QuantMinds International, 2024: Event Guide Featured Article

Beating the Correlation Breakdown, for Pearson's and Beyond: Robust Inference and Flexible Scenarios and Stress Testing for Financial Portfolios

JD Opdyke, Chief Analytics Officer, Partner, Sachs Capital Group Asset Management, LLC
JDOpdyke@gmail.com, 2024

NOTE: This article summarizes a chapter in my forthcoming monograph for Cambridge University Press.

(link to chapter provided here: **http://www.datamineit.com/DMI_publications.htm**)

INTRODUCTION

We live in a multivariate world, and effective modeling of financial portfolios, including their construction, allocation, forecasting, and risk analysis, simply is not possible without explicitly modeling the dependence structure of their assets. Dependence structure can drive portfolio results more than many other parameters in investment and risk models – sometimes even more than their combined effects – but the literature provides relatively little to define the finite-sample distributions of dependence measures in useable and useful ways under challenging, real-world financial data conditions. Yet this is exactly what is needed to make valid inferences about their estimates, and to use these inferences for a myriad of essential purposes, such as hypothesis testing, dynamic monitoring, realistic and granular scenario and reverse scenario analyses, and mitigating the effects of correlation breakdowns during market upheavals (which is when we need valid inferences the most).

The attached is a summary of a chapter of my forthcoming monograph (of the same title) that introduces a new and straightforward method – Nonparametric Angles-based Correlation ("NAbC") – for defining the finite-sample distributions of a very wide range of dependence measures for financial portfolio analysis. These include ANY whose matrix of pairwise associations is positive definite, such as the foundational Pearson's product moment correlation matrix, rank-based measures like Kendall's Tau and Spearman's Rho, as well as measures designed to capture highly non-linear and/or cyclical dependence such as the tail dependence matrix, Chatterjee's correlation, Lancaster's correlation, and Szekely's distance correlation, along with their many variants.

Motivation for NAbC's development has been its effective application to real-world financial portfolios (as opposed to textbook settings), so the solution is characterized by seven critically necessary results that no other method provides simultaneously:

1. NAbC remains valid under challenging, real-world data conditions, with marginal asset distributions characterized by notably different and varying degrees of serial correlation, (non-)stationarity, heavy-tailedness, and asymmetry[1]

2. NAbC can be applied to ANY positive definite dependence measure, including those listed above

3. NAbC remains "estimator agnostic," that is, valid regardless of the sample-based estimator used to estimate any of the above-mentioned dependence measures

4. NAbC provides valid confidence intervals and p-values at both the matrix level and the pairwise cell level, with analytic consistency between these two levels (i.e. the confidence intervals for all the cells define that of the entire matrix, and the same is true for the p-values; this effectively facilitates attribution analyses)

5. NAbC provides a one-to-one quantile function, translating a matrix of all the cells' cdf values to a (unique) correlation/dependence measure matrix, and back again, enabling precision in reverse scenarios and stress testing

6. all the above results remain valid even when selected cells in the matrix are 'frozen' for a given scenario or stress test – that is, unaffected by the scenario – thus enabling flexible, granular and realistic scenarios

7. NAbC remains valid not just asymptotically, i.e. for sample sizes presumed to be infinitely large, but rather, for the specific sample sizes we have in reality,[2] enabling reliable application in actual, real-world, non-textbook settings

METHOD SUMMARY

The key to NAbC's utility and broad range of application is its use of the angles between the pairwise vectors of returns, rather than the values of the pairwise correlations/dependence measures themselves.[3] One angle corresponds to one correlation/dependence measure value, and the entire matrix of angles uniquely identifies the matrix of correlation/dependence measure values, and vice versa. There are four important reasons for the use of angles here:

A. **Automatic Positive Definiteness**: First, proper use of these angles places us on the **unit hyper(-hemi)sphere, where only positive definite samples exist**. This not only ensures that the sample space is valid, but also makes sampling from it efficient and fast.

[1] These obviously are not the only defining characteristics of such data, but from a distributional and inferential perspective, they remain some of the most challenging, especially when occurring concurrently as they do in non-textbook settings.

[2] This is conditional upon n>p, that is, the matrix is full rank, with more observations than assets. It cannot be positive definite otherwise.

[3] The bivariate case of this is simply the widely known and used "cosine similarity." The multivariate case, i.e. the matrix analogue to the bivariate case, is well established and widely used in the literature (see Pinheiro and Bates, 1996, Rebonato and Jackel, 2000, Rapisarda et al., 2007, Pouramadi and Wang, 2015, and Cordoba et al., 2018).

B. **<u>Distributional Independence</u>**: Secondly, and crucially, **the distribution of each of these angles is *<u>independent</u>* with respect to those of the others**. This is critically important for practical usage as it enables the straightforward construction of the multivariate distribution of a matrix of angles, and thus, that of the correlation/dependence measure, which is the more important objective here (vs merely sampling).

C. **<u>Full Information</u>**: Thirdly, **the angles between pairwise data vectors contain ALL the information that exists regarding dependence between the two variables**. The only information we lose in our jump to the hyper(-hemi)sphere is scale, and by design, scale remains irrelevant for dependence measures.

D. **<u>General Conditions</u>**: Finally, **the relationship between angles and the values of correlations/dependence measures holds under the most general conditions: the pairwise matrix simply needs to be symmetric positive definite**. This weak condition remains true for essentially all dependence measures used in quantitative finance, and beyond, so NAbC has an extremely broad range of application.

The NAbC solution is sometimes even available in fully analytic form, such as for the narrow but foundational case of the Gaussian identity matrix (see below, and the link for the spreadsheet containing this solution: http://www.datamineit.com/DMI_publications.htm).

$$f_X(x) = c_k \cdot \sin^k(x),\ x \in (0,\pi),\ k = 1,2,3\ldots\#\text{columns} - 1, \text{ and } c_k = \frac{\Gamma(k/2+1)}{\sqrt{\pi}\,\Gamma(k/2+1/2)}$$

$$F_X(x;k) \sim \frac{1}{2} - \left(\frac{1}{2}\right) \cdot F_{Beta}\left[\cos^2(x); \frac{1}{2}, \frac{1+k}{2}\right] \text{ for } x < \frac{\pi}{2},$$

$$\sim \frac{1}{2} + \left(\frac{1}{2}\right) \cdot F_{Beta}\left[\cos^2(x); \frac{1}{2}, \frac{1+k}{2}\right] \text{ for } x \geq \frac{\pi}{2}$$

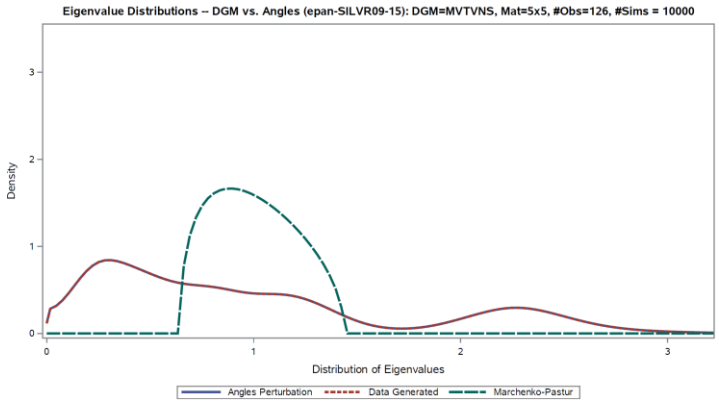
$$F^{-1}(p;k) = \arccos\left(\sqrt{F_{Beta}^{-1}\left(1-2p; \frac{1}{2}, \frac{1+k}{2}\right)}\right) \text{ for } p < 0.5;$$

$$= \pi - \arccos\left(\sqrt{F_{Beta}^{-1}\left(1-2[1-p]; \frac{1}{2}, \frac{1+k}{2}\right)}\right) \text{ for } p \geq 0.5$$

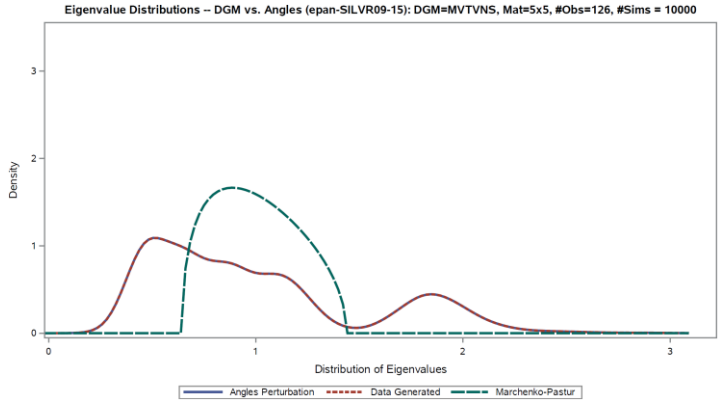
But NAbC extends way beyond this specific case to check all seven of the 'objectives' boxes above simply by estimating the distributions of the angles nonparametrically, via kernels. The implementation details are well established in the literature, straightforward, and shown in the attached article, which also presents a complete example of NAbC's implementation on Kendall's Tau under challenging data conditions, step-by-step. This example includes both the unrestricted and scenario-restricted cases, as described in the next section (I also include below spectral distributions of just a few of the dependence measures covered in the full chapter article attached).

Graph 1: Spectral Distribution-NAbC Angles Kernel v Data Simulations v Marchenko Pastur

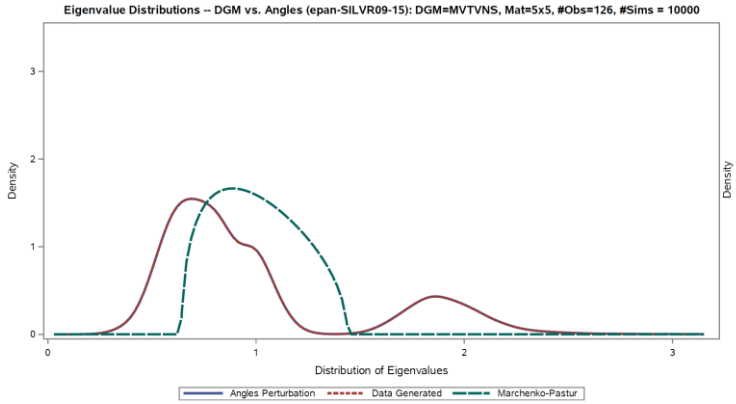
Pearson's Rho



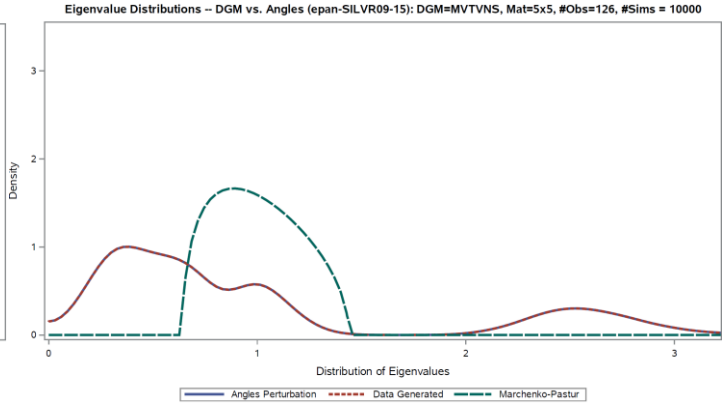
Kendall's Tau



Chatterjee's



Spearman's Rho+Chatterjee



Flexible Scenarios, Reverse Scenarios, and Realistic Stress Testing

Within the framework of the matrix of all pairwise associations,[4] NAbC exploits several results to provide full flexibility for scenario analytics. First, as mentioned above, i. independence of the angles distributions allows us to vary individual cells; second, ii. as established in the literature, the distributions of individual correlation cells, as well as the distribution of the entire correlation matrix, both remain invariant to the ordering of the rows and columns of the matrix (see Pourahmadi and Wang, 2015, and Lewandowski et al., 2009). Third, iii. based on i. and ii., we can exploit the simple mechanics of matrix multiplication so that only selected cells of the matrix are affected, and the rest frozen, as required for a given scenario: all that is required for this is a simple reordering of the rows and columns of the matrix. Taken together, these three results, and NAbC's use of the "all-pairwise" framework, allow us to specify that ANY subset of cells within the structure of the all-pairwise matrix remain 'frozen,' i.e. unaffected by the scenario, thus eliminating the effects of so-called 'peripheral' correlations/associations. No other

[4] Note that some of the abovementioned dependence measures can be implemented on a multivariate basis, and sometimes even in arbitrary and differing dimensions (e.g. Szekely's distance correlation, and variants of Chatterjee's correlation). However, multivariate dependence (as distinct from "all-pairwise" bivariate dependence) imposes limitations that NAbC avoids, as discussed in more detail in the attached chapter.

method provides anything close to this level of (valid) scenario targeting and flexibility. What's more, NAbC provides ancillary but potentially game-changing benefits beyond its immediate design purposes, including a new "Generalized Entropy," as well as effective use within the paradigms of Causal Modeling.

Generalized Entropy

NAbC provides p-values, consistent across both the cell level and matrix level, that demonstrate a remarkable correspondence with the state-of-the-art entropy of the correlation/dependence matrix as derived and calculated in Felippe et al. (2021 and 2023). Yet the latter remains restricted to the case of perfect (in)dependence, whereas NAbC can provide the same entropy calculation using ANY values of the correlation/dependence measure as its baseline. Furthermore, NAbC's 'generalized entropy' is more granular and robust, as it is based on p(p-1)/2 cells, as opposed to only p eigenvalues. Finally, as a measure of 'distance,' NAbC's generalized entropy has multiple advantages over commonly used norms (e.g. the taxi, Chebychev's, and Euclidean/Frobenius norms (collectively, the Minkowski norm)) as it rests on a solid probabilistic foundation, while norms do not and consequently, often lack interpretation in this setting. Entropy has been used increasingly in the literature to measure, monitor, and analyze financial markets (see Meucci, 2010b, Almog and Shmueli, 2019, Chakraborti et al., 2020, and Vorobets, 2024a, 2024b, for several examples), so this 'generalized entropy' is not only highly relevant, but also very intriguing with possibly far-reaching consequences in this setting.

Causal Modeling

Finally, even as an association-based method that broadens, enables, and enhances robust statistical inference in challenging, real-world financial settings, NAbC can be used to tackle questions posed within causal modeling paradigms. Its broad range of application allows for its use on asymmetric, DIRECTIONAL dependence measures, including Chatterjee's new correlation coefficient (Chatterjee, 2021), the improved Chatterjee's coefficient (Xia et al., 2024), Zhang's (2023) combined correlation measure, the QAD measure of Junker et al. (2021), the asymmetric tail dependence measure (Deidda et al, 2023), and others. Because these all are DIRECTIONAL, we can map their inferential results – that is, their individual, cell-level p-values – to the different variable effect classifications of models based on directed acyclical graphs (DAGs)): the mediators, confounders, and colliders, as well as the vanilla causal and 'caused by' covariates (see MacKinnon & Lamp, 2021). All it takes is two runs of NAbC, one in each 'direction.' The subsequent mapping of results is exhaustive and mutually exclusive, so we can proceed with a rigorous, inferential analysis that identifies, probabilistically, the 'causal' relationships between the variables. This obviously does not address the bigger question, however, of whether DAGs can be used reliably within "self-referencing open systems like capital markets" (Polakow et al., 2023); only that it appears NAbC can play a role in recovering them if the answer to this question is "yes" or "under some conditions."

CONCLUSION

To date, financial portfolio analytics in practice very often relies on ad hoc, largely qualitative, and 'judgmental' approaches to specifying and utilizing dependence structure, and when quantitative approaches are used, their valid application largely has been restricted to narrow cases. With NAbC, however, we now have a powerful, applied approach enabling us to treat an extremely broad class of ubiquitous dependence measures with the same level of analytical rigor as the other major parameters in our (finite sample) financial portfolio models. NAbC's utility holds under the most challenging, real-world financial data conditions, and for extremely flexible and targeted scenarios. We can use NAbC in frameworks that identify, probabilistically measure and monitor, and even anticipate critically important events, such as correlation breakdowns, and mitigate and manage their effects. It can even be used within causal paradigms! It should prove to be a very useful means by which we can better understand, predict, and manage portfolios in our multivariate world.