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

Post 2 of 4: Pearson's Under The Gaussian Identity Matrix

NOTE: These posts summarize a chapter in my forthcoming monograph for Cambridge University Press.

INTRODUCTION

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 these dependence measures under challenging, real-world financial data conditions. Yet this is exactly what is needed to make valid inferences about their estimates, and to use these inferences for a myriad of essential purposes, such as hypothesis testing, dynamic monitoring, realistic and granular scenario and reverse scenario analyses, and mitigating the effects of correlation breakdowns during market upheavals (which is when we need valid inferences the most).

This is the second in a series of four posts which introduces a new and straightforward method – Nonparametric Angles-based Correlation ("NAbC") – for defining the finite-sample distributions of a very wide range of dependence measures for financial portfolio analysis. These include ANY 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 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).[1]

This post focuses on NAbC's application to a narrow but foundational case, which is used as a baseline to greatly expand its range of application in Posts 3 and 4. The core method itself, however, remains little changed under very general conditions.

[1] Note that "positive definite" throughout these four posts refers to the dependence measure calculated on the matrix of all pairwise associations in the portfolio, that is, on a bivariate basis. Some of these dependence measures (eg Szekely's correlation) 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 pairwise associations.

POST 2: NAbC applied to Pearson's under the Gaussian identity matrix.
POST 3: NAbC applied to Pearson's under ALL correlation matrix values and ALL relevant, challenging, real-world financial returns data conditions.[2]
POST 4: NAbC applied to ALL matrix values and ALL positive definite measures of portfolio dependence measures, under ALL relevant, challenging, real-world financial data conditions.

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), 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 presume is Gaussian data under no correlation: that is, Pearson correlation values of zero off the diagonal of the matrix as in (1).[3]

(1) identity matrix = $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ for p = 4 assets

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):[4]

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

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

[3] 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.

[4] Recall that Pearson's requires that the first two moments (the mean and the variance) of the distributions of X and Y are finite.

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 post, 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 $-1/(\mathrm{p}-1) \le r \le 1,\ \mathrm{p} = \dim(r).$

ANGLE VALUES vs CORRELATION VALUES

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):[5]

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

(4)

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

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,[6] 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.

(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:

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

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

[6] Note that this is true not only for Pearson's, but also for all relevant dependence measures in this setting, as will be discussed in Posts 3 and 4.

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

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

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

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

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

(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;
```

The above all is well-established and straightforward, and demonstrates, as we know intuitively, that scale does not (and should not) matter when it comes to dependence measures;[7] 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[8] (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).

[7] 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).

[8] 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 Metropolis-Hastings and Metropolis algorithms of Cordoba et al. (2018), and the restricted Wishart distribution approach of Wang et al. (2018).

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

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 the distribution of each is ***independent*** with respect to those of the others (see Pourahmadi and Wang, 2015, Tsay and Pourahmadi, 2017, 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 the dependence between the two variables (see Fernandez-Duren & Gregorio-Dominguez, 2023, and Zhang & Songshan, 2023, as well as Opdyke, 2024). This will be covered more extensively in subsequent posts.

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 – 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).

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

(9)

We see directly from (6) that $\sin^k(x)$, suitably normalized in (10), satisfies this requirement (see Pourahmadi and Wang, 2015, and Makalic and Schmidt, 2018).

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

(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 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).

(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,[9] but it is admittedly cumbersome mathematically. But Opdyke (2022, 2023, and 2024) has shown that (11) can be simplified further, based on some arguably obscure hypergeometric identities:

[9] 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).

(12)

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 $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ = 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

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

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

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:

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

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

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

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

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

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

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

(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 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.[10]

The above (12) now provides a fully analytic solution,[11] and in fact is so straightforward as to be readily implemented in a spreadsheet, and one is provided for download via the link below and included as a file upload in this Post 2.

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 2024) and obtained over 30% decrease in runtime.

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 $1-F_X\left(x;k\right)$; 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.

---

[10] 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).

[11] 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.

FINITE-SAMPLE DISTRIBUTION OF THE CORRELATION MATRIX

As mentioned previously, a key characteristic of the angles distributions is that they are independent vis-à-vis each other, 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 of the identity matrix (under the presumption of Gaussian data here), the (2-sided) p-value of the entire matrix is simply one minus the probability of no false positives, which is the definition of controlling the family-wise error rate (FWER) of the matrix (15).

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

These critical alphas, when inserted as values in the cdf functions, 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. Independence of the angles distributions again 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. 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.

P-VALUES vs ENTROPY: USING P-VALUES AS A MEASURE OF MATRIX DISTANCE/DISPERSION

Before describing how NAbC, unlike competing methods, enables granular, highly flexible scenarios for dependence measures (key result #6 of POST 1), lets take a moment to examine the meaning and implications of the cell-level p-values derived above in (12) and (13).

The (2-sided) p-value of (13) provides what can be viewed as a distance metric that has some advantages over more traditional distance metrics, such as norms. Some commonly used norms in this setting for measuring correlation 'distances' are listed below in (17).

(17)
$$\|x\| = \left(\sum_{i=1}^{d} |x_i|^m \right)^{1/m}$$

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 0.01 from an original or presumed correlation value of, say, 0.97, means something very different than the same shift from 0.07. NAbC attributes 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 these cell-level p-values from (13) is the natural log of the product of the p-values, dubbed 'LNP' in (18) below:

(18)
$$\text{"LNP"} = \ln\left(\prod_{i=1}^{q} p\text{-}value_i \right) = \sum_{i=1}^{q} \ln\left[p\text{-}value_i \right] \text{ where } q = p(p-1)/2 \text{ and } p\text{-}value_i \text{ is 2-sided.}$$

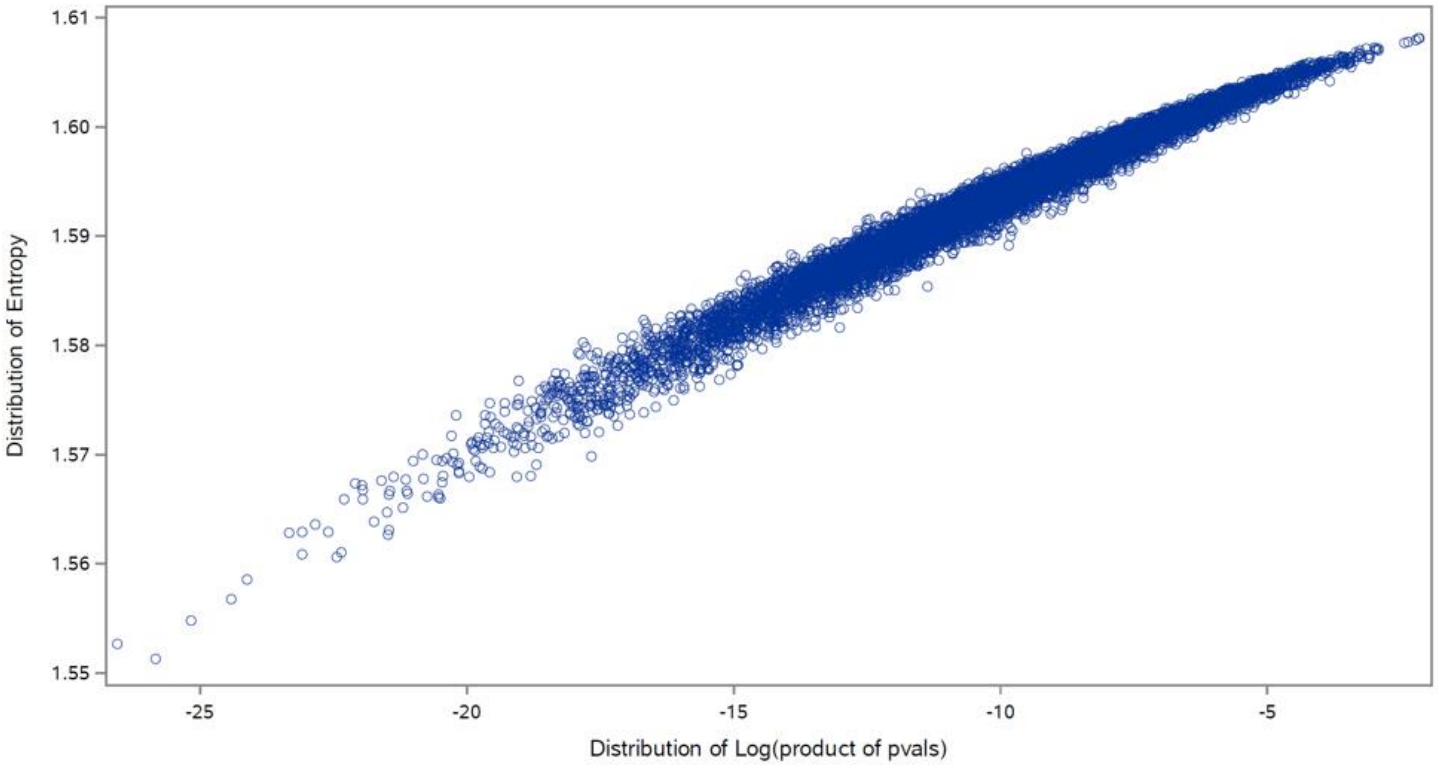
Intriguingly, LNP shows a remarkable correspondence with the entropy of the correlation matrix, defined by Felippe et al. (2021 and 2023) as (19) below:

(19)
$$\text{Entropy} = Ent(R/p) = -\sum_{j=1}^{p} \lambda_j \ln(\lambda_j)$$

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. (Note that this result (19), like NAbC, is valid for ANY positive definite measure of dependence, not just Pearson's, as will be discussed in POSTs 3 and 4).

Graph 1 compares LNP to the entropy of the correlation matrix in 10,000 simulations under the Gaussian identity matrix. The resulting Pearson's correlation between them is just shy of 0.99.

GRAPH 1: Identity Matrix Simulations -- LNP v Entropy



What makes this result worthy of further investigation is that it indicates a broad and useful generalizability of LNP. As will be discussed in Posts 3 and 4, LNP can be calculated for ANY correlation/dependence matrix, not just the identity matrix. Entropy, on the other hand, can be calculated only with reference to the identity matrix as a baseline. Yet the correspondence of LNP to entropy under this specific case speaks to LNP's natural interpretation as a meaningful measure of deviation/distance/dispersion, 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. This topic will be treated in subsequent posts, but is mentioned here as it provides further validation of this approach under this narrow case, as well as much more general conditions.

GRANULAR, HIGHLY FLEXIBLE SCENARIOS

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) the p(p-1)/2 cells for different scenarios.[12] Correlation (dependence) matrices under a tech market bubble (2000) vs those under a

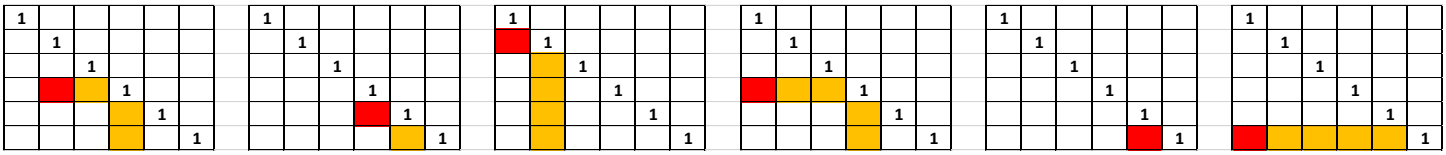
[12] Importantly, when we deviate from the identity matrix (covered in Posts 3 and 4), spectral distributions are far less robust than angles distributions. The latter are bounded, typically unimodal, smooth, not excessively asymmetric, and stable as

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.

Focus only on the lower triangle of the correlation matrices below in Graphs 2-4, 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 (in this narrow case for this Post 2, using inverse transform), 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 *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 2 below.[13]

GRAPH 2: 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 3 below.

matrices approach singularity; in contrast, the former remain unbounded, often are multi-modal, and are far less stable as dependence matrices approach singularity, which is more the rule than the exception when portfolio sizes are not small.

[13] 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.

GRAPH 3: 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 4: 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	

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 attached spreadsheet.

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 only arguable drawback of this approach is that it can be limited by the structure of measuring dependence in pairwise associations. As shown in Graph 4 above, for the p=5 asset matrix, 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 pairwise structure of the matrix. Some scenarios, therefore, could conceivably be required to include for perturbation some few additional cells in the 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, 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 much more difficult to estimate with the same levels of accuracy, let alone to manipulate for purposes of intervention or mitigation. In contrast, pairwise associations are directly identifiable, typically more easily and accurately estimated, and interventions more targeted and transparent.

CONCLUSION

In Post 1 I listed the seven characteristics of the full NAbC solution, and for completeness I list them here below:

1. validity under challenging, real-world financial data conditions, with marginal asset distributions characterized by notably different degrees of serial correlation, non-stationarity, heavy-tailedness, and asymmetry

2. application to ANY positive definite dependence measure, including, for example, Pearson's product moment correlation, rank-based measures like Kendall's tau and Spearman's rho, the kernel-based generalization of Szekely's distance correlation, and the tail dependence matrix, among others.

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

4. it provides valid confidence intervals and p-values at both the matrix-level and the pairwise cell-level, with analytic consistency between these two levels (ie 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. it 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, enabling granular and realistic scenarios

7. it remains valid not just asymptotically, ie for sample sizes presumed to be infinitely large, but rather, for the specific sample sizes we have in reality, enabling reliable application in actual, imperfect, non-textbook settings

This Post 2 covers 4, 5, 6, and 7 above. The next Post 3 expands NAbC to cover 1 as well, using exactly the same angles-based framework. The utility of using the foundational, but undeniably narrow case of the Gaussian identity matrix in this Post 2 rests in establishing the framework and proving out the mechanics of how and why it works, so that we can expand its range of application to the real-world cases of challenging, financial portfolio data. Finally, in Post 4, I expand NAbC's range of application to Characteristics 2 and 3 above, not only to challenging, real-world data conditions, but also simultaneously beyond Pearson's to ALL positive definite measures of dependence.

REFERENCES

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.

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."

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

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).

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

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).

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

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

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

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

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.

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

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.

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

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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