j) denotes an index measuring first order stochastic dominance of x_j over (x.ref). We compute the aggregate area above the ECDF(x_j) as a Stieltges summation (integral) directly computed from adding width times height discussed above as:

$$SD1(x_j) = \Sigma_{i=1}^{Nj} R_{1j.t}, \text{ where } R_{1j.t} = \Delta(x_{j.(t)})\, h_{j.t}, \tag{5}$$

where $R_{1j.t}$ is a product of $\Delta(x_{j.(t)})$ the first difference among x_j data order statistics $((x_{j.(t)})-(x_{j.(t-1)})$ and corresponding heights $h_{j.t} = (1, (Nj-1)/Nj, (Nj-2)/Nj, \ldots, 1/Nj)$. The notation $R_{1j.t}$ suggests the 'R'ight-hand side of SD1(x_j) in (5). We shall define analogous right-hand sides $R_{kj.t}$ of SDk(x_j) for stochastic dominance of higher orders, k=2, 3, 4, in the sequel.

It can be verified for the toy example that (SD1(x1)+m0) or (9.3867579+1.184670677) equals the mean of x1= 10.57143. Similarly, SD1(x2)=11.672472, plus m0 equals the mean of x2= 12.85714. Thus ordering by SD1 is close to the ordering by respective means of the series and satisfies the "compensation principle" that the winner comes ahead even after paying off the loser.

Now we quantify the integration in SD2. We focus on the right-hand side (RHS) of (5). It has a summation of Nj components denoted by $R_{1j.t}$ computed from certain widths times heights. Denote the order statistics of $R_{1j.t}$ by $R_{1j.(t)}$, ordered from the smallest to the largest. They induce an ECDF needed for SD2 quantification. As before, we compute exact areas from products of component differences denoted by $\Delta(.)$ times relevant heights $h_{j.t}$ of pillars above their ECDF steps. Thus our second-order stochastic dominance is

measured by analogous areas using the $\Delta^2(.)$ applied to the RHS components.

$$SD2(x_j) = \Sigma_{i=1}^{Nj} R_{2j.t}, \text{where } R_{2j.t} = \Delta^2(.)\, h_{j.t}, \tag{6}$$

The left panel of Figure 3 depicts computations leading up to SD1(x_j) of (5). The right panel of Figure 3, which refers to second-order dominance, implements (6).

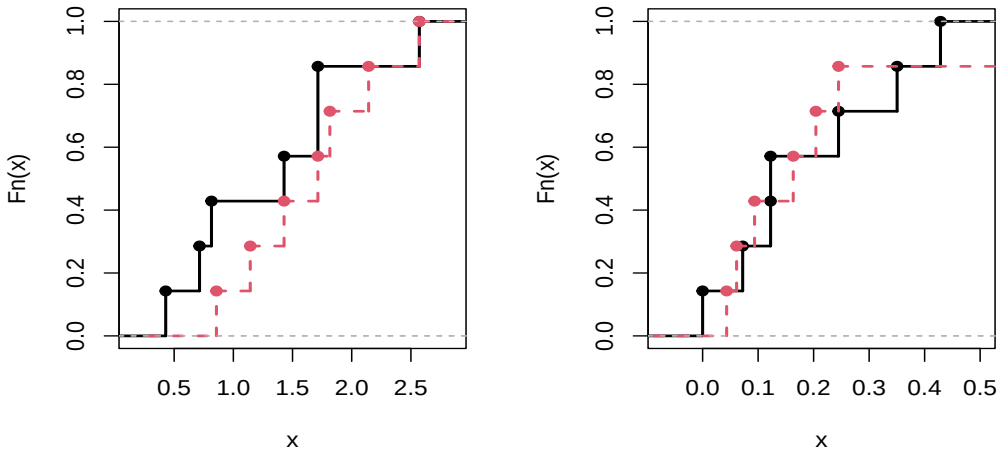

Figure 3: Toy example SD1, order-1 stochastic dominance over (x.ref), is in the left panel depicting x1 (solid line), x2 (dashed line). Similar SD2 is in the right panel.

One can define higher-order SD3(x_j) and SD4(x_j) sequentially by sorting the RHS components produced by the SD(k-1) calculations. The left panel of Figure 4 depicts computations leading up to SD3(x_j), while the right panel depicts SD4(x_j) for j=1,2.

Table 3 reports toy example numerical comparison of dominance areas from step widths for four orders. The areas are higher for x2, as expected. Our higher-order SDk(x_j) involves sequential sorting of areas. They are consistent with the presumption that x2 dominates x1. Note that the dashed line (for x2) depicted in the right panel of Figure 3 does not always stay to the right side of the solid line (for x1).

Unlike Anderson's, our area computations are exact: The quantification in Anderson (1996) has analogous aims. Just as our ECDF interval's widths are unequal, Anderson's more complicated distances are unequal d_j values carefully defined over the merged data from both x1 and x2. He computes certain integrals by using the "trapezoidal approximation," subject to the well-known truncation error. Our computations of the areas above ECDF(x_j) proposed here are exact because we work with areas of rectangles defined by step-functions, not trapezoids.

Comparing SD1(x_j) across j=1,2, .., n: For any data vector x_j with Nj items, the step-function representing their ECDF has Nj rectangles above it comprising the summation in eq. (5). Our algorithm computes SD1(x_j) for all j=1,2, .., n data vectors, ready for ranking. It goes on to compute SD2, SD3 and SD4. They are summarized by the command

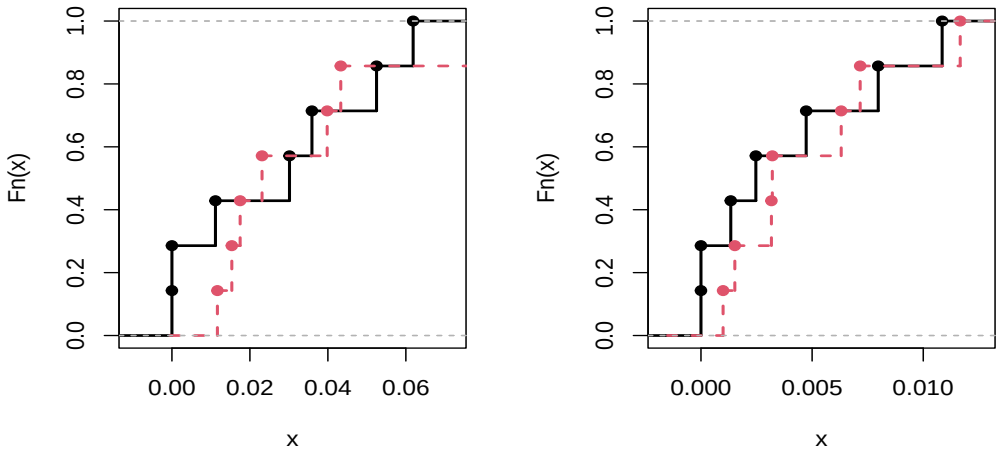

Figure 4: Toy example stochastic dominance over (x.ref) order 3 SD3 in the left panel for x1 (solid line), x2 (dashed line), and order 4 dominance SD4 in the right panel.

`e1=exactSdMtx(cbind(x1,x2))`, and `summaryRank(e1$out)`. The time series differencing is done by the last line of the code below using `dif4mtx()`.

Table 3: Stochastic dominance of four orders for the toy example with N=7. The ranks in the lower panel show x2 dominates x1

	x1	x2
SD1	9.38676	11.67247
SD2	1.34097	1.66750
SD3	0.19157	0.23821
SD4	0.02737	0.03403
SD1.1	2	1
SD2.1	2	1
SD3.1	2	1
SD4.1	2	1
sumRanks	8	4
choice	2	1

3 R code for producing various tables above

```
rm(list=ls())
options(prompt = " ", continue = "  ", width = 68,
useFancyQuotes = FALSE)
library(generalCorr)
options(np.messages=FALSE)
set.seed(99)
x1=c(2,5,6,9,13,18,21)
x2=c(3,6,9,12,14,19,27)
y=2*x1+3*x2+rnorm(7,mean=1)
sy=scale(y)
sx1=scale(x1)
sx2=scale(x2)
coef(lm(sy~sx1+sx2-1))
mtx=cbind(sy,sx1,sx2)
colnames(mtx)=c("sy", "sx1","sx2")
sudoCoefParcor(mtx)
sudoCoefParcorH(mtx)

decileVote(cbind(x1,x2))

momentVote(cbind(x1,x2))

e1=exactSdMtx(cbind(x1,x2))
summaryRank(e1$out)

dif4mtx(cbind(x1,x2))
```

References

Anderson, G. (1996), "Nonparametric Tests of Stochastic Dominance in Income Distributions," *Econometrica*, 64(5), 1183–1193.

Blaug, M. (1962), *Economic Theory in Retrospect*, Homewood, Illinois: Richard D Irwin, Inc, URL https://www.amazon.com/Economic-Theory-Retrospect-Mark-Blaug/dp/0521577012.

Davidson, R. and Duclos, J.-Y. (2000), "Statistical Inference for Stochastic Dominance and for the Meaurement of Poverty and Inequality," *Econometrica*, 68(6), 1435–1464.

Kopa, M. and Petrova, B. (2018), "Strong and Weak Multivariate First-Order Stochastic Dominance," *SSRN eLibrary*, URL https://ssrn.com/abstract=3144058.