Guide for the Extended Approach

Schmid and Schmidt (2007) introduced the generalized Spearman's ρ to directly model high dimensional associations among random variables instead of approximating from pairwise measures. This paper is a great example on conducting rigorous statistical research with proofs from first principle, and it aided our development in understanding intrinsically what copula methods ask for pairs selection.

However, it's quite some technical material to go over and understand thoroughly in a short time. Therefore for this skillset challenge we provide further guidance for you by summing up the key parts you need to know.

What is in the paper?

Spearman's ρ is a nonparametric rank statistic. All the calculation is done using rank data, instead of real values and hence it is distribution free. This quantity is key for selecting pairs for copula-based statistical arbitrage strategies. Kendall's τ is another commonly used quantity that has similar properties. But the original formula only works for 2 random variables.

This paper provided 3 proposed estimators for high dimensional generalization for Spearman's ρ. They are pretty similar to each other, and once you understand one of them, the rest two are trivial.

The way the notations are set in the paper might pose some confusions for people who have not worked in this field. All you need to read is page 4, and here are some further breakdowns.

Procedure

There are d number of random variables (think about stocks daily returns) observed from day 1 to day n. X_{ij} means stock returns data for i-th stock at day j. X_i means the i-th stock's return as a random variable. In your case $d = 4$, since you are dealing with 4 stocks as a cohort.

1. Find the empirical cumulative density function (ECDF) $\hat{F}_i$ for stock i. The formula is given below but you should use `ECDF` from `statsmodels` package.

$$\hat{F}_i(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{X_{ij} \leq x}, \quad \text{where } x \in \mathbb{R}$$

2. Calculate quantile data (or equivalently pseudo-observations) for each X_i, by

$$\hat{U}_i = \frac{1}{n}(\text{rank of } X_i) = \hat{F}_i(X_i)$$

Note this quantity is in $[0, 1]$. And you should use ECDF to calculate it.

3. Ignore the empirical copula definition for now. You don't need to understand the concept to implement the algorithm.

4. The formula for the three estimators are given below, as in the paper. Be aware that even if $d = 2$ they are not the traditional Spearman's ρ.

$$\hat{\rho}_1 = h(d) \times \left\{ -1 + \frac{2^d}{n} \sum_{j=1}^{n} \prod_{i=1}^{d} (1 - \hat{U}_{ij}) \right\}$$

$$\hat{\rho}_2 = h(d) \times \left\{ -1 + \frac{2^d}{n} \sum_{j=1}^{n} \prod_{i=1}^{d} \hat{U}_{ij} \right\}$$

$$\hat{\rho}_3 = -3 + \frac{12}{n\binom{d}{2}} \times \sum_{k<l} \sum_{j=1}^{n} (1 - \hat{U}_{kj})(1 - \hat{U}_{lj})$$

Where:

$$h(d) = \frac{d+1}{2^d - d - 1}$$

The first two are pretty straight forward, the third one is a bit daunting. Let me give an example for $d = 3, j = 2$ in the double sum part. To avoid confusion and simplify notations I write $\hat{U}_{ij}$ as $U_{i,j}$ temporarily.

$$\begin{aligned}\sum_{k<l} \sum_{j=1}^{2} (1 - U_{k,j})(1 - U_{l,j}) &= (1 - U_{1,1})(1 - U_{2,1}) + (1 - U_{1,2})(1 - U_{2,2}) \\ &+ (1 - U_{1,1})(1 - U_{3,1}) + (1 - U_{1,2})(1 - U_{3,2}) \\ &+ (1 - U_{2,1})(1 - U_{3,1}) + (1 - U_{2,2})(1 - U_{3,2})\end{aligned}$$

Think about how to implement this double sum elegantly. This is a typical coding interview question.

Have fun!