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.

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.[1] 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

[1] I take 'real-world' financial returns data to be multivariate with marginal distributions that can vary notably from each other in their degrees of heavy-tailedness, serial correlation, asymmetry, and (non-)stationarity. 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.

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 following 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 (file downloads at http://www.datamineit.com/DMI_publications.htm). These include ANY that are positive definite, such as the foundational Pearson's product moment correlation matrix (Pearson, 1895), rank-based measures like Kendall's Tau (Kendall, 1938) and Spearman's Rho (Spearman, 1904), as well as measures designed to capture highly non-linear and/or cyclical dependence such as the tail dependence matrix (see Embrechts, Hofert, and Wang, 2016, and Shyamalkumar and Tao, 2020), Chatterjee's correlation (Chatterjee, 2021), Lancaster's correlation (Holzmann and Klar, 2024), and Szekely's distance correlation (Szekely, Rizzo, and Bakirov, 2007) and their many variants (such as Sejdinovic et al., 2013, and Gao and Li, 2024).[2]

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[3]

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

---

[2] Note that "positive definite" herein refers to the dependence measure calculated on the matrix of all pairwise associations in the portfolio, that is, calculated on a bivariate basis. Some of these dependence measures (e.g. Szekely's correlation and variants of Chatterjee's) can be applied on a multivariate basis, in arbitrary dimensions, for example, to test the hypothesis of multivariate independence. But "positive definite" herein is not applied in this sense, and I explain below some of the reasons for using the dependence framework of all pairwise associations, which is highly flexible, and allows for more precise, targeted, and hence more effective attribution and intervention analyses.

[3] 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.

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,[4] enabling reliable application in actual, real-world, non-textbook settings

To date, financial portfolio analysis 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. But practitioners, academics, and regulators have a long history of bringing analytic and probabilistic rigor to bear when estimating and analyzing the other parameters of our portfolio risk and investment models. As shown below, NAbC now ensures this same level of rigor can be applied to modeling dependence structure while simultaneously dramatically increasing both robustness and scenario-based flexibility.

PEARSON'S CORRELATION, GAUSSIAN DATA, AND THE IDENTITY MATRIX

We begin with Pearson's product moment correlation matrix, the oldest and arguably most broadly used measure of dependence. Although its limitations often are mischaracterized or misunderstood, especially as they relate to widely held views classifying it strictly as a measure of linear association (see van den Heuvel & Zhan, 2022), in many settings it remains either optimal or centrally relevant for wide-ranging purposes (e.g. robust asset allocation (Welsch and Zhou, 2007), Black-Litterman variants (Meucci, 2010a, Qian and Gorman, 2001), entropy pooling with fully flexible views (Meucci, 2010b; Vorobets, A., 2024a & 2024b), portfolio optimizations combined with random matrix theory (Pafka and Kondor, 2004), stress testing (Bank for International Settlements, Basel Committee on Banking Supervision, 2011), and even non-linear, tail-risk-aware trading algorithms (Li et al., 2022, and Thakkar et al., 2021) to name a few). Consequently, Pearson's is the foundational dependence measure we start with, and the data and correlation structure we initially presume is Gaussian data under no correlation: that is, Pearson correlation values of zero off the diagonal of the matrix as in (1).[5]

(1) identity matrix =

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

for p = 4 assets

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

[5] Note, of course, that a zero value for Pearson's correlation does not imply independence, but independence does imply a zero value for Pearson's correlation.

If we take two variables, such as the returns of two assets, X and Y, over a time period with n observations, we calculate Pearson's correlation coefficient for this sample as (2):[6]

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X\right)\left(Y_i - \frac{1}{n}\sum_{i=1}^{n}Y\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X\right)^2}\sqrt{\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}Y\right)^2}} = \frac{Cov(X,Y)}{s_X s_Y} \quad \textbf{(2)}$$

For the corresponding matrix of all pairwise correlations, we have:

$$R = \begin{bmatrix} 1 & r_{1,2} & r_{1,3} & r_{1,4} \\ r_{2,1} & 1 & r_{2,3} & r_{2,4} \\ r_{3,1} & r_{3,2} & 1 & r_{3,4} \\ r_{4,1} & r_{4,2} & r_{4,3} & 1 \end{bmatrix}$$, with the usual, following characteristics:

i. Symmetry: $r_{i,j} = r_{j,i}$

ii. Unit diagonal entries: $r_{i=j} = 1$

iii. Bounded non-diagonal entries: $-1 \le r_{i,j} \le 1$

iv. The matrix is positive definite, i.e. all eigenvalues $\lambda_i > 0$

For completeness and for reference throughout this article, we define eigenvalues here:

If there exists a nonzero vector *v* such that $Rv = \lambda v$ then λ is an eigenvalue of *R* and *v* is its corresponding eigenvector. λ and *v* can be obtained by solving

$\det(\lambda I - R) = 0,$ then $\det(\lambda I - R)v = 0,$ where I is the identity matrix and det is the determinant

The eigenvalue can be thought of as the magnitude of the (portfolio) variance in the direction of the eigenvector. Also note that with iii. above, this range can be tighter under specific circumstances, such as for equicorrelation matrices where $-1/(\mathrm{p}-1) \le r \le 1,\ \mathrm{p} = \dim(r).$

Correlations to Angles, Angles to Correlations

[6] Recall that Pearson's requires that both the first raw moment (the mean) and the first central moment (the variance) of the distributions of X and Y are finite.

The key to the NAbC approach rests in its use of the ANGLE, $\boldsymbol{\theta}$, between the two mean-centered data vectors of X and Y, as opposed to directly and only using of the values of the correlations themselves. For a single pair of variables, providing a single bivariate correlation value, the relationship between angle value and correlation value is most readily seen in the widely known "cosine similarity," where the cosine of the angle equals the inner product divided by the product of the two vectors' (Euclidean) norms as in (4):[7]

$$\cos(\theta) = \frac{\text{inner product}}{\text{product of norms}} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\|\mathbf{X}\|\|\mathbf{Y}\|} = \frac{\sum_{i=1}^{N}\left(X_i - E(X)\right)\left(Y_i - E(Y)\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i - E(X)\right)^2}\sqrt{\sum_{i=1}^{N}\left(X_i - E(X)\right)^2}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho, \quad \text{with } 0 \le \theta \le \pi$$

(4)

If a portfolio has p assets, the number of its pairwise relationships is npr=p(p-1)/2. For all these npr relationships, the matrix analogue to (4), as long as the matrix is symmetric positive definite,[8] is well established in the literature (Pinheiro and Bates, 1996, Rebonato and Jackel, 2000, Rapisarda et al., 2007, Pouramadi and Wang, 2015, and Cordoba et al., 2018) and shown below, formulaically in (5)-(7) and in code in Table A. The steps for translating between correlations and angles, in both directions, are shown in A.-C. below.

A. estimate the correlation matrix from sample data
B. obtain the Cholesky factorization of the correlation matrix
C. use inverse trigonometric and trigonometric functions on B. to obtain corresponding spherical angles

and in reverse:

C. start with a matrix of spherical angles
B. apply trigonometric functions to obtain the Cholesky factorization
A. multiply B. by its transpose to obtain the corresponding correlation matrix

(see Rebonato & Jaeckel, 2000, Rapisarda et al., 2007, and Pourahmadi and Wang, 2015, but note a typo in the formula in Pourahmadi and Wang, 2015, for the first 3 steps)

Central to this correlation-angle translation mechanism is obtaining the Cholesky factor of the correlation/dependence matrix, which is usually a hard-coded function in most statistical and mathematical software. The relevant formulae are included below for completeness.

[7] While *r* typically is used to represent Pearson's calculated on a sample, *ρ* typically is used to represent Pearson's calculated on a population.

[8] Note that this is true not only for Pearson's, but also for all relevant dependence measures in this setting, as will be discussed below.

(5) A correlation matrix R will be real, symmetric positive-definite, so the unique matrix B that satisfies $R = BB^T$ where B is a lower triangular matrix (with real and positive diagonal entries), and B^T is its transpose, is the Cholesky factorization of R. Formulaically, B's entries are as follows:

$$B_{j,j} = (\pm)\sqrt{R_{j,j} - \sum_{k=1}^{j-1} B_{j,k}^2}, \qquad B_{i,j} = \frac{1}{B_{j,j}}\left(R_{i,j} - \sum_{k=1}^{j-1} B_{i,k}B_{j,k}\right) \text{ for } i > j$$

The Cholesky factor can be viewed as a matrix analog to the square root of a scalar, because like a square root the product of it and its transpose yields the original matrix. Importantly, the Cholesky factor places us on the UNIT hyper-(hemi)sphere (where scale does not matter) because the sum of the squares of its rows always equals one. Next, we recursively apply inverse trigonometric and trigonometric functions to each cell of the Cholesky factor to obtain each cell's angle value; or in reverse, we obtain a correlation/dependence value from each cell's angle value (see Pourahmadi and Wang, 2015, as well as Rapisarda et al., 2007, for a meticulous derivation of these formulas). Note that this relationship is one-to-one, with a unique correlation/dependence matrix yielding a unique angles matrix, and vice versa.

(6)

For R, a p x p correlation matrix,

$$R = \begin{bmatrix} 1 & r_{1,2} & r_{1,3} & \cdots & r_{1,p} \\ r_{2,1} & 1 & r_{2,3} & \cdots & r_{2,p} \\ r_{3,1} & r_{3,2} & 1 & \cdots & r_{3,p} \\ r_{4,1} & r_{4,2} & r_{4,3} & \cdots & r_{4,p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{p,1} & r_{p,2} & r_{p,3} & \cdots & 1 \end{bmatrix},$$

$R = BB^t$ where B is the Cholesky factor of R and

$$B = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \cos(\theta_{2,1}) & \sin(\theta_{2,1}) & 0 & \cdots & 0 \\ \cos(\theta_{3,1}) & \cos(\theta_{3,2})\sin(\theta_{3,1}) & \sin(\theta_{3,2})\sin(\theta_{3,1}) & \cdots & 0 \\ \cos(\theta_{4,1}) & \cos(\theta_{4,2})\sin(\theta_{4,1}) & \cos(\theta_{4,3})\sin(\theta_{4,2})\sin(\theta_{4,1}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \cos(\theta_{p,1}) & \cos(\theta_{p,2})\sin(\theta_{p,1}) & \cos(\theta_{p,3})\sin(\theta_{p,2})\sin(\theta_{p,1}) & \cdots & \prod_{k=1}^{n-1}\sin(\theta_{p,k}) \end{bmatrix}$$

for $i > j$ angles $\theta_{i,j} \in (0, \pi)$.

To obtain an individual angle $\theta_{i,j}$, we have:[9]

$$\text{For } i>1:\quad \theta_{i,1} = \arccos\left(b_{i,1}\right) \text{ for } j=1;\ \text{ and } \theta_{i,j} = \arccos\left(b_{i,j} \Big/ \prod_{k=1}^{j-1}\sin\left(\theta_{i,k}\right)\right) \text{ for } j>1$$

(7) To obtain an individual correlation, $r_{i,j}$, we have, simply from $R = BB^T$:

$$r_{i,j} = \cos\left(\theta_{i,1}\right)\cos\left(\theta_{j,1}\right) + \prod_{k=2}^{i-1}\cos\left(\theta_{i,k}\right)\cos\left(\theta_{j,k}\right)\prod_{l=1}^{k-1}\sin\left(\theta_{i,l}\right)\sin\left(\theta_{j,l}\right) + \cos\left(\theta_{j,i}\right)\prod_{l=1}^{i-1}\sin\left(\theta_{i,l}\right)\sin\left(\theta_{j,l}\right) \text{ for } 1 \le i < j \le n$$

SAS/IML code translating correlations to angles and angles to correlations is shown in Table A below:

TABLE A:

Correlations to Angles

```
* INPUT rand_R is a valid correlation matrix;

cholfact = T(root(rand_R, "NoError"));

rand_corr_angles = J(nrows,nrows,0);
  do j=1 to nrows;
    do i=j to nrows;
      if i=j then rand_corr_angles[i,j]=.;
      else do;
        cumprod_sin = 1;
        if j=1 then rand_corr_angles[i,j]=arcos(cholfact[i,j]);
        else do;
          do kk=1 to (j-1);
            cumprod_sin = cumprod_sin*sin(rand_corr_angles[i,kk]);
          end;
          rand_corr_angles[i,j]=arcos(cholfact[i,j]/cumprod_sin);
        end;
      end;
    end;
  end;

* OUTPUT rand_corr_angles is the corresponding matrix of angles;
```

SAS/IML code (v9.4)

Angles to Correlations

```
* INPUT rand_angles is a valid matrix of correlation angles;

Bs=J(nrows, nrows, 0);
do j=1 to nrows;
  do i=j to nrows;
    if j>1 then do;
      if i>j then do;
        sinprod=1;
        do gg=1 to (j-1);
          sinprod = sinprod*sin(rand_angles[i,gg]);
        end;
        Bs[i,j]=cos(rand_angles[i,j])*sinprod;
      end;
      else do;
        sinprod=1;
        do gg=1 to (i-1);
          sinprod = sinprod*sin(rand_angles[i,gg]);
        end;
        Bs[i,j]=sinprod;
      end;
    end;
    else do;
      if i>1 then Bs[i,j]=cos(rand_angles[i,j]);
      else Bs[i,j]=1;
    end;
  end;
end;
rand_R = Bs*T(Bs);

* OUTPUT rand_R is the corresponding correlation matrix;
```

[9] Note that a similar recursive relationship exists between partial correlations (Madar, 2015), although its sample-generating algorithm it is not generalizable beyond Pearson's correlations, i.e. to all positive definite measures of dependence, as is NAbC as shown below in later sections.

The above all is well-established and straightforward,[10] and demonstrates, as we know intuitively, that **scale does not (and should not) matter when it comes to dependence measures**;[11] again, in this setting, this is because geometrically, the Cholesky factor places us on the UNIT hyper-(hemi)sphere. Importantly, the Cholesky factor also ensures that sampling based directly on the resulting angles will yield only positive definite matrices, as the Cholesky factor remains undefined otherwise. This **automatic enforcement of positive definiteness makes this approach much more efficient than others** that require post-sample verification of positive definiteness, and subsequent resampling when this requirement is violated[12] (see Makalic and Schmidt, 2018, Cordoba et al. 2018, and Papenbrock et al., 2021). This inefficiency grows very rapidly with the size of the matrix/portfolio, as shown in the ratio below in (8) (see Bohn and Hornik, 2024, and Pourahmadi and Wang, 2015).

$$\Pr\left(rand"R" \sim PosDef\right) = X = \frac{\prod_{j=1}^{p-1}\left[\sqrt{\pi}\Gamma\left(\frac{j+1}{2}\right)\right]^{j}}{2^{p(p-1)/2}} < \prod_{j=1}^{p-1}\left[\frac{\sqrt{\pi}}{2}\right]^{j} = \left[\frac{\sqrt{\pi}}{2}\right]^{p(p-1)/2}; \lim_{p\to\infty}\left[X\right] = 0 \quad \textbf{(8)}$$

Even for relatively small matrices of dimension p=25, the odds of successfully randomly generating a single valid positive definite correlation matrix, by uniformly sampling the off-diagonal correlation values themselves across values ranging from –1.0 to 1.0, are less then 2 in 10 quadrillion, leading to prohibitively inefficient sampling. Consequently, even when sampling-rejection algorithms achieve some efficiency gains, realistically the sampling approach in this setting should possess automatic enforcement of positive definiteness. Conceptually, an imperfect but apt analogy is to a rubiks cube: the colored stickers on the cube cannot simply be peeled off and repasted, even some of the time, to solve the cube. The valid solution must be obtained by (always) following the rules governing shifts in the cube, each of which affects many of the individual cubes (cells), not just the one we need to reposition. Similarly with sampling the correlation/dependence matrix: converting to the Cholesky factor (en)forces positive definiteness by forcing the matrix onto the UNIT hyper-(hemi)sphere, where we can subsequently use the distributions of the angles to perturb it and obtain, after re-translation, the distribution of the original correlation/dependence matrix, without violating positive definiteness, simply by following steps A., B., and C., and C., B., and A., above.

Another crucial characteristic of these angles is that **they are random variables whose multivariate relationship is one of independence** (see Pourahmadi and Wang, 2015, Tsay and Pourahmadi, 2017,

[10] Reliance on spherical angles and the hyper(hemi)sphere is not uncommon in quantitative finance, in large part due to its scale invariance: it has even been used effectively to define entire markets (see Kim and Lee, 2016).

[11] Scale invariance is widely proved and cited for Pearson's, Kendall's, and Spearman's (see Xu et al., 2013, and Schreyer et al., 2017 examples).

[12] As shown below, this approach also much more straightforward, not to mention more generalizable, than the other, more complex sampling algorithms that have been proposed, such as the vine and extended onion algorithms of Lewandowski et al. (2009), the similar chordal sparsity method of Kurowicka (2014), the Metropolis-Hastings and Metropolis algorithms of Cordoba et al. (2018), and the restricted Wishart distribution approach of Wang et al. (2018).

and Ghosh et al., 2020). This is critically important for practical usage as it enables the straightforward construction of the multivariate distribution of a matrix of angles, which is the more important objective here (vs merely sampling) and essential for the application of NAbC below.

Finally and most critically, the above demonstrates that **the angles between pairwise data vectors contain ALL the information that exists regarding dependence between the two variables** because the only information we lose by jumping to the unit hyper(hemi-)sphere is scale (see Fernandez-Duren & Gregorio-Dominguez, 2023, and Zhang & Songshan, 2023, as well as Opdyke, 2024a). This will be covered more extensively below.

So with all this in mind we proceed with the use of the angles as described and defined above. The goal is to use the angles as the basis for 1. sample generation of the correlation matrix (dependence measure matrix); and more importantly, 2. definition of the multivariate distribution of the correlation matrix (dependence measure matrix).

Fully Analytic Angles Density, and Efficient Sample Generation

Once we have the matrix of angles, one for each pairwise correlation (dependence measure), we use the well-established finding that, to sample uniformly from the space of positive definite matrices, the probability density function (pdf) must be proportional to the determinant of the Jacobian of the Cholesky factor (9) (see Cordoba, 2018, Pourahmadi and Wang, 2015, Lewandowski et al., 2009).

$$\det\left[J\left(U\right)\right] = 2^{p}\prod_{i=1}^{p-1}u_{ii}^{i} \quad \text{where } U \text{ is the Cholesky factorization of correlation matrix } R = UU^{t} \tag{9}$$

We see directly from (9) that $\sin^{k}\left(x\right)$, suitably normalized in (10), satisfies this requirement (see Pourahmadi and Wang, 2015, and Makalic and Schmidt, 2018).

$$f_{X}\left(x\right) = c_{k}\cdot\sin^{k}\left(x\right),\ x\in\left(0,\pi\right),\ k = 1,2,3,\ldots,\left(\#\text{columns}-1\right),\ \text{and } c_{k} = \frac{\Gamma\left(k/2+1\right)}{\sqrt{\pi}\,\Gamma\left(k/2+1/2\right)} \tag{10}$$

Although not mentioned in Makalic and Schmidt (2018), importantly note that k = #columns – column# (so for the first column of a p=10x10 matrix, k=9; for the second column, k=8, etc.).

However, we need both the cumulative distribution function (cdf) and its inverse, the quantile function, to make use of this density for sampling and other purposes. The most widely used and straightforward method of sampling is inverse transform, whereby the values of a uniform random variate are passed to the quantile function to generate values. Yet regarding the cdf corresponding to (10) above, Makalic and Schmidt (2018) state, “Generating random numbers from this distribution is not straightforward as the corresponding cumulative density [sic] function, although available in closed form, is defined recursively

and requires O(k) operations to evaluate. The nature of the cumulative density [sic] function makes any procedure based on inverse transform sampling computationally inefficient, especially for large k."

Fortunately, that turns out not to be the case, as Opdyke (2020) derived an analytic, non-recursive expression of the cdf below in **(11)**.

$$F_X(x;k) \sim \frac{1}{2} - c_k \cdot \cos(x) \cdot {}_2F_1\left[\frac{1}{2}, \frac{1-k}{2}; \frac{3}{2}; \cos^2(x)\right] \text{ for } x < \frac{\pi}{2},$$

$$\sim \frac{1}{2} + c_k \cdot \cos(x) \cdot {}_2F_1\left[\frac{1}{2}, \frac{1-k}{2}; \frac{3}{2}; \cos^2(x)\right] \text{ for } x \geq \frac{\pi}{2}$$

where the Gaussian hypergeometric function ${}_2F_1[a,b;c;r] = \sum_n^{\infty} \frac{(a)_n (b)_n}{(c)_n} \cdot \frac{r^n}{n!}$

where $(h)_n = h(h+1)(h+2)\cdots(h+n-1),\ n \geq 1,\ (h)_0 = 1,$ and $|r| < 1,\ c \neq 0, -1, -2, \ldots$

Interestingly, the Gaussian hypergeometric function makes many appearances in this setting,[13] but it is admittedly cumbersome mathematically. But Opdyke (2022, 2023, and 2024a) has shown that (11) can be simplified further, based on some arguably obscure hypergeometric identities in **(12)** below:

For $c = a+1$ and $0 < r < 1$ simultaneously, which holds in this setting, we have ${}_2F_1[a,b;c;r] = B(r;a,1-b)\left(a/r^a\right)$

where $B(r;a,b) = \int_0^r u^{a-1}(1-u)^{b-1}\,du$ = the incomplete beta function

(see DLMF, 2024)

In addition we have

$F_{Beta}(r;a,b) = B(r;a,b)/B(a,b)$ where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ = the complete beta function, so

$B(r;a,b) = F_{Beta}(r;a,b) \cdot B(a,b)$

(see Weisstein, E., 2024a and 2024b)

Combining terms we have

[13] The (Gaussian) hypergeometric function appears in derivations of the distribution of individual correlations (see Muirhead, 1982, and Taraldsen, 2021), moments of the spectral distribution under some conditions (see Adams et al. 2018, and https://reference.wolfram.com/language/ref/MarchenkoPasturDistribution.html), and in the definition of positive definite functions (see Franca & Menegatto, 2022).

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

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

Recognizing that the complete Beta function is the inverse of the normalization factor of c(k) for these values, their product equals 1 and cancels, as do the two cosine terms, and we obtain the following signed beta cdf:

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

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

And now, with this straightforward, fully analytic, non-recursive cdf, we can obtain a straightforward, fully analytic quantile function of the angle distribution:

$$\text{Let } p = \Pr(x \geq X). \text{ Then for } x < \frac{\pi}{2},$$

$$p = \frac{1}{2} - \left(\frac{1}{2}\right) \cdot F_{Beta}\left[\cos^2(x); \frac{1}{2}, \frac{1+k}{2}\right]$$

$$-2p = -1 + F_{Beta}\left[\cos^2(x); \frac{1}{2}, \frac{1+k}{2}\right]$$

$$1 - 2p = F_{Beta}\left[\cos^2(x); \frac{1}{2}, \frac{1+k}{2}\right]$$

$$F_{Beta}^{-1}\left(1 - 2p; \frac{1}{2}, \frac{1+k}{2}\right) = \cos^2(x)$$

$$\sqrt{F_{Beta}^{-1}\left(1 - 2p; \frac{1}{2}, \frac{1+k}{2}\right)} = \cos(x)$$

$$\arccos\left(\sqrt{F_{Beta}^{-1}\left(1 - 2p; \frac{1}{2}, \frac{1+k}{2}\right)}\right) = x$$

(Note that arcos is arc-cosine, the inverse of the cosine function.)

We must reflect the symmetric angle density for p≥0.5, so we have

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

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

Importantly, although often ignored in the sampling literature (see Makalic and Schmidt, 2018), note that properly adjusting for sample size, n, and degrees of freedom gives $k \leftarrow k+n-\#cols-2$, so consequently, $k = n - column\# \ -2$.

So now from (12) above we have for the angles distribution, under the Gaussian identity matrix, for the first time together, the pdf, cdf, and quantile function:

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

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

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

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

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

Apparently the first (and only other) presentation of this quantile function result comes from an anonymous blog post in March, 2018, although it was obtained via a different derivation, which serves to further validate the result.[14]

[14] See Xi'an, March, 2018 (https://stats.stackexchange.com/questions/331253/draw-n-dimensional-uniform-sample-from-a-unit-n-1-sphere-defined-by-n-1-dime/331850#331850
and https://xianblog.wordpress.com/2018/03/08/uniform-on-the-sphere-or-not/). In the interest of proper attribution, a reference on the website to the book "The Bayesian Choice" hints that the Xi'an pseudonym is Christian Robert, a professor of Statistics at Université Paris Dauphine (PSL), Paris, France, since 2000 (https://stats.stackexchange.com/users/7224/xian).

The above (12) now provides a fully analytic solution,[15] and in fact is so straightforward as to be readily implemented in a spreadsheet, and one is provided for download via the link below.

http://www.datamineit.com/JD%20Opdyke--The%20Correlation%20Matrix-Analytically%20Derived%20Inference%20Under%20the%20Gaussian%20Identity%20Matrix--02-18-24.xlsx

So contrary to the assertions of Makalic and Schmidt (2018), the straightforward approach of inverse transform sampling CAN be used in this setting, for this narrow case, to efficiently sample the correlation matrix. And in fact, this is the most efficient way to sample it. Roman (2023) has compared Makalic and Schmidt (2018) to the above method (defined in Opdyke, 2022, 2023, and 2024a) and obtained over 30% decrease in runtime when using Opdyke (2022, 2023, and 2024a).

But sampling arguably is the less important of our two goals, because with a fully analytic finite-sample distribution, we can define, exactly for a given sample size, the p-value of a given cell, and the confidence interval of a given cell. The one-sided p-value simply is the CDF value for the lower tail, or [1 – (CDF value)] for the upper tail (13), and due to this pdf's symmetry, the two-sided p-value is simply two times either one-sided value. Correspondingly, the confidence interval for the critical value alpha is based on the quantile function as in (14)

(13) one-sided p-value = $F_X\left(x;k\right)$ or $1-F_X\left(x;k\right)$ where k = n – column# – 2;

two-sided p-value = 2 x one-sided p-value

(14) $F^{-1}\left(\alpha/2;k\right)$ and $F^{-1}\left(1-\alpha/2;k\right)$ where, for a 95% confidence interval for example, α = 0.05

Notably, because the angles distributions are independent, the density of the entire matrix is simply the product of the densities of all the cells. This means we can readily define the p-value and confidence intervals of the entire matrix such that they are analytically consistent with those of the cells, because they are determined based directly on the cell level p-values and confidence intervals, respectively, as shown below.

Matrix-level p-values and Confidence Intervals

As mentioned above, a key characteristic of the angles is that they are independent random variables, which makes defining their multivariate distribution straightforward: it is simply the product of all the angles' pdf's. But what does this mean for the p-value and confidence intervals for the entire matrix? Given the null hypothesis (i.e. the Gaussian identity matrix up to this point, although these results also apply to the more general case), the (2-sided) p-value of the entire matrix is simply one minus the

[15] Note that we use the term 'analytic' as opposed to 'closed-form' because we are unaware of a closed-form algorithm for the inverse cdf of the beta distribution (see Sharma and Chakrabarty, 2017, and Askitis, 2017). However, for all practical purposes this is essentially a semantic distinction since this quantile function is hard-coded into all major statistical / econometric / mathematical programming languages.

probability of no false positives, which is the definition of controlling the family-wise error rate (FWER) of the matrix (15).[16]

(15) $$\text{matrix (2-sided) } pvalue = \left[1 - \prod_{i=1}^{p(p-1)/2} \left(1 - p\text{-}value_i\right)\right]$$ where $p\text{-}value_i$ is the 2-sided p-value.

Again, because the cell-level distributions are independent, their p-values are independent, and otherwise statistically more powerful approaches for calculating the FWER that rely on, for example, resampling methods (see Westfall and Young, 1993, and Romano and Wolf, 2016), do not apply here. In other words, they provide no power gain over (15) because under independence, there is no dependence structure for them to exploit. So the straightforward calculation above in (15) is, by definition, the most powerful for FWER control.

Similarly, calculation of the confidence interval for the entire matrix (16) is essentially the same as that of the p-value, but of course it is divided in half to account for each tail, and the root of the critical values is taken, rather than the product. Otherwise, the calculations are identical to obtain the critical alphas for these 'simultaneous confidence intervals.'

(16) $$\alpha_{crit-simult-LOW} = \left(1 - \left[1 - \alpha/2\right]^{\left(1/\left\lfloor p(p-1)/2 \right\rfloor\right)}\right)$$ and $$\alpha_{crit-simult-HIGH} = 1 - \alpha_{crit-simult-LOW}$$

These critical alphas, when inserted in the quantile function (12), provide the two correlation matrices that define and capture, say, (1-alpha)=(1-0.05)=95% of randomly sampled matrices under the null hypothesis, which in this case is the identity matrix. Again, it is the independence of the angles that makes these simultaneous confidence intervals very straightforward to calculate.

Importantly, again note that because we derived the quantile (inverse cdf) function in (12) above, we can go in either direction regarding these results: we can specify a correlation matrix and, under the null hypothesis of the identity matrix, obtain its p-values, both for the individual cells and the entire matrix, simultaneously. We also can specify a matrix of cdf values and obtain its corresponding correlation matrix, which is extremely useful and straightforward when constructing reverse scenarios. Finally, we can use simultaneous confidence intervals to obtain the two correlation matrices that form the matrix-level confidence interval.

Note that all these calculations are included in the downloadable spreadsheet, with visible formulae corresponding to each step of these calculations for full transparency. In the next section below I expand these results for Pearson's to apply to all data conditions, and all values of the null hypothesis (i.e. any values for the matrix, not just the identity matrix).

[16] Note that this approach has been used in the literature for addressing very closely related problems (see Fang et al., 2024).

PEARSON'S CORRELATION, REAL-WORLD FINANCIAL DATA, ANY MATRIX

Currently, the extant literature does not provide analytic forms for the angles distributions under general conditions. Deriving these appears to be a non-trivial problem. Spectral (eigenvalue) distributions, which many researchers turn to in this setting, have been shown to vary dramatically when data is characterized by different degrees of heavy-tailedness (see Burda et al., 2004, Burda et al., 2006, Akemann et al., 2009; Abul-Magd et al., 2009, Bouchaud & Potters, 2015, Martin & Mahoney, 2018; and Opdyke, 2024a), as well as by different degrees of serial correlation (see Burda et al., 2004, 2011, and Opdyke, 2024a), and the literature provides no general analytic form under general, real-world financial data conditions – certainly nothing that is analogous to convergence to the Marchenko-Pastur distribution under iid independence (Marchenko and Pastur, 1967).[17] If angles distributions are of similar complexity, deriving their general analytic form under general conditions, if possible, currently appears to be a non-trivial, unsolved problem.

However, this need not be a showstopper for our purposes, in part because angles distributions have many characteristics that make them useful here generally, and more useful specifically than spectral distributions in this setting, by multiple criteria, including structurally, empirically, and distributionally.

Structurally: Aggregation level becomes relevant and important here. For a given correlation matrix *R* there are many more angles distributions than there are spectral distributions (i.e. p(p-1)/2 cells vs p eigenvalues, a factor of (p-1)/2 more). As a matrix approaches singularity (non-positive definiteness (NPD)), which arguably is the rule rather than the exception for non-small investment portfolios, a much smaller *proportion* of angles distributions will approach degeneracy (i.e. maximum/minimum values of **π** and zero) than is true for eigenvalue distributions (where more values will wrongly fall below zero).

Consequently, the overall construction of the correlation matrix via $R = BB^T$ generally will remain much more stable than one based on an eigen-decomposition of $R = V\Lambda V^{-1}$ where *V* is a matrix with column eigenvectors and Λ is a diagonal matrix of the corresponding eigenvalues.

Empirically: If an angle distribution approaches degeneracy, most of its values typically will approach 0 or **π**. But the relevant trigonometric functions (sin, cos) of these values are stable, and will simply approach -1, 0, or 1. This makes $R = BB^T$ a relatively stable calculation empirically, even if it produces an *R* that is approaching NPD. In contrast, eigenvalue/vector estimations are more numerically involved compared to the application of simple trigonometric functions, and this, combined with the fact that they have no upper bound (in the general case), makes their computation comparatively less numerically stable as matrices approach NPD.

[17] Note that some exceptions to convergence to this celebrated distribution do exist (see Li and Yao (2018), Hisakado and Kaneko (2023), and Maltsev and Malysheva (2024) for examples).

Distributionally: As shown graphically below under challenging, real-world financial data conditions, the angles distributions are relatively "well behaved," both in the general sense and relative to spectral distributions. They are relatively smooth and typically unimodal, and clearly bounded on $\theta \in (0, \pi)$. Spectral distributions, based on the same data, very often are spikey[18] and highly multimodal, and their unboundedness (in the general case) translates into larger variances and less tail accuracy. Simply put, they typically are much more complex and challenging to estimate precisely and accurately compared to individual angles distributions for a given correlation matrix *R* under real-world financial data.

All of this adds up to a more robust and granular basis for inference and analysis when relying on angles distributions as opposed to spectral distributions. As discussed in more detail below, spectral distributions simply are at the wrong level of aggregation for these purposes: they remain at the (higher) level of the p assets of a portfolio – NOT at the granular level of the p(p-1)/2 pairwise associations of that portfolio, which is where the angles distributions (and correlations!) lie. Consequently, while potentially very useful for things like portfolio factor analysis, spectral analysis simply is too blunt a tool for our purposes here: we need to be able to make inferences and monitor processes and conduct (reverse) scenario analyses and customized stress tests on ALL aspects of the dependence structure measured by the correlation matrix, at the granular level at which it is defined. The specific need for this in scenario and reverse scenario analyses is covered in more detail below.

So given the useful characteristics of the angles distributions (on both a general basis and relative to the alternative of spectral distributions), not to mention the fact that they remain at the right level of aggregation for granular analysis of the correlation matrix, we are able to proceed WITHOUT their analytic derivation: rather, we can use a time-tested nonparametric approach, such as kernel estimation, to reliably define them. All this requires is a single simulation (say, N=10,000) based on the known or well-estimated correlation matrix, and the known or well-estimated data generating mechanism. Then, after translating all N simulated correlation matrices to N matrices of angles, we fit a kernel to each empirical angle distribution, i.e. the empirical distribution of each angle for each cell of the matrix. We now have not only the densities of all the individual angles, but also the multivariate density of the matrix, which is just the product of all the individual densities due to their independence. Note that this goes in both directions: we can perform 'look-ups' on the empirically defined distribution to obtain the cdf value(s) corresponding to particular angle value(s), or use cdf value(s) to 'look up' corresponding angle (quantile) value(s). The kernel fitting smooths this empirical density to all (continuous) values. This process is described step by step below.

1. Simulate samples (say, N=10k) based on the specified/known or well estimated correlation matrix and the specified/known or well estimated data generating mechanism.

[18] In fact, one of the most commonly encountered covariance (correlation) matrices under real world financial data conditions is the spiked matrix (see Johnstone, 2001), where one or few eigenvalues dominate and the majority of eigenvalues are close to zero, i.e. not reliably estimated. This further demonstrates that spectral approaches are far too limited and limiting to effectively solve this problem under real-world conditions.

2. Calculate the corresponding N correlation matrices, and their Cholesky factorizations, and transform each of these into a lower triangle matrix of angles (as described above in (6)).
3. Fit kernel densities to each of the p(p-1)/2 empirical angle distributions, each having N observations.
4. Generate samples based on the densities in 3.[19]
5. Convert the samples from 4. back to a re-parameterized Cholesky factorization, and then multiply by its transpose to obtain a set of N validly sampled correlation matrices (as described above in (7)). Positive definiteness is enforced automatically as the Cholesky factor places us on the **unit** hyper-hemisphere.

The distribution of correlation matrices from 5. is identical to that of 2., but after the kernel densities are fit once in 3., generating samples in 4. is orders of magnitude faster than relying on direct simulations in steps 1. and 2. And of course, using 3.-5. rather than 1. and 2. allows for correct probabilistic inference both at the cell level and at the matrix level, due to the independence of the angles distributions (remember the correlations themselves are NOT independent!). This reliance on the angles, and their subsequent transformation to correlations, allows us to isolate specifically the distribution of the entire correlation matrix, for probabilistic inference, without touching any other distributional aspect of the data, which is the point of the methodology. A cavalier 'bootstrap' of correlation matrices via direct data simulation fails at this objective, because the non-independence of the cells undermines the validity of any empirically-based inference. In other words, direct simulation does not preserve inferential capabilities, but simulation of the angles does.

So this framework is essentially identical to that for the specific case of the Gaussian identity matrix, with the only difference being it is based on nonparametrically defined, as opposed to parametrically defined, angles distributions. Before covering implementation details below, I show some examples of graphs of the angles distributions and the corresponding spectral distribution under challenging, simulated financial returns data. The multivariate returns distribution of the portfolio is generated based on the t-copula of Church (2012), with p=5 assets, varying degrees of heavy-tailedness (df=3, 4, 5, 6, 7), skewness (asym. parm.=1, 0.6, 0, -0.6, -1), non-stationarity (std. dev.=3σ, σ/3, σ; n/3 obs each), and serial correlation (AR1=-0.25, 0, 0.25, 0.50, 0.75), with a block correlation structure shown in (17) below and n=126 observations.[20] The spectral distribution is compared against Marchenko-Pastur as a baseline.

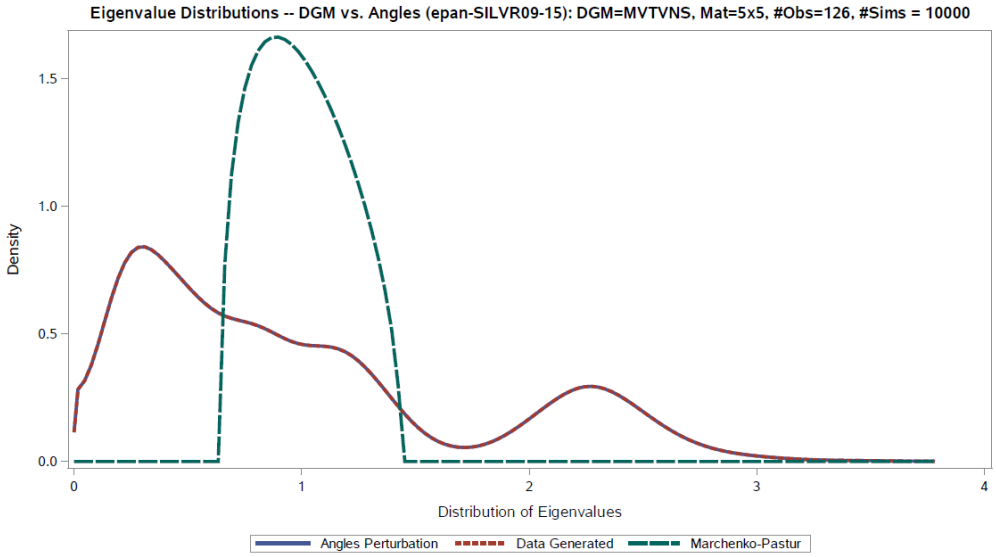
(17)

1	-0.3	-0.3	0.2	0.2
-0.3	1	-0.3	0.2	0.2
-0.3	-0.3	1	0.2	0.2
0.2	0.2	0.2	1	0.7
0.2	0.2	0.2	0.7	1

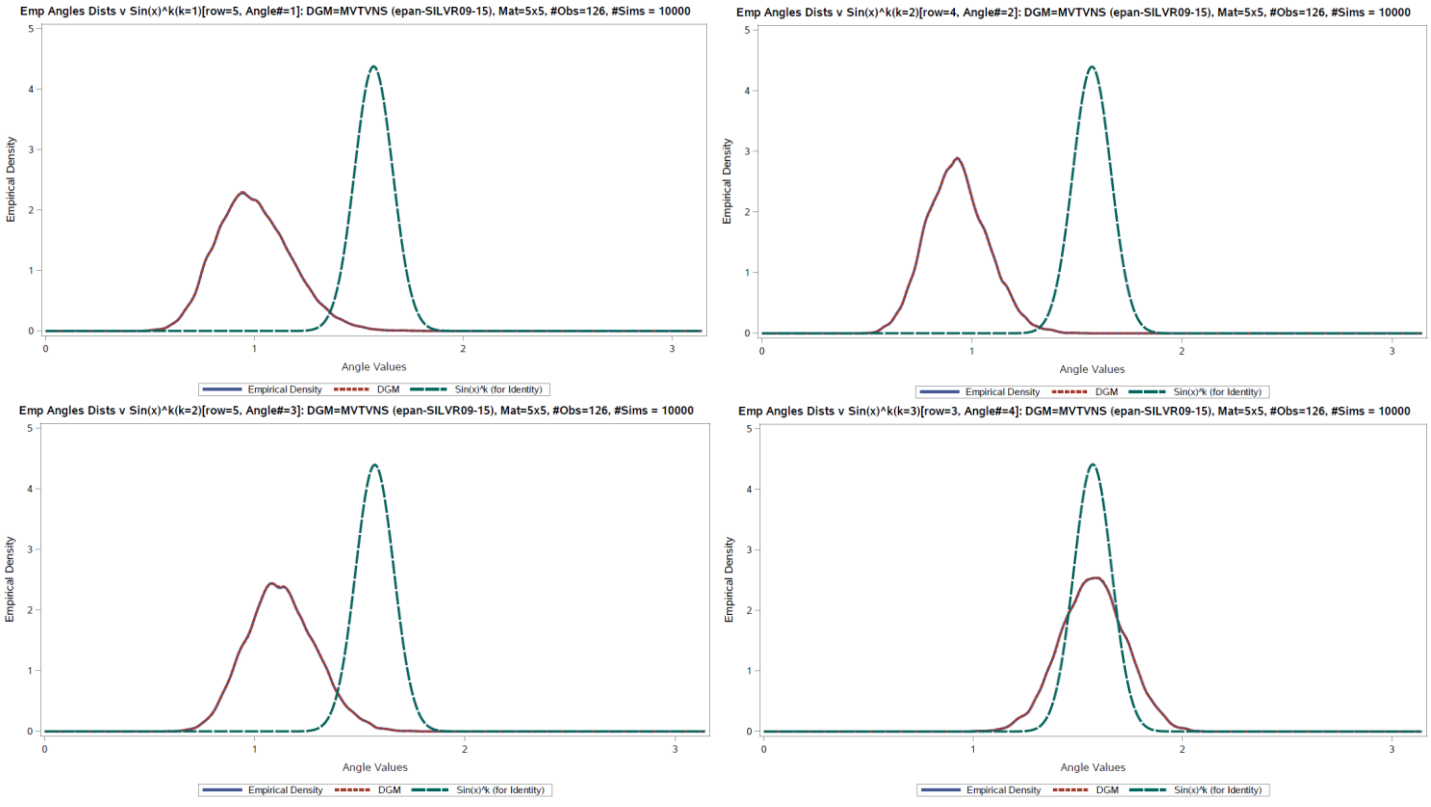
[19] Algorithms for sample generation based on commonly used kernels (e.g. the Gaussian and Epanechnikov) are widely known. An example of the latter is simply the median of three uniform random variates (see Qin and Wei-Min, 2024).

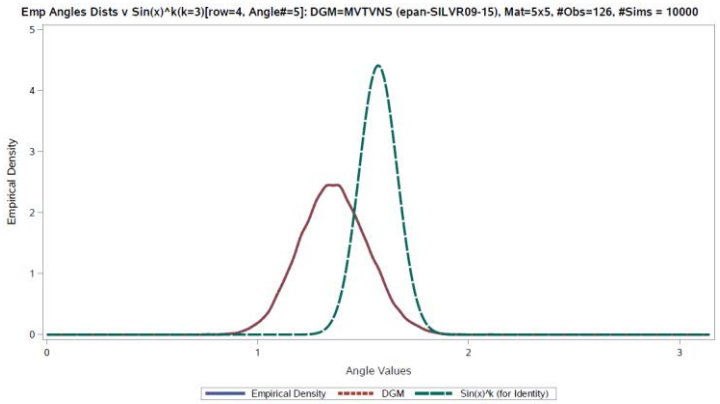
[20] Note that this is only approximately Church's (2012) copula, which incorporates varying degrees of freedom (heavy-tailedness) and asymmetry, because I also impose serial correlation and non-stationarity on the data (and then empirically rescale the marginal densities).

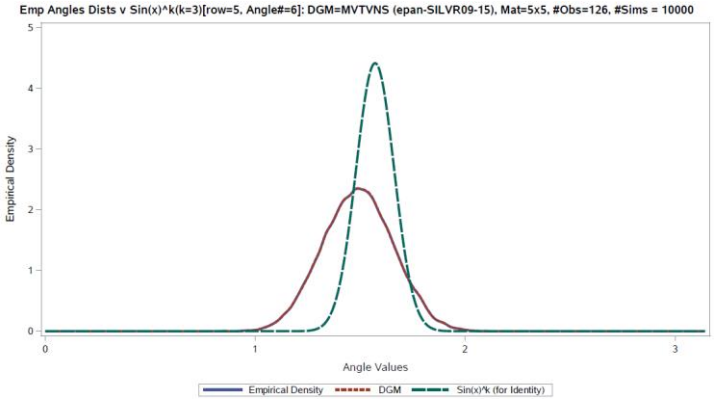
Graph 1: Spectral Distribution – Angles/Kernel Perturbation v Data Simulations v Marchenko Pastur

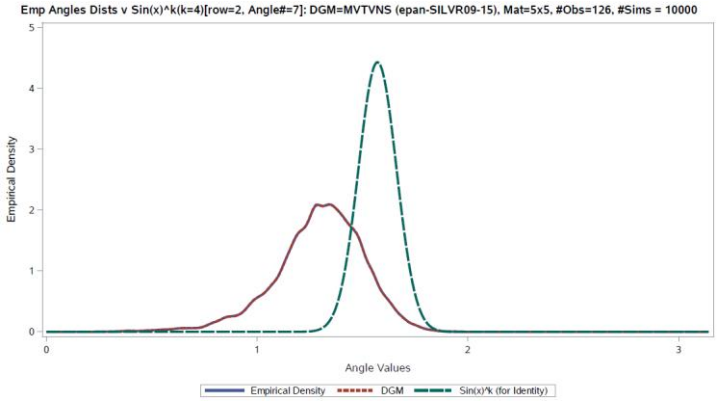


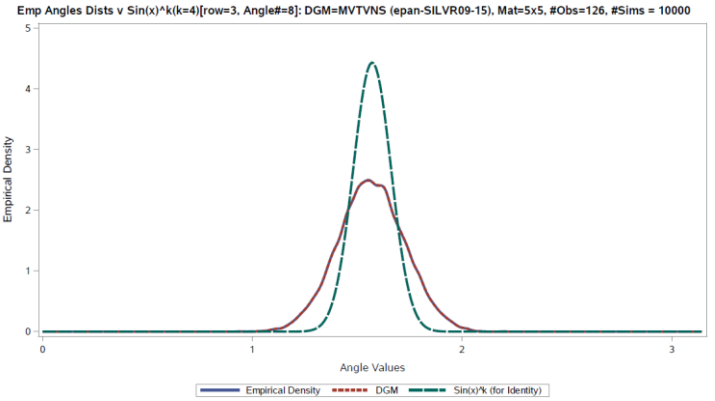
Graphs 2-10: Angles Distributions – Angles/Kernel Perturbation v Data Simulations v Independence

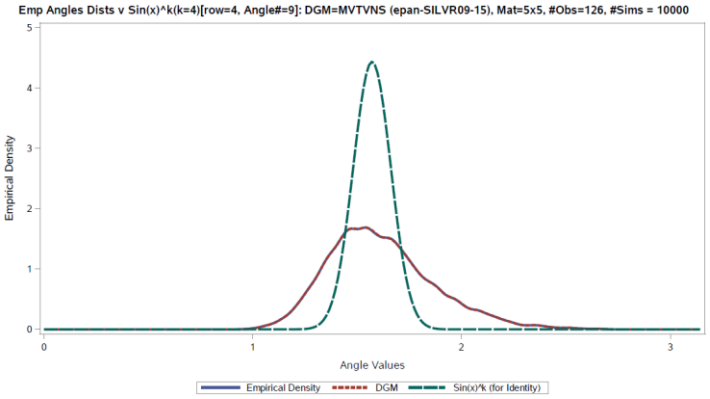


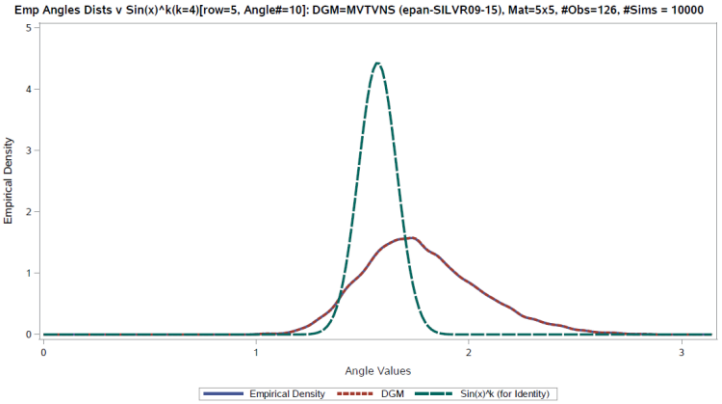












Several points are worth noting and reemphasizing from these graphs. First, the graphs of the angles distributions contain three densities: A. one based on angles perturbation (i.e. sampling from the fitted kernel) as described above in steps 3.-5., B. one based on direct data simulations (steps 1.-2.), and C. the analytical density under the (Gaussian) identity matrix as a comparative baseline. Note that the only reason I include B. is to demonstrate the validity of A., and as expected, the angles distributions from A. and B. are empirically identical (with A. being orders of magnitude faster and more computationally efficient). The spectral distributions based on the samples generated in both A. and B. also are identical, as are a wide range of additional aggregated metrics not presented herein (e.g. various norms, VaR-based economic capital, and 'generalized entropy' as described below). This empirically validates that the angles-perturbation approach is an efficient and correct method for isolating and generating the density of the correlation matrix, and unlike steps 1. and 2., one that preserves inferential capabilities. In other

words, these results empirically validate that the angles contain all extant information regarding dependence structure here (see Fernandez-Duren & Gregorio-Dominguez, 2023, and Zhang & Songshan, 2023, as well as Opdyke, 2024a).

Second, note again that a nonparametric approach works in practice here at least in part because the angles distributions are 'well behaved.' Since they are relatively smooth and unimodal and well bounded, N=10,000 simulations almost always suffice to provide a precise and accurate measure of their densities. Poorly behaved distributions that are very spikey, highly multi-modal, and unbounded could require numbers of simulations orders of magnitude larger. If N=10,000,000 or even 1,000,000 for example, this approach could be computationally prohibitive in many cases for real-world-sized portfolios, which often exceed p=100 and p(p-1)/2=4,950 pairwise associations/cells.

Finally, as described above, note the multi-modal and unbounded nature of the spectral distribution for this portfolio compared to the angles distributions, where the biggest thing approaching an estimation challenge is a slight asymmetry. But this speaks only to estimation issues. More notable is the fact that on a cell-by-cell basis, the angles distributions deviate materially i. not only from central values of $\boldsymbol{\pi}/2$, and less dramatically from perfect symmetry, when compared to their (analytic) distributions under the (Gaussian) identity matrix, but also ii. from each other! Each angle's distribution can vary quite notably compared to the other angles' distributions, especially under smaller sample sizes. There simply is no way that a single spectral density, even if perfectly estimated, will be able to capture and reflect all the richness of dependence structure reflected here at the granular level of the pairwise cells, for any useful purposes, including cell-level attribution analyses, granular scenario and reverse scenario analyses, cell-level intervention 'what if' analyses, and customized stress testing, let alone precise and correct inference at either the cell level OR the matrix level. I now leave comparisons to spectral distributions behind[21] to cover implementation issues below.

Nonparametric Kernel Estimation

Due to the bounded nature of the angles distributions on $\theta \in (0, \pi)$, the kernel must be appropriately reflected at the boundary (see Silverman, 1986) via: $\text{if } \theta < 0 \text{ then } \theta \leftarrow -\theta; \text{if } \theta > \pi \text{ then } \theta \leftarrow (2\pi - \theta)$, which is asymptotically valid. As per the standard implementation, the kernel itself is defined as

$$f_h(\theta) = \frac{1}{N}\sum_{i=1}^{N} K_h(\theta - \theta_i) = \frac{1}{hN}\sum_{i=1}^{N} K_h\left(\left[\theta - \theta_i\right]/h\right)$$ with

[21] Continued reliance on spectral approaches for this specific problem brings to mind a quotation attributed to John M. Keynes: "the difficulty lies not so much in developing new ideas as in escaping from old ones."

Gaussian: $K(\theta) = \left(1/\sqrt{2\pi}\right) \cdot e^{-\theta^2/2}$, Epanechnikov: $K(\theta) = (3/4) \cdot \left(1-\theta^2\right),\ |\theta| \le 1$.

Both the Gaussian and the Epanechnikov kernels have been tested extensively in this setting, along with three different bandwidth estimators, *h*, from Silverman (1986) and one from Hansen (2014), respectively:

$h = 1.06 \cdot \hat{\sigma} \cdot N^{-1/5}$, $h = 0.79 \cdot \text{IQR} \cdot N^{-1/5}$, $h = 0.9 \cdot \min\left(\text{IQR}/1.34, \hat{\sigma}\right) \cdot N^{-1/5}$, and

$h = 2.34 \cdot \hat{\sigma} \cdot N^{-1/5}$ for Epanechnkov only, where $\hat{\sigma} =$ sample standard deviation and

$\text{IQR} =$ sample interquartile range. As with virtually all kernel implementations, the choice of kernel matters less than the choice of bandwidth, although in this setting, across a broad range of data conditions and correlation values, the Epanechnikov kernel appears to perform slightly 'better' (i.e. with slightly less variance, thus providing slightly more statistical power) than the Gaussian, perhaps because its sharp bounds require reflection at the boundary less often than the Gaussian kernel. The bandwidth that appears to perform best across wide-ranging conditions is $h = 0.9 \cdot \min\left(\text{IQR}/1.34, \hat{\sigma}\right) \cdot N^{-1/5}$. Additionally, for larger matrices (e.g. p=100), bandwidths need to be tightened by multiplying *h* by a factor of 0.15. When there are many cells (e.g. for p=100, #cells=p(p-1)/2=4,950) this tightening avoids a slight drift in metrics that are aggregated across all the cells (e.g. correlation matrix norms, spectral distributions, and LNP (a type of 'generalized entropy' defined below)). Multiplying by this factor for smaller matrices does not adversely affect the density estimation in any way, so this factor always is used. For matrices much larger than p=100, a further tightening of this factor may be required, and this is readily determined by empirically comparing the distributions of these aggregated metrics under i. direct data simulation vs. ii. NAbC's kernel-based sampling.

Once the kernels have been estimated and the angles distributions generated by perturbing/sampling based on those kernels, the p-values and confidence intervals for both the individual correlation cells and the entire correlation matrix are the same as those derived for the Gaussian identity matrix. The only difference, aside from their now-nonparametric basis, is that the angles distributions are no longer symmetric by definition, as is true under the (Gaussian) identity matrix. This can be seen in the graphs of the angles distributions provided above. The p-value calculation, however, remains very straightforward, and it requires just a bit of care to properly account for asymmetry. The one-sided p-value remains simply (13):

(13) one-sided p-value = $F_X(x;k)$ or $1 - F_X(x;k)$ for lower and upper tails, respectively,
where k = n – column# – 2

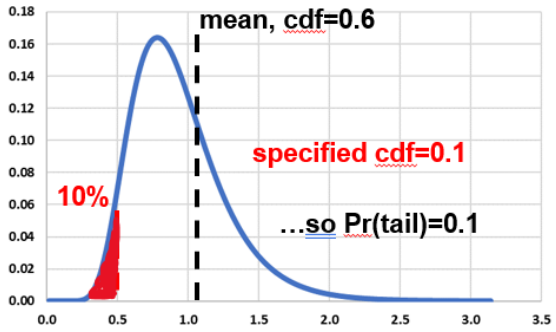
However, due to possible (probable) asymmetry, the two-sided p-value is different, requiring first the calculation of the empirical mean correlation matrix from the simulations in step 2. of the five kernel-

based sampling steps above. This mean correlation matrix is then translated into a matrix of angles, and we obtain the empirical cdf values corresponding to these "mean angles" with a "look-up" on the entire angles distributions generated in step 4. These cdf's will be close to 0.5 when the angles distributions are close to symmetry, and they will deviate from 0.5 under asymmetry. The two-sided p-values are based on the difference between the cdf values of each of the angles of the specified correlation matrix being 'tested,' and those of the "mean angles." Specifically, the two-sided p-values are the sum of the probability in the tails BEYOND this difference.[22] Formulaically this is simply (18):

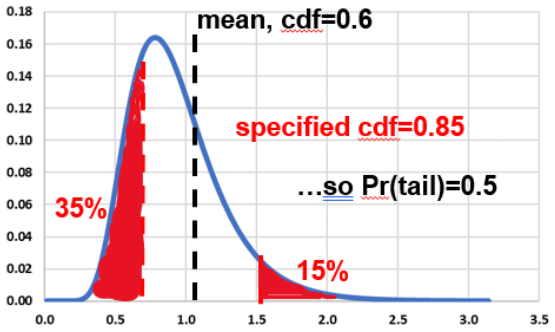
(18) two-sided p-value = max[0, Mcdf – d] + max[0, 1 – (Mcdf + d)], where
d = abs(Mcdf – cdf), Mcdf = mean angle cdf, cdf = cdf of specified angle

This usually results in summing both tails, but under notable asymmetry, sometimes only one tail is used. Below is a graphical example of both cases, where the cdf of the "mean angle" is 0.6 and the cdf of the relevant angle in the specified correlation matrix (i.e. the correlation matrix for which we are obtaining p-values, confidence intervals, etc.) is cdf=0.1 in the single-tail case (Graph 11) and cdf=0.85 in the double-tail case (Graph 12). In the statistical sense, however, both cases remain two-sided p-values.

Graph 11: p-value for a single specified (more) extreme angle cdf



Graph 12: p-value for a single specified non-extreme angle cdf



Note that while cdf=0.1 is hardly more 'extreme' than cdf=0.85 in absolute terms, relative to the mean angle cdf=0.6, it is twice as 'extreme,' i.e. twice as far probabilistically from the mean cdf=0.6, with an absolute difference of 0.5 for Graph 11, and 0.25 for Graph 12. Moreover, a value as extreme as the case of Graph 11 is associated with only 1/5 the probability of being observed compared to that of Graph 12 (compare the red shaded areas). This example demonstrates why asymmetry must be properly taken into account in this setting, but the two-sided p-value still remains a very straightforward calculation, and the "mean angles" matrix is used for additional, important purposes below, as discussed in the Scenarios section.

Cell-level confidence intervals still are simply calculated as in (14), which automatically takes asymmetry into account. This is identical to the same calculation under the (Gaussian) identity matrix. The matrix-level p-value, again, is simply one minus the probability of no false positives (15), which is the

[22] So this difference is 0.5 for Graph 11 and 0.25 for Graph 12.

definition of controlling the family-wise error rate (FWER) of the matrix. Also, just as under the (Gaussian) identity matrix, calculation of the confidence interval for the entire matrix remains (16) as previously.

Importantly, again note that we can go in either direction regarding these results: we can specify a correlation matrix and, under the null hypothesis of the specified correlation matrix, obtain the p-values of an observed matrix, both for the individual cells and the entire matrix, simultaneously. We also have the matrix-level quantile function: we can specify a matrix of cdf values and obtain its corresponding, unique correlation matrix, which can be extremely useful and straightforward when constructing reverse scenarios. Finally, we can use simultaneous confidence intervals to obtain the two correlation matrices that form the matrix level confidence interval. An example with all these results is shown in the "One Example" section below, but first I discuss the scenario-restricted case.

GRANULAR, FULLY FLEXIBLE SCENARIOS, REVERSE SCENARIOS, & CUSTOMIZED STRESS TESTS

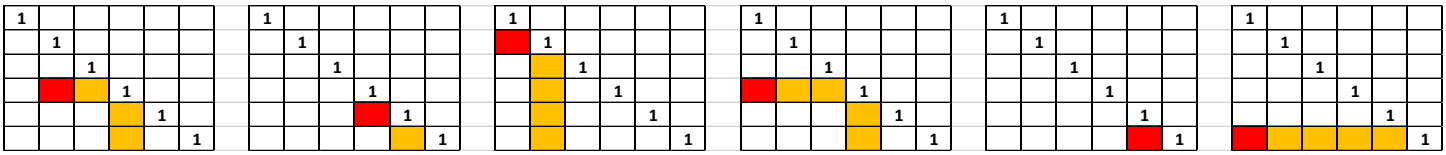
I have taken a very granular, 'bottom up' approach to defining the finite-sample distribution of the correlation matrix here, based on the distributions of the individual correlation cells. In addition to analytical consistency, this provides a flexibility that other approaches, such as those based on the spectrum of the dependence measure's matrix, cannot provide, because with only p eigenvalues, they simply are at the wrong level of aggregation to flexibly vary (or freeze) individual cells, as well as specific combinations of the p(p-1)/2 cells for different scenarios. Correlation (dependence) matrices under a tech market bubble (2000) vs those under a housing bubble (2008) vs those under Covid (2020) will change very different individual cells, and very different combinations of cells, in very different ways, often in terms of both direction and magnitude, while leaving many cells strongly affected under one upheaval completely unaffected under another. In other words, while correlation 'breakdowns' will occur under all of these extreme conditions, the granular nature of pairwise association matrices ensures that the fundamentally different nature of these breakdowns will be captured and reflected empirically in all related analyses. The only way to flexibly and realistically model this is at the most granular level – that of the individual correlation cells.

Fortunately, when using NAbC, several results allow for this. First, 1. independence of the angles distributions allows us to vary individual cells. Second, 2. 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, based on 1. and 2., 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.

To explain 3., I focus only on the lower triangle of the correlation matrices below in Graphs 13-15, since the upper triangle is just its reflection. Note again that using NAbC, we only perturb angles. We never perturb the correlation values directly. We must always convert to angles, perturb the angle values, and then translate back to correlation values. In doing so, when multiplying the Cholesky factor by its

transpose, $R = BB^T$, changing a given angle cell in matrix B will affect other cells, but only those cells to the right of it in the same row, and those below the diagonal of the corresponding column, as shown graphically for several examples in Graph 13 below.[23]

GRAPH 13: Mechanics of Matrix Multiplication



This means that we can simply reorder the matrix so that the targeted cells we want to vary all end up in the rightmost triangle of the lower triangle, according to the fill order in Graph 14 below.

GRAPH 14: Rightmost Triangle Fill Order

Rightmost Triangle Fill Order

11					
12	7				
13	8	4			
14	9	5	2		
15	10	6	3	1	

If we only change in matrix B the angle values of cells 1, 2, and 3 above, no other cells in the correlation matrix R will be affected, simply by virtue of the mechanics of matrix multiplication from $R = BB^T$. Below I show another example. Reorder the correlation matrix so that rows 1-6 are now 6-1 and columns 1-6 are now 6-1, so that the original cells 1,2 and 1,3 and 2,3 and 4,3 are now in the rightmost triangle of the lower triangular matrix, in the fill order shown above.

GRAPH 15: Example of Mechanics of Matrix Multiplication Applied to Rightmost Triangle Fill Order

Determine Targeted Change Cells

1,2					
1,3	2,3				
		4,3			

Reorder Rows/Cols to Fill Rightmost Triangle with Targets According to Fill Order

11					
12	7				
13	8	4			
14	9	5	2		
15	10	6	3	1	

Changes in Corresponding Angles Cells ONLY change Same in Resorted Matrix

11					
12	7				
13	8	4,3			
14	9	5	2,3		
15	10	6	1,3	1,2	

[23] Note that not all of these (orange) cells will necessarily change if values of zero are involved, but none OTHER than these (orange) cells CAN change when only the red cell changes.

Changes to the corresponding cells in the angles matrix B (the orange cells) will only change these same cells , after $R = BB^T$, in the resulting correlation matrix, leaving the rest unaffected. Note that the green cells to be targeted for change do not even have to be contiguous, nor do they have to completely 'fill' the rightmost (orange) triangle (note that cells 5 and 6 are not targeted): they only must fill the rightmost triangle according to the order of the middle matrix above. Note also that the "rightmost triangle" rule is nested/hierarchical: if I wanted to perform 'what if' analyses on only one of those cells (e.g. cell "1,2") without changing the other three, I order the original correlation matrix to place that cell as the 'first' in the lower triangle of the B matrix, as shown. Then, subsequent changes to it will not affect the other (orange) cells. In contrast, changes to cell "4,3" will affect the values of the other orange cells. Readers are encouraged to test this in the interactive spreadsheet (url link provided above).

So we can exploit these four simultaneous conditions – 1. independence of the angles distributions; 2. (correlation) distribution invariance to row and column order; 3. the mechanics of matrix multiplication; and 4. the granular, cell-level geometry of NAbC – to obtain great flexibility in defining scenarios wherein some cells vary and some do not. No other approach allows this degree of flexibility, which is what is required for defining correlation/dependence matrices for use in realistic, plausible, and sometimes extreme stress market scenarios. This also greatly simplifies attribution analyses, isolating and making transparent the identification of effects due to specific pairwise associations, which is something spectral analyses cannot do in this setting.

The above allows for the specification of ANY scenario within the structure of the pairwise matrix. Note, however, that some scenarios can include combinations of cells which are forced to include (in the lower right triangle) one or a few cells not affected by the scenario. This is unavoidable due to the structure of the pairwise matrix: for example, in the matrix above, there are only p! (ie 5!=120) ways to sort the rows and columns, but there are [p(p1-)/2]! (ie 15!= 1,307,674,368,000) ways to sort the 15 cells. The matrix obviously cannot accommodate freely sorting the individual cells in this way because it breaks the structure of the matrix. Some scenarios, therefore, could conceivably be required to include for perturbation some few additional cells in the lower rightmost triangle that are not relevant to the scenario and otherwise should be held constant. Fortunately, in practice, especially with large matrices, this appears to be a relatively rare occurrence, and when it happens, the effects are identifiable so that materiality can be assessed. But dealing with these potential cases appears to be well worth the price of the unmatched flexibility that this approach provides,[24] not to mention the other advantages it maintains over more complex, strictly multivariate dependence structures. For usage with actual market data, the latter typically are more difficult to estimate with the same levels of precision, let alone to manipulate for

[24] Most of the related scenario literature perturbs scenario-based cells and simply ignores their (notable) effects on the rest of the matrix (which should remain 'frozen,' but doesn't), not to mention the effects of the rest of the matrix on the scenario-related cells, and euphemistically refers to the former as 'peripheral' correlations (see Ng et al. (2013) and Yu et al. (2014)).

purposes of intervention or mitigation. In contrast, pairwise associations are directly identifiable, typically more easily and accurately estimated,[25] and interventions are more targeted and transparent.

To conclude this section I deal with one final implementation issue. When the matrix is scenario-restricted, and we only perturb a subset of the matrix while keeping the remaining cells fixed, what values do we use for those 'frozen' cells? This is where the mean angles matrix, used for calculating the two-sided p-values in the previous section, comes into play. When the matrix angles are sampled using the fitted kernel densities, a sample is drawn from the entire matrix, and if it is scenario restricted, the sampled values for those cells that are 'frozen' are simply overwritten with their means. So after N=10,000 samples, all 10,000 values of the 'frozen' cells unaffected by the scenario will have the same mean value for that specific cell, and when translated via $R = BB^T$ back into correlation matrices, all the correlation values for those cells will be their mean correlation value. In other words, their values will not change, and will remain 'frozen,' based on a reasonably robust estimator of their true value (note that these values are not based on one estimated matrix, but rather the mean of N=10,000 matrices). The order of magnitude of empirical accuracy of these values is inversely related to the number of samples drawn, N. In the example in the "One Example" section below, we observe accuracy to the fourth decimal place for these frozen cells when N=25,000, as expected. Alternately, the values could be treated as truly known constants from the beginning, but it is more conservative (and realistic) to use estimates based on the mean of all the samples.

Finally, now, we are able to revisit the fact that all of the above findings are generalizable not only to ANY data conditions, but also to ANY dependence measure, as long as its pairwise matrix is symmetric positive definite. Again, this is due to the fact that the relationship between angles and correlation/dependence matrices holds under this condition, regardless of the dependence measure that has generated the matrix, as shown below.

BEYOND PEARSON'S WITH NAbC: ALL POSITIVE DEFINITE DEPENDENCE MEASURES

The only condition required for the relationships between angles and dependence measure values, as shown in (6) and (7) above, is the symmetric positive definiteness of the dependence measure. Because this approach uses the framework of all pairwise comparisons, measuring dependence on a bi-variate basis, the requirement of symmetric positive definiteness, more precisely, is the symmetric positive definiteness of the matrix of the dependence measure calculated on every pairwise association of the all the assets in the portfolio. This distinction is important to make as many dependence measures can be

[25] They also can be **estimated** rigorously, and with targeted precision and flexibility, with well-established methods such as vine copulas (see Czado and Nagler, 2022)). Ironically, however, when used for **inference or sampling** for this problem specifically, vine copulas and similar methods become extremely unwieldy and much more complex and less transparent than NAbC, not to mention not generalizable beyond Pearson's (see the vine and extended onion algorithms of Lewandowski et al. (2009), and the similar chordal sparsity method of Kurowicka (2014)).

calculated not only on a bi-variate basis, but also on a multivariate basis, such as Szekely's distance correlation (Szekely, Rizzo, and Bakirov, 2007) and variants of Chatterjee's correlation (see Huang et al., 2022, Gamboa et al., 2022, Fuchs, 2024, and Pascual-Marqui et al., 2024; see as well as Chatterjee, 2022 for a summary of the recent literature on multivariate measures). We keep to the framework of the all-pairwise matrix here for numerous reasons, including its tremendous flexibility, ease and directness of application, ease, if not increased power, in estimation, and ease and transparency in intervention and what-if analyses. But the main point here is that all references to positive definiteness herein refer to the framework of the all-pairwise matrix.

This positive definiteness (numerical issues aside) has been long proven for the "the big three," that is, for the three most widely used dependence measures – Pearson's rho, Kendall's tau, and Spearman's rho (see Sabato et al., 2007). The values of these measures all range from –1 to 1,[26] but many other measures range from 0 to 1. These include Szekely's, Lancaster's, the Tail Dependence Matrix, Chatterjee's and its many variants (see Gao and Li, 2024) and many others. Proving that these, too, are positive definite is very straightforward, and was done by Embrechts et. al. (2016) regarding the tail dependence matrix. Recall the definition of positive definiteness (for a matrix of dimension p):

$\text{if } x'Rx > 0 \text{ for all } x \in \mathbb{R}^p \setminus \mathbf{0}$, then R is positive definite.

Because all of the (0,1) dependence measures described above are defined by $0 \le R_{i,j} \le 1 \text{ for all } i \ne j \text{ and } R_{i,i} = 1 \text{ and } R_{i,j} = R_{j,i}$, $x'Rx$ can be written in quadratic form as

(18) $$x'Rx = \sum_{i=1}^{p} x^2 + 2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} R_{i,j}x_i x_j$$

As long as $0 < R_{i,j} < 1 \text{ for all } i \ne j$, that is, the coefficients on the cross terms (the second term of (18)) all remain BETWEEN 0 and 1, then

(19) $\sum_{i=1}^{p} x^2 + 2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} R_{i,j}x_i x_j > 0$ and so $x'Rx > 0$, always, and so R is positive definite.

In the p = 2 case, for example, R is positive definite if $R_{1,1} > 0 \text{ and } \left(R_{1,1}R_{2,2} - R_{1,2}^2\right) > 0$, which is always true when $0 < R_{i,j} < 1 \text{ for all } i \ne j \text{ and } R_{i,i} = 1$. For the boundary cases, if $R_{i,j} = 0 \text{ for all } i \ne j$, R obviously remains positive definite as the first term of (18) always is greater than zero and the second term disappears; and if $R_{i,j} = 1 \text{ for all } i \ne j$ then R is positive semi-definite, although this case of perfect multivariate dependence is only textbook relevant. In practice, empirically, positive semi-definiteness only is relevant as a boundary condition, as it relates to empirical matrices that approach singularity.

[26] Of course, these are maximal bounds and many conditions exist under which actual bounds are tighter. For example, for Pearson's under the equicorrelation matrix E (all equal correlations), the lower bound is -1/(dim[E] -1) rather than -1.

Consequently, this means that all dependence measures with values ranging from 0 to 1 are, in practice, positive definite, and that NAbC can be applied to them to define their finite sample distributions. Empirical examples of this are shown in the next section.

Operationally, implementing NAbC on these (0, 1) measures is no different from implementing it on Pearson's or Kendall's or Spearman's; the (0, 1) instead of (–1, 1) range does not even change how we reflect at the boundary when fitting the nonparametric kernel. This is because specific cells of the Cholesky factor can validly be negative, making the assignation in the last line of the "Correlations to Angles" code in Table A above sometimes assign an angle value slightly above $\boldsymbol{\pi}/2$, even though $\boldsymbol{\pi}/2$ corresponds to a measure value of zero.[27] So this is a soft upper boundary in this case, even though the measure's range of (0,1) typically is not.[28] So when NAbC generates angle $\boldsymbol{\theta}$, we continue to reflect based on: $\text{if } \theta < 0 \text{ then } \theta \leftarrow -\theta; \text{if } \theta > \pi \text{ then } \theta \leftarrow \left(2\pi - \theta\right)$

since for measures with a (0,1) range, the upper bound of $\boldsymbol{\pi}$ will never be reached, and the lower bound of zero remains valid and hard. So NAbC applies in exactly the same way, for all of these positive definite dependence measures, whether their range of values is (–1, 1) or (0, 1).

Finally, again note that the condition of symmetric positive definiteness holds not only for all relevant dependence measures, as shown above, but also under all relevant real-world data conditions: that is, multivariate financial returns data whose marginal distributions typically are characterized by different degrees of asymmetry, heavy-tailedness, (non-)stationarity, and serial correlation. So this is a very weak and general condition, allowing for the extremely wide-ranging application of NAbC.

Spectral and Angles Distributions

I present below the angles distributions for some of the dependence measures discussed above, under simulated data reflecting challenging, real-world data conditions (see Opdyke, 2024a for the application of NAbC to a larger number of different data conditions). Briefly, as above, the multivariate returns distribution of the simulated portfolio is generated based on the t-copula of Church (2012), with p=5 assets, varying degrees of heavy-tailedness (df=3, 4, 5, 6, 7), skewness (asymmetry parameter=1, 0.6, 0, -0.6, -1), non-stationarity (standard deviation=3σ, σ/3, σ; 1/3 observations each), and serial correlation (AR1=-0.25, 0, 0.25, 0.50, 0.75), with a block correlation structure shown in (20) below and n=126

[27] Note that angle values (which range from zero to $\boldsymbol{\pi}$ on the hyper-hemisphere) decrease while dependence measure values increase, so a measure value of -1 corresponds to an angle value of $\boldsymbol{\pi}$, a measure value of zero corresponds to an angle value of $\boldsymbol{\pi}/2$, and a measure value of 1 corresponds to an angle value of zero (see Zhang et al., 2015 and Lu et al., 2019).

[28] On a related issue, note that Chatterjee's correlation, for example, is bounded by (0,1) only asymptotically, and finite sample results can exceed these bounds. However, when applying NAbC to this and other measures in hundreds of thousands of data simulations under widely varying conditions, as an empirical matter such finite sample exceedences never caused NAbC's angles distributions to deviate from those of direct data simulations, nor made empirical matrices not positive definite.

observations, for half a year of daily returns.[29]

(20)

1	-0.3	-0.3	0.2	0.2
-0.3	1	-0.3	0.2	0.2
-0.3	-0.3	1	0.2	0.2
0.2	0.2	0.2	1	0.7
0.2	0.2	0.2	0.7	1

For verification purposes only, I compare those angles distributions based on the data simulation directly against those based on NAbC's angle kernels, and in all cases the results are empirically indistinguishable. The same is true for the spectral distributions, which I also present below against the Marchenko-Pastur distribution as a(n independence) baseline (see Marchenko and Pastur, 1967). The empirical results yield both expected, and some additional interesting findings.

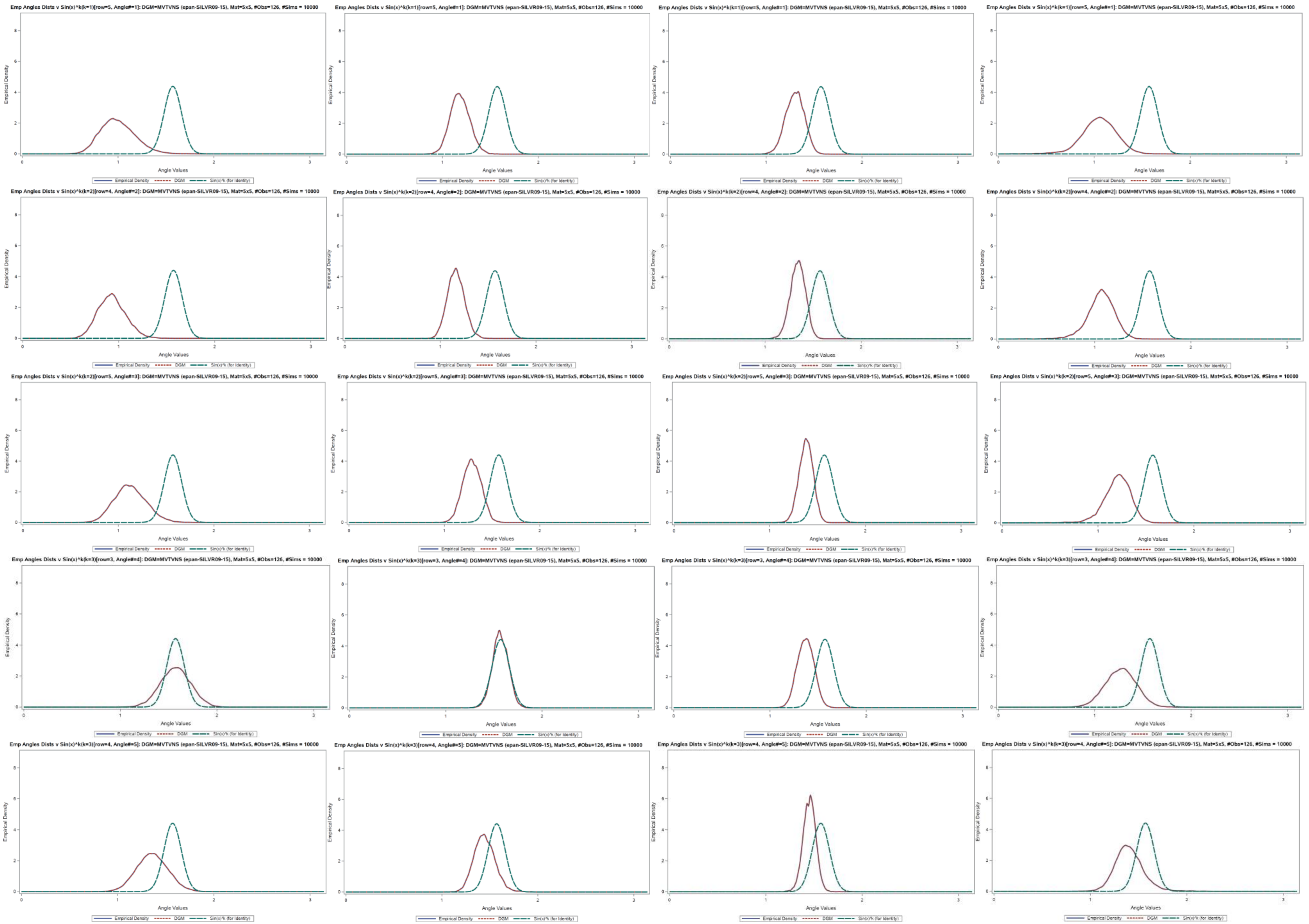
First, note that the spread, and the spread and shifts, of both the spectral and angles distributions, respectively, are larger for Pearson's than for Kendall's, which is consistent with the former's relative sensitivity to more extreme values under many conditions. The shifts and spread of both measures are much larger than those of Chatterjee,[30] although this is largely due to the fact that while Chatterjee is generally more powerful under dependence that is highly nonlinear and/or highly cyclical, it is less powerful under associations that are more monotonic, and the data conditions of this example fall more (but not entirely) into the latter category. The story changes a bit when we use the dependence measure suggested by Zhang (2023), which is essentially a maximum between Spearman's rho and Chatterjee's correlation, its objective being to obtain large, if not the maximum power under both types of dependence structures (i.e. strong monotonic dependence as well as highly nonlinear/cyclical dependence). This shows how readily NAbC can be applied to any (positive definite) dependence measure, and its utility for making cross-measure comparisons, all else equal, using the same, universally applicable method.

[29] Note again that this is only approximately Church's (2012) copula, which incorporates varying degrees of freedom (heavy-tailedness) and asymmetry, because I also impose serial correlation and non-stationarity on the data (and then empirically rescale the marginal densities).

[30] The symmetric version of Chatterjee's correlation coefficient is used here (see Chatterjee, 2021), with the finite sample bias correction proposed by Dalitz et. al., 2024.

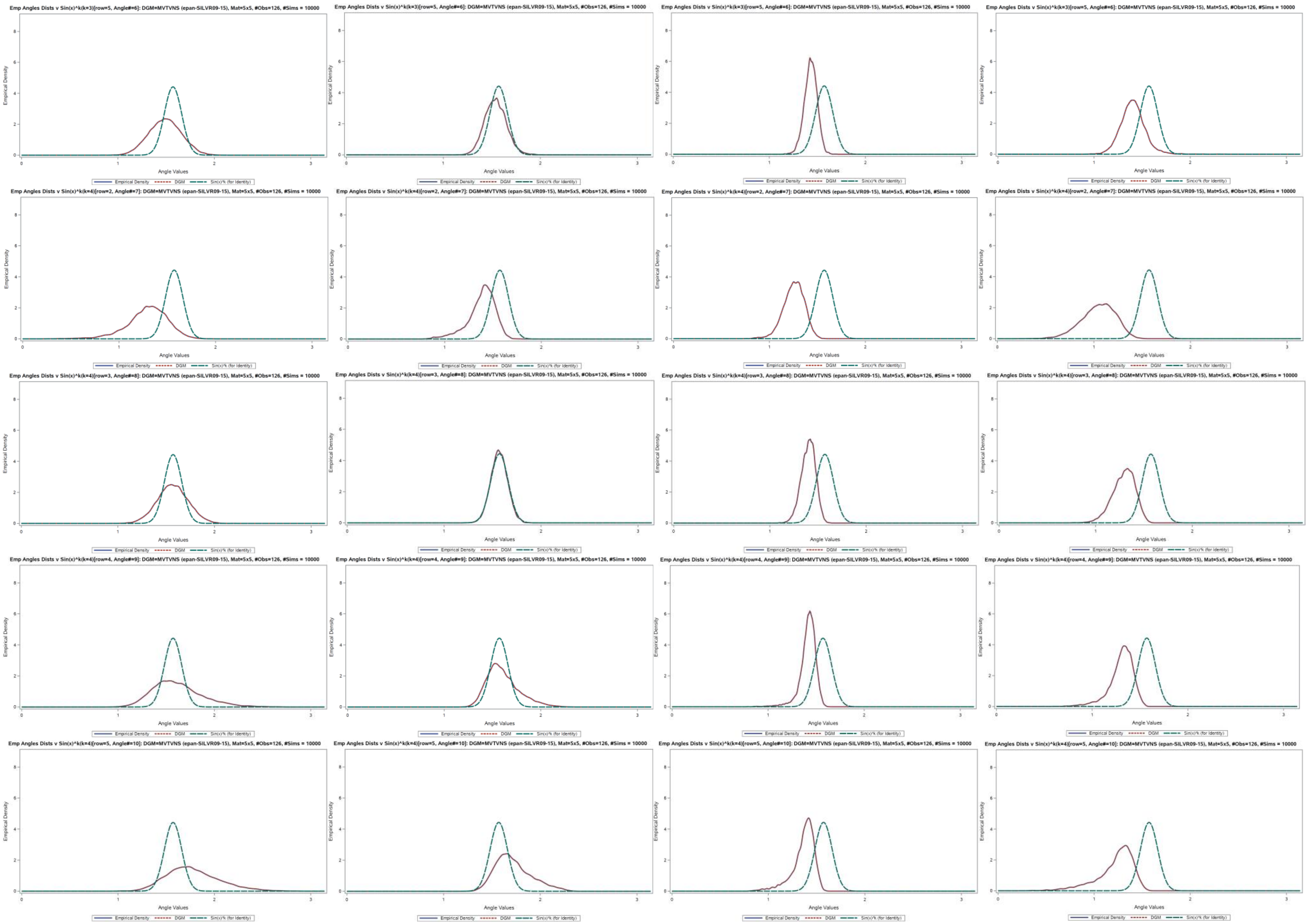
Graph 16a: Angles Distributions--NAbC Angles Kernel v Data Simulations v Identity Matrix

Pearson's Rho	Kendall's Tau	Chatterjee's	Spearman's Rho+Chatterjee



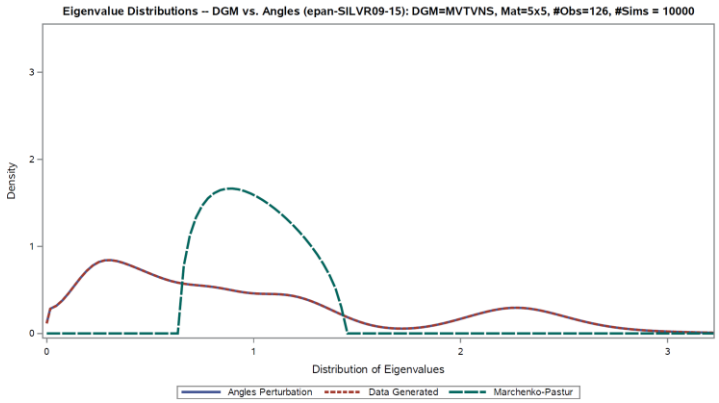
Graphs 16b: Angles Distributions--NAbC Angles Kernel v Data Simulations v Identity Matrix

Pearson's Rho | **Kendall's Tau** | **Chatterjee's** | **Spearman's Rho+Chatterjee**

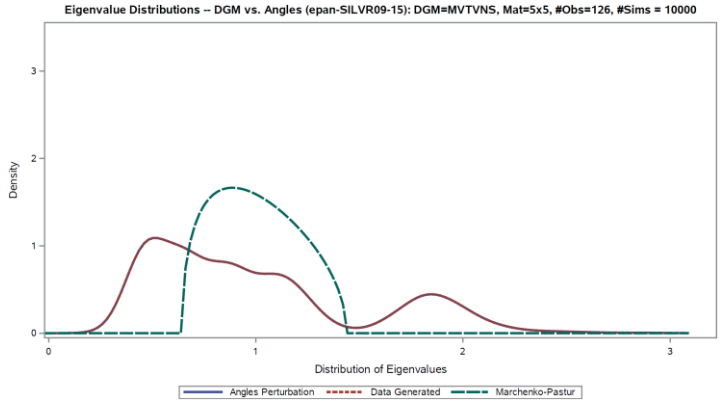


Graph 17: Spectral Distribution-NAbC Angles Kernel v Data Simulations v Marchenko Pastur

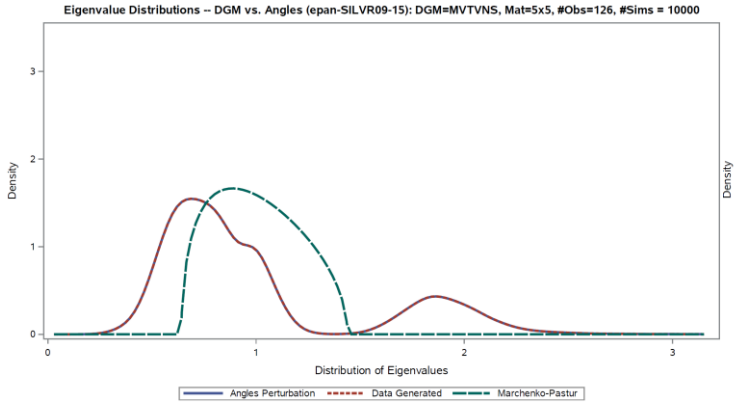
Pearson's Rho



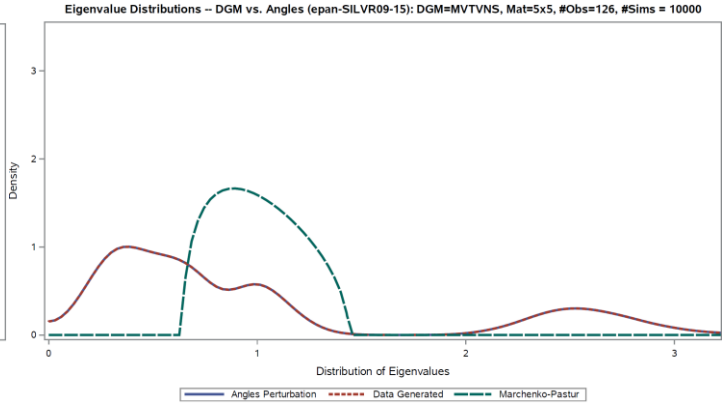
Kendall's Tau



Chatterjee's



Spearman's Rho+Chatterjee



Given all the mechanisms defined and verified above, I can now provide a complete example of the application of NAbC, checking all of the original seven objectives boxes, simultaneously.

One Example: Kendall's Tau p-values & Confidence Intervals, Unrestricted & Scenario-restricted

Here I provide an example of the complete application of NAbC, for Kendall's Tau, under two cases: unrestricted, and scenario-restricted. NAbC provides both p-values and confidence intervals, at both the cell level and matrix level. Solely for ease of replication, the data generating mechanism for this example is simply multivariate standard normal, with N=25k simulations and number of observations n = 160. The values of the matrix [A] are arbitrary, but correspond closely to those obtained when translating from a Pearson's matrix example used in one of my previous and shorter NAbC publications, using $\tau = (2/\pi)\arcsin(r)$ where r = Pearson's, which is generally valid under elliptical data (and which is one of the reasons I used multivariate Gaussian data here; see McNeil et. al., 2005).

UNRESTRICTED CASE: Given a specified or well-estimated correlation matrix [A], and its specified or well-estimated data generating mechanism:

[A]

1				
0.13	1			
-0.06	0.19	1		
0.19	-0.19	-0.06	1	
0.41	0.26	0.00	0.06	1

[B]

0.8				
0.7	0.8			
0.8	0.7	0.7		
0.7	0.8	0.8	0.7	

[C]

1				
0.3	1			
0.1	0.1	1		
0.05	-0.1	0.1	1	
0.5	0.25	0.2	0.15	1

Q1. **Confidence Intervals**: What are the two correlation matrices that correspond to the lower– and upper–bounds of the 95% confidence interval for [A]? What are, simultaneously, the individual 95% confidence intervals for each and every cell of [A]?

Q2. **Quantile Function**: What is the unique correlation matrix associated with [B], a matrix of cumulative distribution function values associated with the corresponding cells of [A]?

Q3. **p-values**: Under the null hypothesis that observed correlation matrix [C] was sampled from the data generating mechanism of [A], what is the p-value associated with [C]? And simultaneously, what are the individual p-values associated with each and every cell of [C]?

SCENARIO-RESTRICTED CASE: Under a specific scenario only selected pairwise correlation cells of [A] will vary (green), while the rest (red) are held constant, unaffected by the scenario (e.g. COVID). This is matrix [D].

[D]

1				
0.13	1			
-0.06	0.19	1		
0.19	-0.19	-0.06	1	
0.41	0.26	0.00	0.06	1

[E]

	0.8			
	0.8	0.8	0.7	

[F]

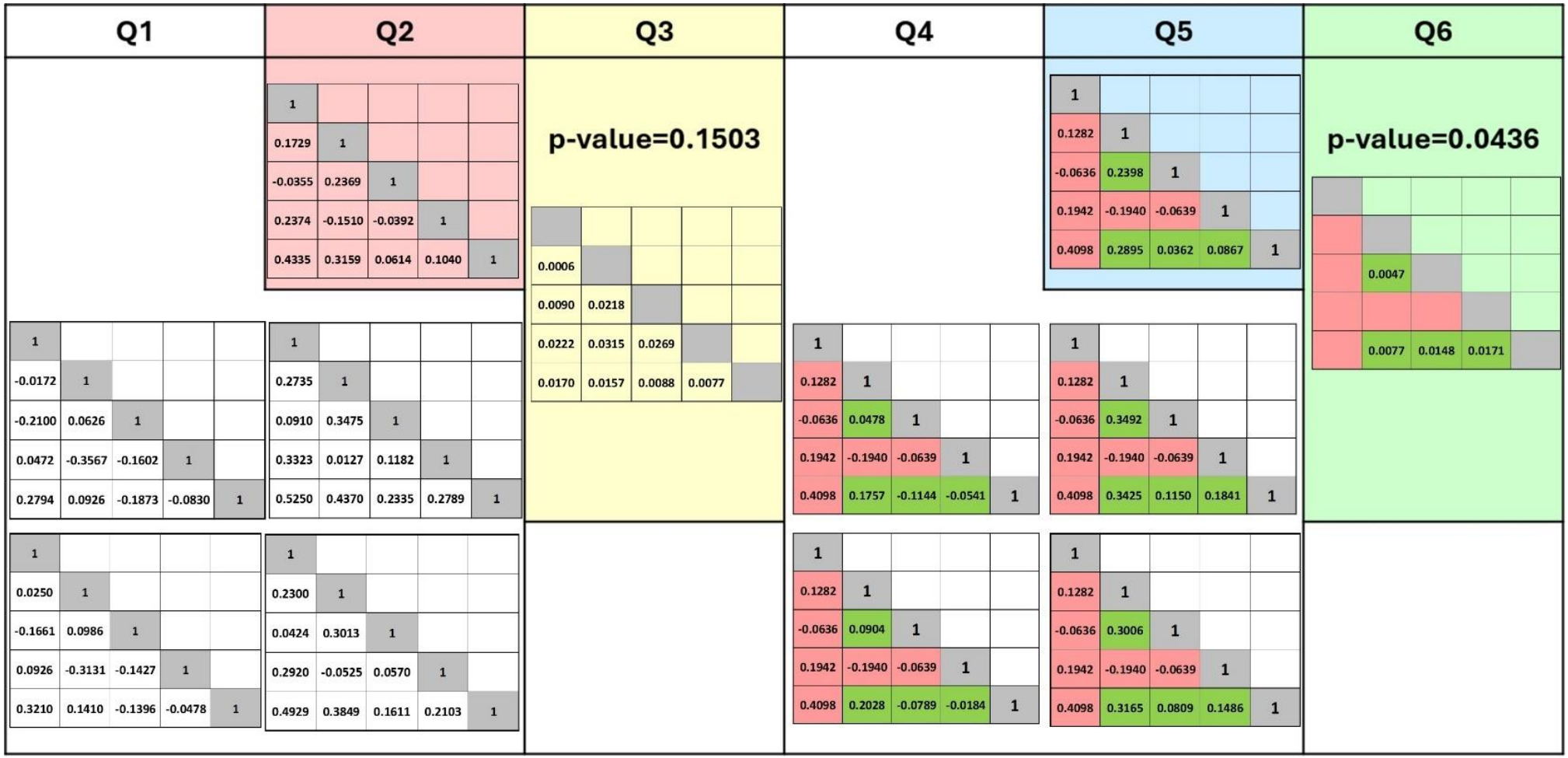
1				
0.13	1			
-0.06	0.350	1		
0.19	-0.19	-0.06	1	
0.41	0.180	0.125	0.215	1

Q4. **Confidence Intervals**: What are the two correlation matrices that correspond to the lower– and upper–bounds of the 95% confidence interval for [D] (holding constant the non-selected red cells)? What are, simultaneously, the individual 95% confidence intervals for only those cells of [D] that are relevant to the scenario (green)?

Q5. **Quantile Function**: What is the unique correlation matrix associated with [E], a matrix of cumulative distribution function values associated with the corresponding cells of [D]?

Q6. **p-values**: Under the null hypothesis that observed correlation matrix [F] was sampled from the (scenario-restricted) data generating mechanism of [D], what is the p-value associated with [F] (with red cells held constant)? And simultaneously, what are the individual p-values associated with every (non-constant, green) cell of [F]?

TABLE B: NAbC Provides Complete Inference, Example of Kendall's Tau – Cell and Matrix Level p-values and Confidence Intervals



Answers to these questions require inference at both the cell and matrix levels, simultaneously and with cross-level consistency, as well as requiring the matrix-level quantile function, all under both the unrestricted and scenario-restricted cases, under any data conditions. Only NAbC can simultaneously answer Q1.-Q6. above under general data conditions, as shown below in Table B.

For Q1 and Q4, the two top matrices correspond to the first (matrix-level) question, and the bottom two matrices correspond to the second (cell-level) question. Note the wider intervals on a cell-by-cell basis for the matrix-level confidence intervals compared to the cell-level confidence intervals, as expected. Also note, for Q3 and Q6, the smaller p-values for the individual cells compared to the respective matrix-level p-values, which are larger, as expected, as they control the family-wise error rate (FWER). Note also that the green cells of Q5 differ from the corresponding cells in Q2: even though the (green) angles distributions themselves remain unaffected by scenario restrictions, the ultimate correlation values of those cells ARE affected due to the matrix multiplication of the Cholesky factor, $R = BB^T$. Finally, note that the empirical values of the red cells in Q4-Q6 differ slightly from those in [D] and [F]. This is due to NAbC's conservative use of the mean of the estimated angles (correlation) matrices, rather than presuming we know the absolute 'true' values of these cells (although this is justified in some specific cases).

NAbC REMAINS "ESTIMATOR AGNOSTIC"

Another important and useful characteristic of NAbC, only addressed so far as one of the original seven objectives, is that it remains "estimator agnostic," that is, valid for use with any reasonable estimator of any of the dependence measures being modeled (e.g. Kendall's or Pearson's or Chatterjee's, etc.). Different estimators will have different characteristics under different data conditions. For example, some will provide minimum variance / maximum power, while others may provide unbiasedness or less bias, while others may provide more robustness, and/or different and shifting combinations of these characteristics. Ideally, we would like to be able to use estimators that provide the best trade-offs for our purposes under the conditions most relevant to our given portfolio. Fortunately, NAbC "works" for any estimator, as the relationship between correlations and angles requires only symmetric positive definiteness. NAbC's finite sample distribution and its resulting inferences obviously will inherit the advantages and disadvantages of the estimator being used, but this is generally an advantage as it provides flexibility to use the 'best' estimator under the widest possible range of conditions.[31]

[31] All empirical results herein use as the estimator the sample correlation matrix, as sample sizes all have exceeded 10p (10 times the dimension of the matrix), which is a widely used threshold for whether a more sophisticated, bias-correcting estimator is needed (see Bongiorno, Challet, and Loeper, 2023). When sample sizes do not meet this threshold, the covariance/correlation matrix estimation literature is rich and deep, but from among the many approaches, I have found the Average Oracle of Bongiorno, Challet, and Loeper (2023) to be among the most compelling theoretically, empirically, and in terms of practicality and transparency of usage under a wide range of real-world data conditions. Notably, Average Oracle outperforms all flavors of non-linear shrinkage a la Ledoit and Wolf (2022a,b) (see Bongiorno and Challet, 2023).

NAbC AND GENERALIZED ENTROPY

In a relevant and validating digression, it is intriguing and important to note that the (two-sided) cell-level p-values NAbC provides (see Q3 and Q6 in Table B above) actually can be used to construct a competitor to commonly used distance metrics, such as norms, and it has a number of advantages over them in this setting. Some commonly used norms for measuring correlation 'distances' include the Taxi, Frobenius/Euclidean, and Chebyshev norms (collectively, the Minkowski norm), listed below in (21).

(21) $$\|x\| = \left(\sum_{i=1}^{d} |x_i|^m \right)^{1/m}$$ where x is a distance from a presumed or baseline correlation value, d=number of observations, and m=1, 2, and ∞ correspond to the Taxi, Frobenius/Euclidean, and Chebyshev norms, respectively.

All of these norms measure absolute distance from a presumed or baseline correlation value. But the range of all relevant and widely used dependence measures is bounded, either from –1 to 1 or 0 to 1, and the relative impact and meaning of a given distance at the boundaries are not the same as those in the middle of the range. In other words, a shift of 0.02 from an original or presumed correlation value of, say, 0.97, means something very different than the same shift from 0.47. NAbC's p-values attribute probabilistic MEANING to these two different cases, while a norm would treat them identically, even though they very likely indicate what are very different events of very different relative magnitudes with potentially very different consequences.

Therefore, a natural, PROBABILISTIC distance measure based directly on NAbC's cell-level p-values is the natural log of the product of the p-values, dubbed 'LNP' in (22) below:

(22) $$\text{"LNP"} = \ln\left(\prod_{i=1}^{q} p\text{-}value_i \right) = \sum_{i=1}^{q} \ln\left[p\text{-}value_i \right] \text{ where } q = p(p-1)/2 \text{ and } p\text{-}value_i \text{ is 2-sided.}$$

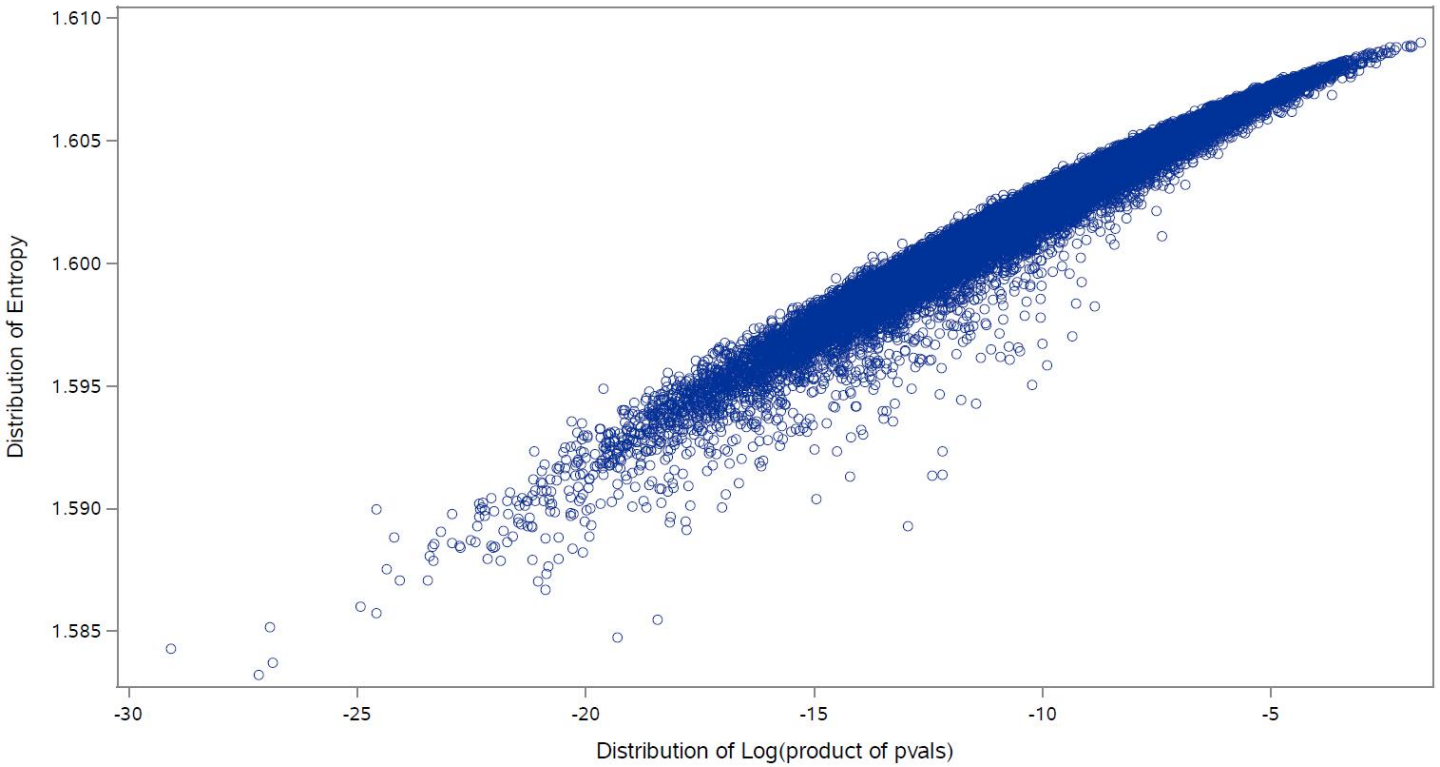
This was shown in a previous publication of this article, using a Pearson's correlation matrix under the (Gaussian) identity matrix, to have a very strong correspondence with the entropy of the correlation matrix,[32] defined by Felippe et al. (2021 and 2023) as (23) below:

(23) $$\text{Entropy} = Ent(R/p) = -\sum_{j=1}^{p} \lambda_j \ln(\lambda_j)$$

where R is the sample correlation matrix and λ_j are the p eigenvalues of the correlation matrix after it is scaled by its dimension, R/p. Importantly, this result (23), like NAbC, is valid for ANY positive definite measure of dependence, not just Pearson's. Graph 18 below compares LNP to the entropy of the Kendall's Tau matrix in 10,000 simulations under the Gaussian identity matrix. The resulting Pearson's correlation between them is 0.98 (virtually identical to the comparison of LNP based on Pearson's rather than Kendall's).

[32] The Pearson's correlation between LNP and the entropy of Felippe et al. (2021 and 2023) was just under 0.99.

Graph 18: Identity Matrix Simulations – LNP (based on Kendall's) v Entropy



It is important to note, however, that entropy here is limited to being calculated relative to the case of independence, which for many dependence measures corresponds only with the identity matrix.[33] In contrast, LNP can be calculated, and retains its meaning, in all cases, based on ANY values of the dependence matrix, not just the case of independence. Yet the correspondence of LNP to entropy under this specific case speaks to LNP's natural interpretation as a meaningful measure of deviation/distance/disorder (depending on your interpretation), and one that also is more flexible and granular than entropy as it is measured cell-by-cell, p(p-1)/2 times, as opposed to only p times for p eigenvalues. As such, LNP might be considered a type of 'generalized entropy' relative to any baseline, as specified by the researcher (i.e. the specified dependence matrix), that is not necessarily perfect (in)dependence. Such measures certainly are relevant in this setting as 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).

Interpretations aside, the use of LNP here warrants further investigation as a matrix-level measure that, unlike widely used distance measures such as norms, has a solid and meaningful probabilistic foundation. Its calculation applies not only beyond the independence case generally, but also to ALL positive definite measures of dependence, regardless of their values. LNP's range of application is as wide as that of NAbC's matrix-level p-value, and the two are readily calculated side-by-side as they are

[33] Recall, of course, that a zero value for Pearson's or Kendall's or Spearman's does not imply independence, but independence does imply a zero value for these measures.

both based on NAbC's cell-level (two-sided) p-values for the entire matrix. These are intriguing results with possibly far-reaching implications.

NAbC AND CAUSAL MODELING

In today's rapidly evolving data science world, the tired mantra of "correlation is not causation" fails to advance our knowledge frontier, and even can be misleading in many cases. While correlation is not causation, neither is it NOT causation! In other words, strong association-based findings from methods like dependence measures far more often than not, when responsibly and rightly implemented, do, in fact, indicate causal mechanisms. This is true even though the big data explosion of the past two decades can make silly and not-so-silly counterexamples seem almost the rule rather than the exception. Without appropriate knowledge and implementation of critically important methods like multiple comparisons adjustments (see Hsu, 1999, and Westfall and Young, 1993, and Efron, 2004, 2007a, 2007b), these counterexamples remain vulnerable strawmen for those who would dismiss or minimize association-based modeling (ABM), and over a century of statistical and econometric research, merely as inferior and limited prologue to the parousia of causal modeling (CM). While this may be effective for promoting a causality book or a new related piece of software, it actually limits the advancement of our applied research knowledgebase, and even the advancement of CM, because it turns out that many of the most effective CM methods are built squarely on ABM methods (see Opdyke, 2024b, for additional examples).

> "… a DAG structure recovering algorithm, which is **based on the Cholesky factorization of the covariance matrix of the observed data** (emphasis added). …achieves the state-of-the-art performance." Cai et al., 2023

Even more importantly, while CM is not new, its widespread application is. We are still learning critically important caveats and cautionary tales about the two primary causal paradigms when attempting their responsible application in real-world settings.

> "The clarion call for causal reduction in the study of capital markets is intensifying. However, in self-referencing and open systems such as capital markets, the idea of unidirectional causation (if applicable) may be limiting at best, and unstable or fallacious at worst." Polakow et al., 2023

> "Most of the literature on causality considers the structural framework of Pearl and the potential-outcomes framework of Neyman and Rubin to be formally equivalent, and therefore interchangeably uses the do-notation and the potential-outcome subscript notation to write counterfactual outcomes. In this paper, we … prove that structural counterfactual outcomes and potential outcomes do not coincide in general – not even in law." De Lara, 2024.

> "… potential outcomes (PO) and structural causal models (SCMs) stand as the predominant frameworks. However, these frameworks face notable challenges in practically modeling counterfactuals … we identify an inherent model capacity limitation, termed as the "degenerative counterfactual problem", emerging from the consistency rule that is the cornerstone of both

> frameworks. …We hope it opens new avenues for future research of counterfactual modeling, ultimately enhancing our understanding of causality and its real-world applications." Gong et al, 2024

> "Try to come up with some form of efficient frontier for the 500 constituents of the SP500 with a causal model. Or ask yourself if that even make sense in the first place.", Alejandro Rodriguez Dominguez, 2024a, Causal Researcher, Head of Quantitative Analysis at Miraltabank

In fact, CM arguably has hit a hype cycle, which endangers its responsible use in the medium to longer-term future when it fails to live up to its over-hyped promise, at least in certain settings.

> "The explicit study of causality in AI fields has officially hit the 'hype cycle', at least according to Gartner" Grimbly, 2022, and Gartner, 2022

Of course, all this is not to say that we shouldn't aggressively pursue promising causality research (see Rodriguez Dominguez & Yadav, 2024, and Rodriguez Dominguez, 2024b), even in settings like finance where it may be extremely challenging to implement responsibly and reliably; only that we should avoid its hype cycle while doing so. We need to refrain from setting up sometimes convenient, but largely useless "correlation vs causation" strawmen, which serve no real research purpose, and objectively and honestly keep causality's very real limitations front-and-center when pursuing its rigorous application in challenging, real-world settings.

So to say "correlation is not causation," while true and very important for an introductory statistics course, most certainly is not the same as stating "association-based modeling is not causal modeling," because, in fact, and in real-world practice, the lines here are rightly blurred, and the latter builds on the former, albeit within compelling and insightful and original paradigms that went underappreciated and underutilized for far too long.

> "…the distinction between prediction and causation, taken to its limit, melts away." Daoud & Dubhashi, 2023

Failing to recognize the blurred lines that melt away likely will limit the scope of applied CM research as we will miss opportunities to combine the best of both worlds (to the extent that they are distinct) to develop the most effective methods to tackle the hardest problems we face in real-world settings, especially in finance.

Enter NAbC. While unarguably an "ABM" method that broadens, enables, and enhances robust statistical inference in challenging, real-world financial settings, it, too, can be used to tackle questions posed within the causal paradigm(s). NAbC's broad range of application becomes critically important and useful here. As shown above, NAbC remains valid for ANY dependence measure whose pairwise matrix is symmetric positive definite. This includes many ASYMMETRIC, directional dependence measures, such as Chatterjee's new correlation coefficient (Chatterjee, 2021), the improved Chatterjee's coefficient (Xia et al, 2024), Zhang's combined correlation measure (Zhang, 2023), the QAD measure of Junker et al. (2021), the asymmetric tail dependence measure (Deidda et al, 2023), and others. Each of

these has different power under different dependence conditions, whether these are monotonic relationships or highly nonlinear relationships or highly cyclical relationships or asymmetric dependence in the extreme tails, or various combinations thereof. But the important point here is that 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 the causal paradigm(s): 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.

But why do we need NAbC to do this? This calculation can be done, albeit ineffectually, individually and directionally for each of the pairwise associations of a covariate with a treatment (X) and its dependent variable (Y) and all the other covariates. NAbC's important contribution here is that its inferences (p-values) are based not merely on an isolated pairwise measure, but rather on the entire matrix of covariates, and their relationships with each other AND X and Y, **simultaneously**.[34] While each of the angles distributions used by NAbC is independent of the others, the distributions of the associated correlation cells most certainly are not, due to $R = BB^T$. So the matrix-based approach, as opposed to the individual covariate-by-covariate approach, is critically important in this setting, because the pairwise (directional) matrices that correspond to a directed acyclic graph (DAG) are not merely groups of otherwise unrelated pairwise relationships: they are entire matrices of typically complex, intertwined, directional relationships, and the only viable way to attempt to recover the DAG accurately is to analyze the entire matrices simultaneously, which is what NAbC does.

NAbC has been applied in this way to data generated by verified DAGs and preliminary classification rates appear very promising. This is an area of further research and application and testing for NAbC. But of course, none of this addresses the question of whether DAGs can be used reliably within "self-referencing open systems like capital markets" (Polakow et al., 2023) to begin with; only that it appears NAbC can play a role in recovering them if the answer to this question is "yes" or "under some conditions."[35]

I'll close this section with the cautionary note that setting is key here. What may be an appropriate and relatively straightforward application of a causal framework in, say, a clinical trial setting, with provably viable assumptions and satisfied methodology constraints, may be wholly unjustified in a finance setting. Just because the mathematics can be coded and the computations performed doesn't mean the underlying requirements are met or the presumptions are valid. As with all (data) science, we must remain cognizant of these restrictions, especially in the face of the pressure exerted by hype cycles to

[34] Of course, multivariate regression models do this too, although they are addressing a somewhat different but closely related set of questions.

[35] Czado (2025) demonstrates that vine copulas, described above as being a very flexible and effective method for real-world portfolio simulation (if not dependence measure inference), also can be remarkably effective in the causal discovery setting.

market the next new thing. The seminal works of Pearl (1996, 2000) and Angrist, Imbens, and Rubin (1996) obviously provided an extremely valuable paradigm shift that generated powerful new tools to address research questions not previously answerable, or at least only partially answerable. But when they are presented as a separate and superior mode of inquiry, unrelated to all that has come before, researchers will end up limiting the very research they seek to promote. Responsible researchers must resist these siren's songs, eschew artificial, strawman divisions between causal and association-based modeling, and agnostically and scientifically embrace the most effective methodological combinations of both. I look forward to further testing NAbC within the causal paradigm(s) to confirm that it can be classified as one of these 'best of both' methods.

CONCLUSIONS

NAbC defines the finite sample distributions of an extremely broad range of dependence measures – all those whose pairwise matrices are positive definite – under challenging, real-world financial data conditions (i.e non-iid multivariate returns data with varying degrees of asymmetry, (non-)stationarity, serial correlation, and heavy-tailedness in the margins). Motivation for its development has been the need for a method that satisfies all seven of the objectives listed below, because to date, no extant method addressed all of these "real-world necessary" requirements simultaneously. Yet anything less than this, when modeling dependence structure in our risk and investment portfolios, fails to rise to the same level of analytical rigor as has been applied to the other parameters of these models: that is indefensible given that its effects can be larger than many, if not all of the other parameters combined. I list again the seven objectives below for the reader's convenience:

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[36]

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)

[36] 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.

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,[37] enabling reliable application in actual, real-world, non-textbook settings

For the specific case of Pearson's under the Gaussian identity matrix, NAbC allows me to provide an interactive spreadsheet that implements the fully analytic solution, with p-values and confidence intervals at both the cell and matrix levels (along with a measure of generalized entropy).

http://www.datamineit.com/JD%20Opdyke--The%20Correlation%20Matrix-Analytically%20Derived%20Inference%20Under%20the%20Gaussian%20Identity%20Matrix--02-18-24.xlsx

But way beyond Pearson's, the fully general NAbC solution presented herein checks all seven boxes, simultaneously. The list of critically important, applied research that NAbC now facilitates, if not makes possible, is not only expansive, but also feasible with an ease of use and interpretability, broad range of application, scalability, and robustness not found in other more limited (spectral) methods with narrow ranges of application. What's more, preliminary tests show NAbC to be directly usable and useful in causal modeling, further expanding its already comprehensive scope.

With NAbC, 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. 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 should prove to be a very useful means by which we can better understand, predict, and manage portfolios in our multivariate world.

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

REFERENCES

Abul-Magd, A., Akemann, G., and Vivo, P., (2009), "Superstatistical Generalizations of Wishart-Laguerre Ensembles of Random Matrices," *Journal of Physics A Mathematical and Theoretical*, 42(17):175207.

Adams, R., Pennington, J., Johnson, M., Smith, J, Ovadia, Y., Patton, B., Saunderson, J., (2018), "Estimating the Spectral Density of Large Implicit Matrices" https://arxiv.org/abs/1802.03451.

Akemann, G., Fischmann, J., and Vivo, P., (2009), "Universal Correlations and Power-Law Tails in Financial Covariance Matrices," https://arxiv.org/abs/0906.5249.

Almog, A., and Shmueli, E., (2019), "Structural Entropy: Monitoring Correlation-Based Networks over time With Application to Financial Markets," *Scientific Reports*, 9:10832.

Angrist, JD, Imbens, G., and Rubin, D., (1996), " Identification of causal effects using instrumental variables (with discussion)," *J. Amer. Statist. Assoc.*, 91 444–472.

Askitis, D., (2017), "Asymptotic expansions of the inverse of the Beta distribution," https://arxiv.org/abs/1611.03573

BIS, Basel Committee on Banking Supervision, Working Paper 19, (1/31/11), "Messages from the academic literature on risk measurement for the trading book."

Bohn, W., Hornik, K., (2014), "Generating random correlation matrices by the simple rejection method: Why it does not work," *Stat. & Prob.Letters*, 87 (C), 27-30.

Bongiorno, C., Challet, D., and Loeper, G., (2023), "Filtering Time-dependent Covariance Matrices Using Time-Independent Eigenvalues," https://arxiv.org/pdf/2111.13109.

Bongiorno, C., and Challet, D., (2023), "Covariance Matrix Filtering and Portfolio Optimisation: The Average Oracle vs Non-Linear Shrinkage and All the Variants of DCC-NLS," https://arxiv.org/abs/2309.17219.

Bouchaud, J, & Potters, M., (2015), "Financial applications of random matrix theory: a short review," The Oxford Handbook of Random Matrix Theory, Eds G. Akemann, J. Baik, P. Di Francesco.

Burda, Z., Jurkiewicz, J., Nowak, M., Papp, G., and Zahed, I., (2004), "Free Levy Matrices and Financial Correlations," *Physica A: Statistical Mechanics and its Applications*.

Burda, Z., Gorlich, A., and Waclaw, B., (2006), "Spectral Properties of empirical covariance matrices for data with power-law tails," *Phys. Rev., E 74*, 041129.

Burda, Z., Jaroz, A., Jurkiewicz, J., Nowak, M., Papp, G., and Zahed, I., (2011), "Applying Free Random Variables to Random Matrix Analysis of Financial Data Part I: A Gaussian Case," *Quantitative Finance*, Volume 11, Issue 7, 1103-1124.

Cai, Y., Li, X., Sun, M., and Li, P., (2023), “Recovering Linear Causal Models with Latent Variables via Cholesky Factorization of Covariance Matrix,” arXiv:2311.00674v1 [stat.ML].

Chakraborti, A., Hrishidev, Sharma, K., and Pharasi, H., (2020), “Phase Separation and Scaling in Correlation Structures of Financial Markets,” *Journal of Physics: Complexity*, 2:015002.

Chatterjee, S., (2021), “A New Coefficient of Correlation,” *Journal of the American Statistical Association*, Vol 116(536), 2009-2022.

Chatterjee, S., (2022), “A Survey of Some Recent Developments in Measures of Association,” ArXiv preprint, arXiv:2211.04702.

Church, Christ (2012). "The asymmetric t-copula with individual degrees of freedom", Oxford, UK: University of Oxford Master Thesis, 2012.

Cordoba, I., Varando, G., Bielza, C., and Larranaga, P., (2018), “A fast Metropolis-Hastings method for generating random correlation matrices,” *IDEAL*, pp. 117-124, part of Lec Notes in Comp Sci., Vol 11314.

Czado, C., (2025), “Vine Copula Based Structural Equation Models,” *Computational Statistics and Data Analysis*, pp.453-477.

Czado, C., and Nagler, T., (2022), “Vine Copula Based Modeling,” *Annual Review of Statistics and Its Application*, pp.453-477.

Dalitz, C., Arning, J., and Goebbels, S., (2024), “A Simple Bias Reduction for Chatterjee’s Correlation,” arXiv:2312.15496v2.

Daoud, A., and Dubhashi, D., (2023), “Statistical Modeling: The Three Cultures,” *Harvard Data Science Review*, Issue 5.1.

De Lara, L., (2023), “On the (in)compatibility between potential outcomes and structural causal models and its signification in counterfactual inference,” arXiv:2309.05997v3 [math.ST].

Deidda, C., Engelke, S., and De Michele, C., (2023), “Asymmetric Dependence in Hydrological Extremes,” *Water Resources Research*, Vol. 59, Issue 12.

Digital Library of Mathematical Functions (DLMF), Section 8.17.ii, Hypergeometric Representations, National Institute of Standards and Technology (NIST), Handbook of Mathematical Functions, US Departement of Commerce, by Cambridge University Press, Online Version 1.2.1; Release date 2024-06-15 (https://dlmf.nist.gov/8.17#ii).

Efron, B. (2004), “Large-scale simultaneous hypothesis testing: The choice of a null hypothesis,” *J. Amer. Statist. Assoc.*, 99 96–104.

Efron, B. (2007a), “Correlation and large-scale simultaneous significance testing,” *J. Amer. Statist. Assoc.*, 102 93–103.

Efron, B. (2007b), "Doing thousands of hypothesis tests at the same time," *Metron*, LXV 3–21.

Embrechts, P., Hofert, M., and Wang, R., (2016), "Bernoulli and Tail-Dependence Compatibility," *The Annals of Applied Probability*, Vol. 26(3), 1636-1658.

Fang, Q., Jiang, Q., and Qiao, X., (2024), "Large-Scale Multiple Testing of Cross-Covariance Functions with Applications to Functional Network," arXiv:2407.19399v1 [math.ST] 28 Jul.

Felippe, H., Viol, A., de Araujo, D. B., da Luz, M. G. E., Palhano-Fontes, F., Onias, H., Raposo, E. P., and Viswanathan, G. M., (2021), "The von Neumann entropy for the Pearson correlation matrix: A test of the entropic brain hypothesis," working paper, arXiv:2106.05379v1

Felippe, H., Viol, A., de Araujo, D. B., da Luz, M. G. E., Palhano-Fontes, F., Onias, H., Raposo, E. P., and Viswanathan, G. M., (2023), "Threshold-free estimation of entropy from a Pearson matrix," working paper, arXiv:2106.05379v2.

Fernandez-Duran, J.J., and Gregorio-Dominguez, M.M., (2023), "Testing the Regular Variation Model for Multivariate Extremes with Flexible Circular and Spherical Distributions," arXiv:2309.04948v2.

Franca, W., and Menegatto, V., (2022), "Positive definite functions on products of metric spaces by integral transforms," *Journal of Mathematical Analysis and Applications*, 514(1).

Fuchs, S., (2024), "Quantifying Directed Dependence via Dimension Reduction," *Journal of Multivariate Analysis*, 201:105266.

Gamboa, F., Gremaud, P., Klein, T., and Lagnoux, A., (2022), "Global Sensitivity Analysis: A Novel Generation of Mighty Estimators Based on Rank Statistics," *Bernoulli*, 28(4):2345–2374.

Gao, M., Li, Q., (2024), "A Family of Chatterjee's Correlation Coefficients and Their Properties," arXiv:2403.17670v1 [stat.ME]

Gartner, (2002), What's New in the 2022 Gartner Hype Cycle for Emerging Technologies. September 10, 2022, https://www.gartner.com/en/articles/what-s-new-in-the-2022-gartner-hype-cycle-for-emerging-technologies

Ghosh, R., Mallick, B., and Pourahmadi, M., (2021) "Bayesian Estimation of Correlation Matrices of Longitudinal Data," *Bayesian Analysis*, 16, Number 3, pp. 1039–1058.

Gong, H., Lu, C., and Zang, Y., (2024), "Distribution-consistency Structural Causal Models" arXiv:2401.15911v2 [cs.AI]

Grimbly, St. John, (2022), "Differential Equations vs. Structural Causal Models," September 10, 2022.

Hansen, B., (2014), Econometrics, Ch. 20 – Nonparametric Density Estimation, p.333

Hisakado, M. and Kaneko, T., (2023), "Deformation of Marchenko-Pastur distribution for the correlated time series," arXiv:2305.12632v1.

Holzmann, H., and Klar, B., (2024) “Lancaster Correlation - A New Dependence Measure Linked to Maximum Correlation,” arXiv:2303.17872v2 [stat.ME].

Hsu, J.(1999), Multiple Comparisons: Theory and Methods, Chapman & Hall / CRC, Boca Raton, FL.

Huang, Z., Deb, N., and Sen, B., (2022), “Kernel Partial Correlation Coefficient – A Measure of Conditional Dependence,” *The Journal of Machine Learning Research*, 23(1):9699–9756.

Johnstone, I., (2001), “On the distribution of the largest eigenvalue in principal components analysis,” *The Annals of Statistics*, 29(2): 295–327, 2001.

Junker. R., Griessenberger, F., and Trutschnig, W., (2021), “Estimating scale-invariant directed dependence of bivariate distributions,” *Computational Statistics & Data Analysis*, Volume 153.

Kendall, M. (1938), "A New Measure of Rank Correlation," *Biometrika*, 30 (1–2), 81–89.

Kim, W., and Lee, Y., (2016), “A Uniformly Distributed Random Portfolio,” *Quantitative Finance*, Vol. 16, No. 2, pp.297-307.

Kurowicka, D., (2014). “Joint Density of Correlations in the Correlation Matrix with Chordal Sparsity Patterns,” *Journal of Multivariate Analysis*, 129 (C): 160–170.

Lewandowski, D.; Kurowicka, D.; Joe, H. (2009). "Generating random correlation matrices based on vines and extended onion method". *Journal of Multivariate Analysis*, 100 (9): 1989–2001.

Ledoit, O., and Wolf, M., (2022a), “Markowitz Portfolios Under Transaction Costs,” *Working paper series/Department of Economics*, (420).

Ledoit, O., and Wolf, M., (2022b), “Quadratic Shrinkage for Large Covariance Matrices,” *Bernoulli*, 28(3):1519-1547.

Li, W. ,Yao, J., (2018), “On structure testing for component covariance matrices of a high-dimensional mixture,” *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 80(2):293-318.

Li, G., Zhang, A., Zhang, Q., Wu, D., and Zhan, C., (2022), “Pearson Correlation Coefficient-Based Performance Enhancement of Broad Learning System for Stock Price Prediction,” *IEEE Transactions on Circuits and Systems—II: Express Briefs*, Vol 69(5), 2413-2417.

Lu, F., Xue, L., and Wang, Z., (2019), “Triangular Angles Parameterization for the Correlation Matrix of Bivariate Longitudinal Data,” *J. of the Korean Statistical Society*, 49:364-388.

MacKinnon, D., and Lamp, S., (2021), “A Unification of Mediator, Confounder, and Collider Effects,” *Prev Sci.*, 22(8), 1185-1193.

Madar, V., (2015), “Direct Formulation to Cholesky Decomposition of a General Nonsingular Correlation Matrix,” *Statistics & Probability Letters*, Vol 103, pp.142-147.

Makalic, E., Schmidt, D., (2018), “An efficient algorithm for sampling from sin(x)^k for generating random correlation matrices,” arXiv: 1809.05212v2 [stat.CO].

Maltsev, A., and Malysheva, S. (2024), “Eigenvalue Statistics of Elliptic Volatility Model with Power-law Tailed Volatility,” arXiv:2402.02133v1 [math.PR].

Marchenko, A., Pastur, L., (1967), "Distribution of eigenvalues for some sets of random matrices,” *Matematicheskii Sbornik*, N.S. 72 (114:4): 507–536.

Martin, C. and Mahoney, M., (2018), “Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning,” *Journal of Machine Learning Research*, 22 (2021) 1-73.

McNeil, A., Frey, R., and Embrechts, P., (2005). Quantitative Risk Management: Concepts, Techniques, and Tools, Princeton, NJ: Princeton University Press.

Meucci, A., (2010a), “The Black-Litterman Approach: Original Model and Extensions,” The Encyclopedia of Quantitative Finance, Wiley, 2010

Meucci, A., (2010b), “Fully Flexible Views: Theory and Practice,” arXiv:1012.2848v1

Muirhead, R., (1982), Aspects of Multivariate Statistical Theory, Wiley Interscience, Hoboken, New Jersey.

Ng, F., Li, W., and Yu, P., (2014), “A Black-Litterman Approach to Correlation Stress Testing,” *Quantitative Finance*, 14:9, 1643-1649.

Opdyke, JD, (2020), “Full Probabilistic Control for Direct & Robust, Generalized & Targeted Stressing of the Correlation Matrix (Even When Eigenvalues are Empirically Challenging),” QuantMinds/RiskMinds Americas, Sept 22-23, Boston, MA.

Opdyke, JD, (2022), “Beating the Correlation Breakdown: Robust Inference and Flexible Scenarios and Stress Testing for Financial Portfolios,” QuantMindsEdge: Alpha and Quant Investing: New Research: Applying Machine Learning Techniques to Alpha Generation Models, June 6.

Opdyke, JD, (2023), “Beating the Correlation Breakdown: Robust Inference and Flexible Scenarios and Stress Testing for Financial Portfolios,” Columbia University, NYC–School of Professional Studies: Machine Learning for Risk Management, Invited Guest Lecture, March 20.

Opdyke, JD, (2024a), Keynote Presentation: “Beating the Correlation Breakdown, for Pearson’s and Beyond: Robust Inference and Flexible Scenarios and Stress Testing for Financial Portfolios,” QuantStrats11, NYC, March 12.

Opdyke, JD, (2024b), “Association-based vs Causal Research: the Hype, the Contrasts, and the Stronger-than-expected Complementary Overlaps,” QuantStrats 11, NYC, March, 12, 2024.

Pafka, S., and Kondor, I., (2004), “Estimated correlation matrices and portfolio optimization,” Physica A: Statistical Mechanics and its Applications, Vol 343, 623-634.

Papenbrock, J., Schwendner, P., Jaeger, M., and Krugel, S., (2021), “Matrix Evolutions: Synthetic Correlations and Explainable Machine Learning for Constructing Robust Investment Portfolios,” *Journal of Financial Data Science*, 51-69.

Pascual-Marqui, R., Kochi, K., and Kinoshita, T. (2024), “Distance-based Chatterjee Correlation: A New Generalized Robust Measure of Directed Association for Multivariate Real and Complex-Valued Data,” arXiv:2406.16458 [stat.ME].

Pearl, J., (2000), Causality: Models, Reasoning, and Inference, Cambridge University Press.

Pearl, J., (1996), “The Art and Science of Cause and Effect,” Lecture given Thursday, October 29, 1996, UCLA 81st Faculty Research Lecture Series

Pearson, K., (1895), “VII. Note on regression and inheritance in the case of two parents,” *Proceedings of the Royal Society of London*, 58: 240–242.

Pinheiro, J. and Bates, D. (1996), “Unconstrained parametrizations for variance-covariance matrices,” *Statistics and Computing*, Vol. 6, 289–296.

Polakow, D., Gebbie, T., and Flint, E., (2023), “Epistemic Limits of Empirical Finance: Causal Reductionism and Self-Reference,” arXiv:2311.16570v2 [q-fin.GN]

Pourahmadi, M., Wang, X., (2015), “Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor,” *Statistics and Probability Letters*, 106, (C), 5-12.

Qin, T., and Wei-Min, H., (2024), “Epanechnikov Variational Autoencoder,” arXiv:2405.12783v1 [stat.ML] 21 May 2024.

Qian, E. and Gorman, S. (2001). “Conditional Distribution in Portfolio Theory.” *Financial Analysts Journal*, 44-51.

Rapisarda, F., Brigo, D., & Mercurio, F., (2007), “Parameterizing Correlations: A Geometric Interpretation,” *IMA Journal of Management Mathematics*, 18(1), 55-73.

Rebonato, R., and Jackel, P., (2000), “The Most General Methodology for Creating a Valid Correlation Matrix for Risk Management and Option Pricing Purposes,” *Journal of Risk*, 2(2)17-27.

Rodriguez Dominguez, A., and Yadav, O., (2024), “Measuring causality with the variability of the largest eigenvalue,” *Data Science in Finance and Economics*.

Rodriguez Dominguez, A., (2024a), LinkedIn Commentary, permission granted for citation.

Rodriguez Dominguez, A., (2024b), “Geometric Spatial and Temporal Constraints in Dynamical Systems and Their Relation to Causal Interactions between Time Series,”
SSRN: https://ssrn.com/abstract=4949383 or http://dx.doi.org/10.2139/ssrn.4949383

Romano, J., and Wolf, M., (2016), "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing," *Statistics & Probability Letters*, Vol 113, 38-40.

Rubsamen, Roman, (2023), "Random Correlation Matrices Generation," https://github.com/lequant40/random-correlation-matrices-generation

Sabato, S., Yom-Tov, E., Tsherniak, A., Rosset, S., (2007), "Analyzing systemlogs: A new view of what's important," *Proceedings, 2nd Workshop of ComputingSystems ML*, pp.1–7.

Schreyer, M., Paulin, R., and Trutschnig, W., (2017), "On the exact region determined by Kendall's tau and Spearman's rho," arXiv: 1502:04620.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K., (2013) "Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing," *The Annals of Statistics*, 41(5), 2263-2291.

Sharma, D., and Chakrabarty, T., (2017), "Some General Results on Quantile Functions for the Generalized Beta Family," *Statistics, Optimization and Information Computing*, 5, 360-377.

Shyamalkumar, N., and Tao, S., (2020), "On tail dependence matrices: The realization problem for parametric families," *Extremes*, Vol. 23, 245–285.

Silverman, B., (1986), Density Estimation for Statistics and Data Analysis, New York, Chapman and Hall.

Spearman, C., (1904), "'General Intelligence,' Objectively Determined and Measured," *The American Journal of Psychology*, 15(2), 201–292.

Szekely, G., Rizzo, M., and Bakirov, N., (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35(6), pp2769-2794.

Taraldsen, G. (2021), "The Confidence Density for Correlation," *The Indian Journal of Statistics*, 2021.

Thakkar, A., Patel, D., and Shah, P., (2021), "Pearson Correlation Coefficient-based performance enhancement of Vanilla Neural Network for Stock Trend Prediction," *Neural Computing and Applications*, 33:16985-17000.

Tsay, R., and Pourahmadi, M., (2017), "Modelling structured correlation matrices," *Biometrika*, 104(1), 237–242.

van den Heuvel, E., and Zhan, Z., (2022), "Myths About Linear and Monotonic Associations: Pearson's r, Spearman's ρ, and Kendall's τ," *The American Statistician*, 76:1, 44-52.

Vorobets, A., (2024a), "Sequential Entropy Pooling Heuristics," https://ssrn.com/abstract=3936392 or http://dx.doi.org/10.2139/ssrn.3936392

Vorobets, A., (2024b), "Portfolio Construction and Risk Management," https://ssrn.com/abstract=4807200 or http://dx.doi.org/10.2139/ssrn.4807200

Wang, Z, Wu, Y., and Chu, H., (2018), “On equivalence of the LKJ distribution and the restricted Wishart distribution,” arXiv:1809.04746v1.

Weisstein, E., (2024a), "Beta Distribution." From *MathWorld*--A Wolfram Web Resource. https://mathworld.wolfram.com/BetaDistribution.html

Weisstein, E., (2024b), "Regularized Beta Function." From *MathWorld*--A Wolfram Web Resource. https://mathworld.wolfram.com/RegularizedBetaFunction.html

Welsch, R., and Zhou, X., (2007), “Application of Robust Statistics to Asset Allocation Models,” *REVSTAT–Statistical Journal*, Volume 5(1), 97–114.

Westfall, P., and Young, S., (1993), Resampling Based Multiple Testing, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, New York.

Xia, L., Cao, R., Du, J., and Chen, X., (2024), “The Improved Correlation Coefficient of Chatterjee,” *Journal of Nonparametric Statistics*, pp1-17.

Xi’an, March, 2018 (https://stats.stackexchange.com/questions/331253/draw-n-dimensional-uniform-sample-from-a-unit-n-1-sphere-defined-by-n-1-dime/331850#331850
and
https://xianblog.wordpress.com/2018/03/08/uniform-on-the-sphere-or-not/

Xu, W., Hou, Y., Hung, Y., and Zou, Y., (2013), “A Comparative Analysis of Spearman’s Rho and Kendall’s Tau in Normal and Contaminated Normal Models,” *Signal Processing*, 93, 261–276.

Yu, P., Li, W., Ng, F., (2014), “Formulating Hypothetical Scenarios in Correlation Stress Testing via a Bayesian Framework,” *The North Amer. J. of Econ. and Finance*, Vol 27, 17-33.

Zhang, Q., (2023), “On relationships between Chatterjee’s and Spearman’s correlation coefficients,” arXiv:2302.10131v1 [stat.ME]

Zhang, Y., and Songshan, Y., (2023), “Kernel Angle Dependence Measures for Complex Objects,” arXiv:2206.01459v2

Zhang, W., Leng, C., and Tang, Y., (2015), “A Joint Modeling Approach for Longitudinal Studies,” *Journal of the Royal Stat. Society, Series B*, 77(1), 219-238.