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 4 of 4: ANY Positive Definite Measure, ALL Real-World Financial Data Conditions

NOTE: These posts summarize 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. 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 fourth 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]

[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, calculated on a bivariate basis. Some of these dependence measures (eg 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 attribution and intervention analyses.

This Post 4 expands NAbC's application beyond Pearson's to ANY positive definite dependence measure, under any values, and under all challenging, real-world financial data conditions.

POST 1: NAbC introduced.
POST 2: NAbC applied to Pearson's under the Gaussian identity matrix (fully analytic solution).
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 under ALL relevant, challenging, real-world financial data conditions.

Correlations and Angles (Review of Posts 2 & 3)

To briefly review from Posts 2 & 3, I defined and reviewed the relationship between the correlation cells in a Pearson's correlation matrix and the angles of their corresponding pairwise data vectors. There exists an angle value for every correlation value in the matrix. For a single, bivariate correlation, this can be seen directly via the widely used cosine similarity in (1),[3] but the matrix analog also is well established in the literature as shown in (2.a) and (2.b) (see Pinheiro and Bates, 1996, Rebonato & Jaeckel, 2000, Rapisarda et al., 2007, and Pourahmadi and Wang, 2015, but note a typo in the formula in Pourahmadi and Wang, 2015 corresponding to (2.b) below):

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

(1)

(2.a). For R, a p x p correlation matrix,

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

[2] I take 'real-world' financial returns data to be multivariate with marginal distributions that 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.

[3] 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.

$R = BB^t$ where B is the Cholesky factor (defined in Post 2) 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:[4]

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

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

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

This relationship is one-to-one and bi-directional. I present below straightforward SAS/IML code translating correlations to angles (2.a) and angles to correlations (2.b) in Table A.

The above all is well-established and straightforward. But why are we interested in these angles in this setting? There are several very important reasons:

A. Because they are derived based on the matrix's Cholesky factor, the angles, unlike the correlations themselves, are forced on to the unit hyper-(hemi)sphere, where **positive definiteness automatically is enforced**. This is necessary for efficient sampling, as well as for direct and proper definition of the multivariate sample space (see Post 2 for more detail on this).

B. Crucially, the **distributions of all of the angles are independent**, which makes sampling, and more importantly, construction of their multivariate distribution (and that of the translated correlation matrix), straightforward and useable, where it otherwise would remain intractable.

[4] 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, ie to all positive definite measures of dependence, as shown in my upcoming Post 4.

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;
```

C. **The angles contain all information regarding dependence structure** (see Fernandez-Duren & Gregorio-Dominguez, 2023, and Zhang & Songshan, 2023, as well as Opdyke, 2024). On the UNIT hyper-(hemi)sphere, the only thing we lose is scale, but scale does not and should not matter for any useful and useable measure of dependence.[5]

D. Finally, **angles distributions are more robust and** much better behaved than spectral distributions, and unlike the latter, are **at the right level of aggregation for granular scenarios** (for examples of the dramatic changes of spectral distributions under heavy-tails, see Opdyke, 2024, Burda et al., 2004, Burda et al., 2006, Akemann et al., 2009; Abul-Magd et al., 2009, Bouchaud & Potters, 2015, Martin & Mahoney, 2018), and under serial correlation, see Opdyke, 2024, and Burda et al., 2004, 2011). As discussed in Post 3, I present some empirical examples of this in numerous graphs below under real-world financial data conditions.

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

Fortunately, all of the above advantages of relying on angle values hold not only for Pearson's matrix, but for ANY positive definite dependence measure, under ANY data conditions found in challenging, real-world financial settings.

Beyond Pearson's: Finite Sample Distribution of ANY Positive Definite Dependence Measure

In Post 3 I discuss in more depth why angles distributions are far more appropriate than spectral (eigenvalue) distributions for solving this particular problem, and so do not revisit this comparison here other than to reemphasize points A.-D. above. In this Post 4 I focus on the fact that the only condition required for the relationships between angles and dependence measure values, as shown in (2.a) and (2.b) 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 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, 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: as discussed in Post 3, these include 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 and below 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,[6] 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):

if $x'Rx > 0$ 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
$0 \le R_{i,j} \le 1$ for all $i \ne j$ and $R_{i,i} = 1$ and $R_{i,j} = R_{j,i}$,

[6] 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.

$x'Rx$ can be written in quadratic form as

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

As long as $0 < R_{i,j} < 1 \text{ for all } i \neq j$, that is, the coefficients on the cross terms (the second term of (3)) all remain BETWEEN 0 and 1, then

(4) $\sum_{i=1}^{p} x^2 + 2\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} R_{i,j}x_ix_j > 0 \text{ and so } x'Rx > 0$, 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 $0 < R_{i,j} < 1 \text{ for all } i \neq j \text{ and } R_{i,i} = 1$. For the boundary cases, if $R_{i,j} = 0 \text{ for all } i \neq j$, R obviously remains positive definite as the first term of (3) always is greater than zero and the second term disappears; and if $R_{i,j} = 1 \text{ for all } i \neq j$ 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.

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 above sometimes assign an angle value slightly above **π**/2, even though **π**/2 corresponds to a measure value of zero.[7] So this is a soft upper boundary in this case, even though the measure's range of (0,1) is not.[8] 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 **π** will never be reached, and the lower bound of

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

[8] 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.

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.

Finite Sample Distribution for ANY Dependence Measure, Under ANY Real-world Data Conditions

I present below the angles distributions for some of the dependence measures discussed above, under challenging, real-world data conditions (see Opdyke, 2024 for the application of NAbC to a large number of different data conditions). Briefly, the multivariate returns distribution of the portfolio in this case 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 (5) below and n=126 observations, for half a year of daily returns.[9]

(5)

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

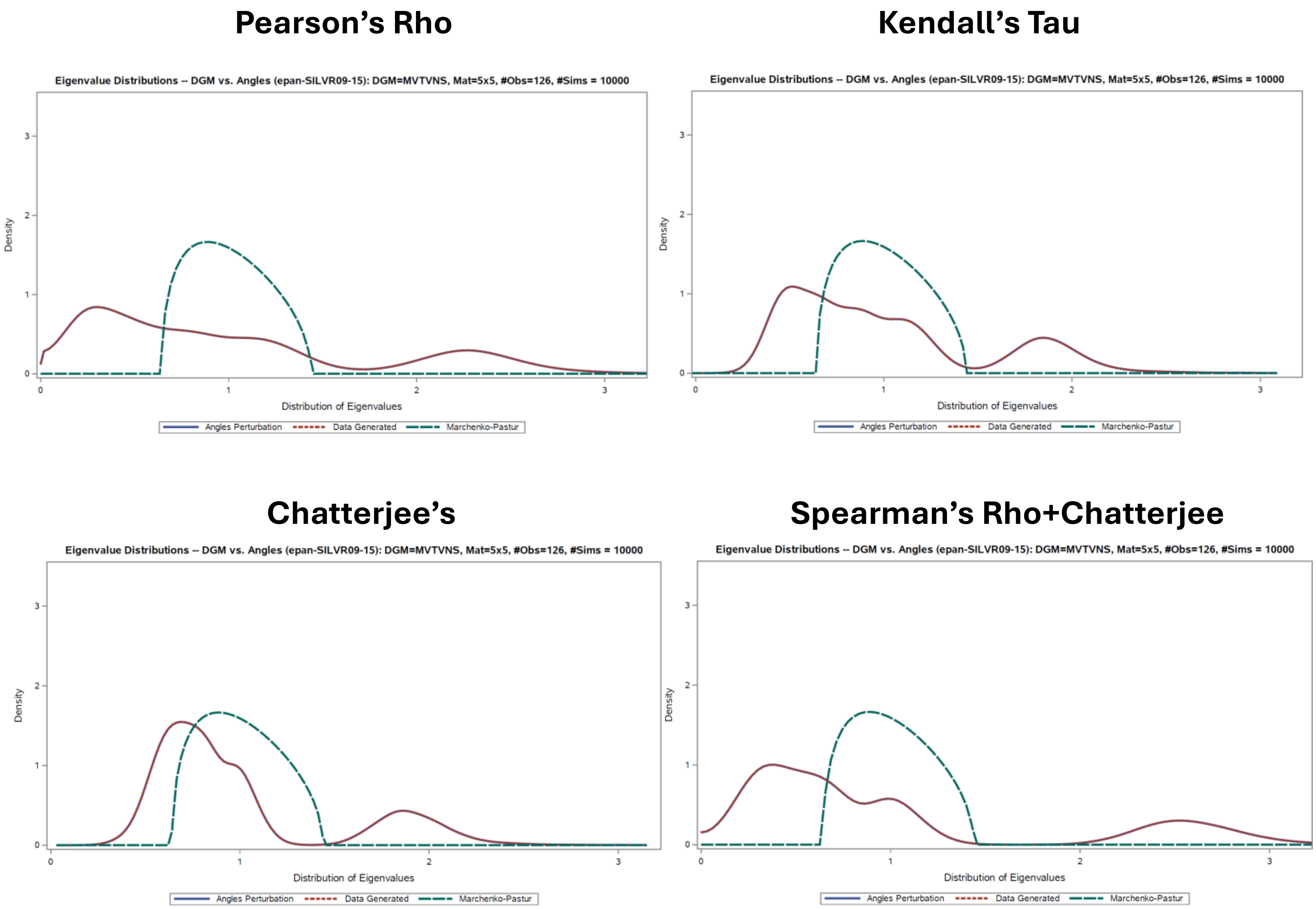
For verification purposes only, I compare those angles distributions based on the data simulation directly against those based on NAbC's angle kernels, and in all cases the results are empirically indistinguishable. The same is true for the spectral distributions, which I also present below against the 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,[10] 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

[9] 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).

[10] 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 1: Spectral Distribution-NAbC Angles Kernel v Data Simulations v Marchenko Pastur

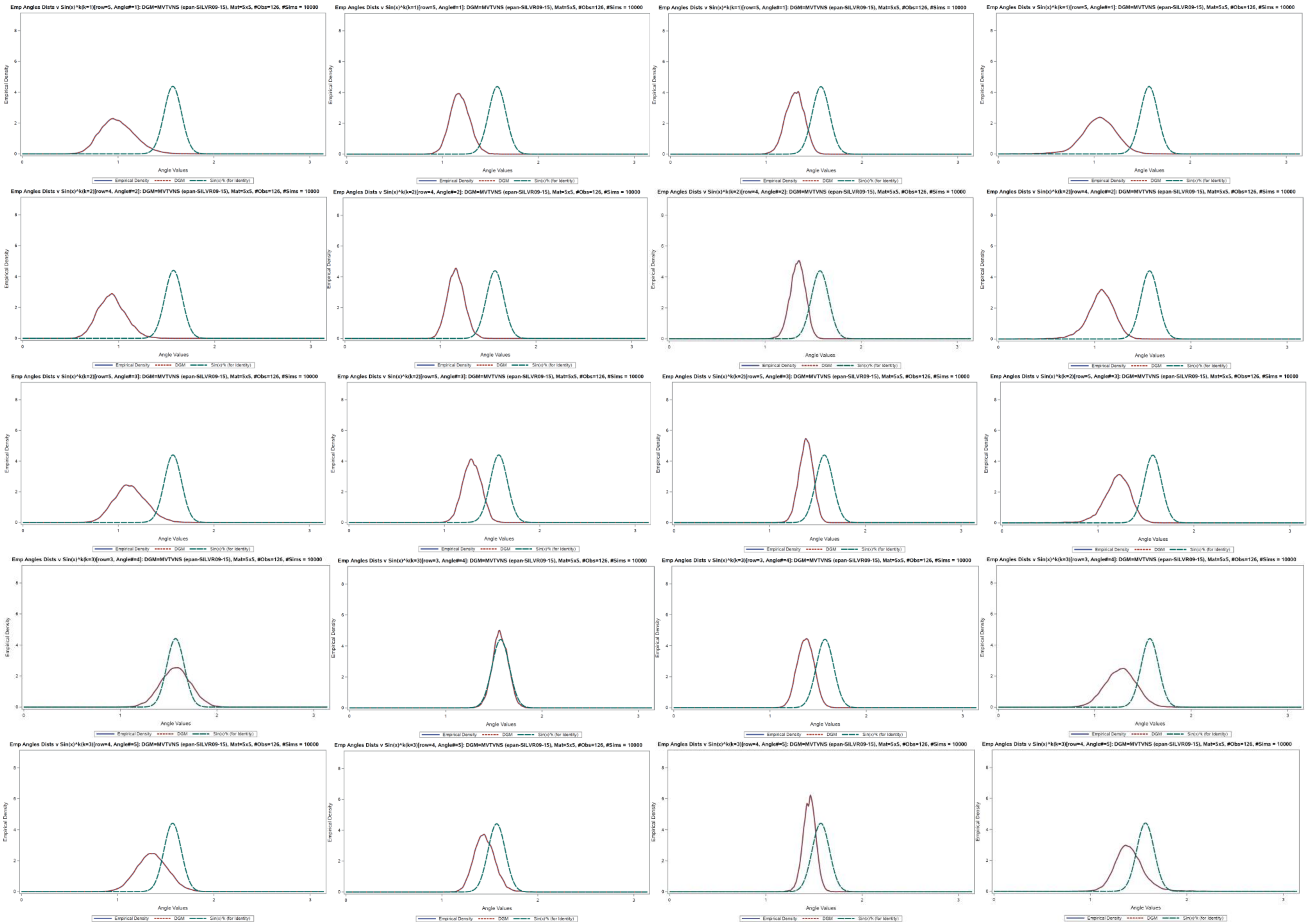


powerful under associations that are more monotonic, and the data conditions of this example fall more (but not entirely) into the latter category. The story changes a bit when we use the dependence measure suggested by Zhang (2023), which is essentially a maximum between Spearman's rho and Chatterjee's correlation, the 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.

Post 3 covers in detail NAbC's calculation of the dependence measure's cell level and matrix level p-values and confidence intervals, which I will not be repeat here because it is identical regardless of the dependence measure being examined (Post 3 covered only Pearson's matrix). However, I take the example from Post 3 for Pearson's matrix, which included such p-values and confidence intervals, and recreate it here using Kendall's Tau. Matrix input values necessarily are slightly different, but all other aspects of the example remain the same to demonstrate the apples-to-apples seamless application of

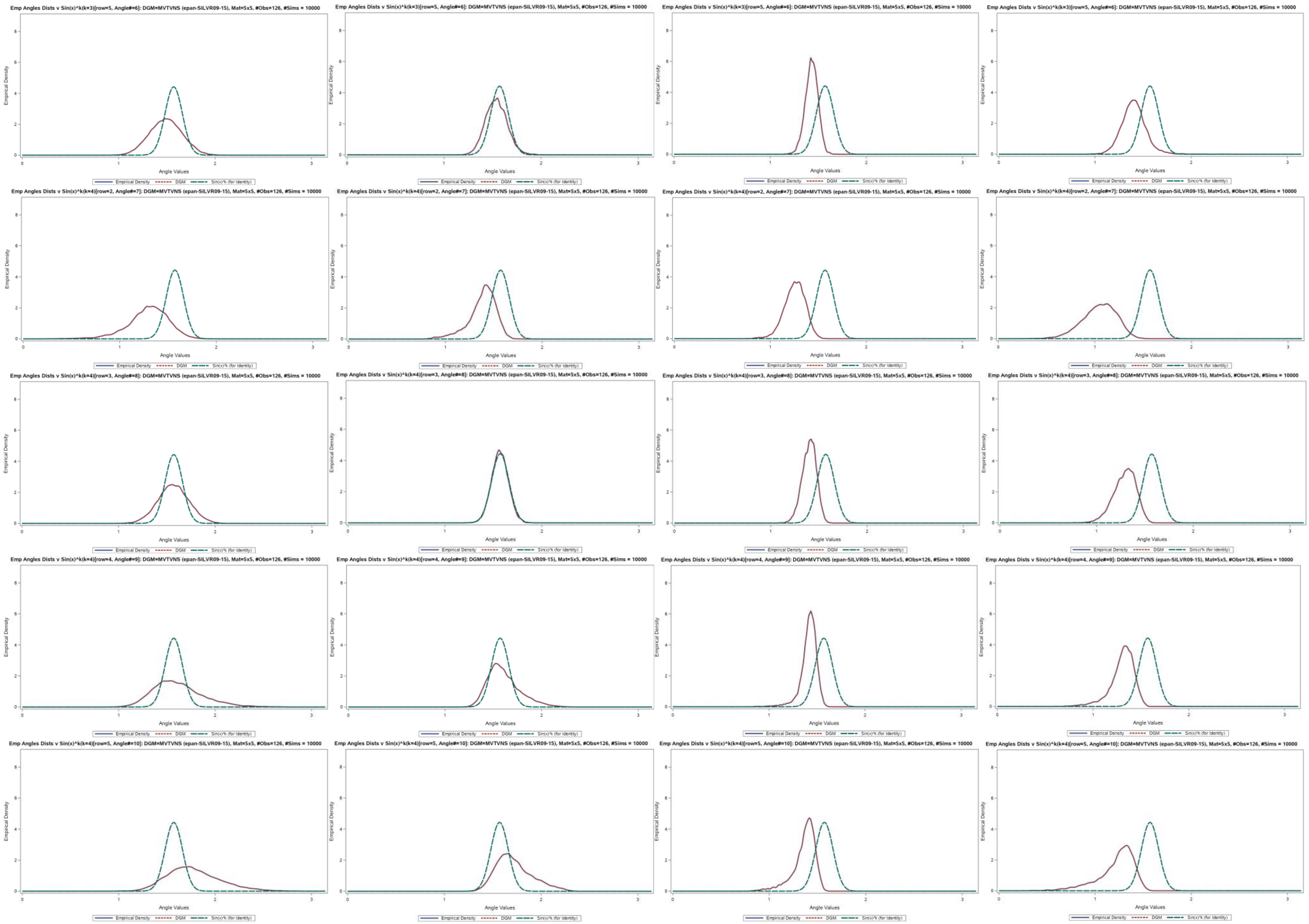
Graphs 1-5: Angles Distributions--NAbC Angles Kernel v Data Simulations v Identity Matrix

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



Graphs 6-10: Angles Distributions--NAbC Angles Kernel v Data Simulations v Identity Matrix

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



NAbC across different dependence measures. [11]

NAbC Applied with Kendall's: Unrestricted + Scenario-Restricted p-values and Confidence Intervals

Below I apply NAbC to obtain both p-values and confidence intervals, for Kendall's Tau, under two cases: unrestricted, and scenario-restricted. Solely for ease of replication, the data generating mechanism for these examples is simply multivariate standard normal, with N=25k simulations and number of observations n = 160.

<u>UNRESTRICTED CASE</u>: 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. **<u>Confidence Intervals</u>**: 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. **<u>Quantile Function</u>**: What is the unique correlation matrix associated with [B], a matrix of cumulative distribution function values associated with the corresponding cells of [A]?

Q3. **<u>p-values</u>**: 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]?

<u>SCENARIO-RESTRICTED CASE</u>: 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

[11] Values used here for Kendall's matrix were close to those obtained when translating from the Post 3 Pearson's example using $\tau = (2/\pi)\arcsin(r)$ where r = Pearson's, which is generally valid under elliptical data (which is one of the reasons I used multivariate Gaussian data here; see McNeil et. al., 2005).

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 (sencario-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]?

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.

Q1

1				
-0.0172	1			
-0.2100	0.0626	1		
0.0472	-0.3567	-0.1602	1	
0.2794	0.0926	-0.1873	-0.0830	1

1				
0.0250	1			
-0.1661	0.0986	1		
0.0926	-0.3131	-0.1427	1	
0.3210	0.1410	-0.1396	-0.0478	1

Q2

1				
0.1729	1			
-0.0355	0.2369	1		
0.2374	-0.1510	-0.0392	1	
0.4335	0.3159	0.0614	0.1040	1

1				
0.2735	1			
0.0910	0.3475	1		
0.3323	0.0127	0.1182	1	
0.5250	0.4370	0.2335	0.2789	1

1				
0.2300	1			
0.0424	0.3013	1		
0.2920	-0.0525	0.0570	1	
0.4929	0.3849	0.1611	0.2103	1

Q3

p-value=0.1503

0.0006				
0.0090	0.0218			
0.0222	0.0315	0.0269		
0.0170	0.0157	0.0088	0.0077	

Q4

1				
0.1282	1			
-0.0636	0.0478	1		
0.1942	-0.1940	-0.0639	1	
0.4098	0.1757	-0.1144	-0.0541	1

1				
0.1282	1			
-0.0636	0.0904	1		
0.1942	-0.1940	-0.0639	1	
0.4098	0.2028	-0.0789	-0.0184	1

Q5

1				
0.1282	1			
-0.0636	0.2398	1		
0.1942	-0.1940	-0.0639	1	
0.4098	0.2895	0.0362	0.0867	1

1				
0.1282	1			
-0.0636	0.3492	1		
0.1942	-0.1940	-0.0639	1	
0.4098	0.3425	0.1150	0.1841	1

1				
0.1282	1			
-0.0636	0.3006	1		
0.1942	-0.1940	-0.0639	1	
0.4098	0.3165	0.0809	0.1486	1

Q6

p-value=0.0436

	0.0047			
	0.0077	0.0148	0.0171	

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 – see Post 3). 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 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"

Although this has been covered in previous Posts, it bears repeating that, regardless of the dependence measure being used, NAbC remains "estimator agnostic," that is, valid for use with any reasonable estimator of that dependence structure. 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.

LNP: A Generalized Entropy for All Positive Definite Dependence Measures

The (two-sided) p-values NAbC provides (see Q3 and Q6 above, and Post 3 for details) actually can be used to construct a competitor to commonly used distance metrics, such as norms (e.g. Taxi, Frobenius/Euclidean, and Chebyshev norms: see Post 3 for definitions), and has a number of advantages over them in this setting. 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'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 (6) below:

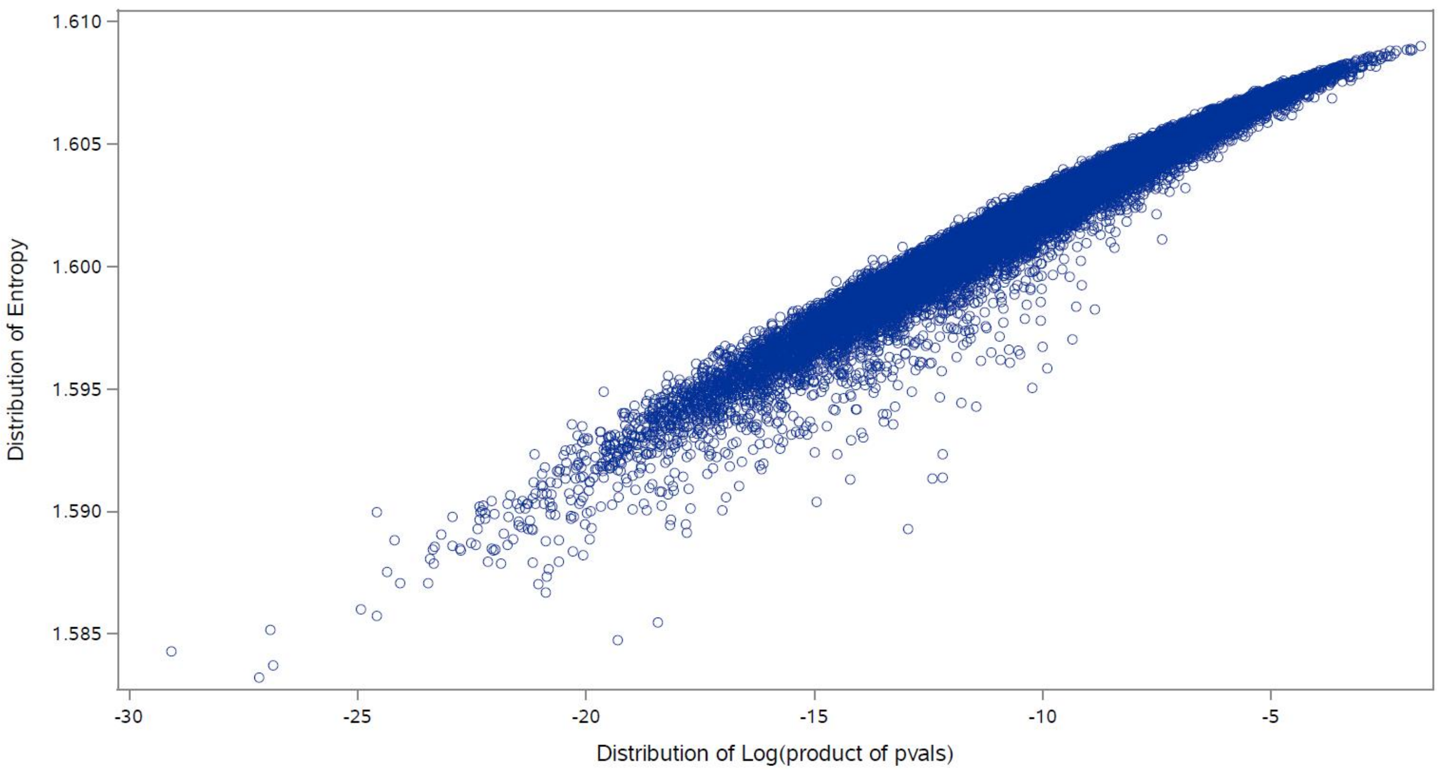
(6) $$\text{"LNP"} = \ln\left(\prod_{i=1}^{q} p\text{-}value_i\right) = \sum_{i=1}^{q} \ln\left[p\text{-}value_i\right] \text{ where } q = p\left(p-1\right)/2$$ and $p\text{-}value_i$ is 2-sided.

This was shown in Post 2, using a Pearson's correlation matrix under the (Gaussian) identity matrix, to have a very strong correspondence with the entropy of the correlation matrix, defined by Felippe et al. (2021 and 2023) as (7) below:

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

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 (7), like NAbC, is valid for ANY positive definite measure of dependence, not just Pearson's. Graph 12 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.

Graph 12: 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.[12] 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/independence/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, 2010, 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 both based on NAbC's cell-level (two-sided) p-values for the entire matrix. These are intriguing results with possibly far-reaching implications.

Conclusion

In Posts 1,2, and 3 I listed the seven characteristics of the full NAbC solution that, taken together, are shared by no other approach, and for completeness I list them again 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 (i.e. the confidence intervals for all the cells define

[12] 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.

that of the entire matrix, and the same is true for the p-values; this also 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, while the rest are allowed to vary, enabling granular and realistic scenarios

7. it 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, enabling reliable application in actual, imperfect, non-textbook settings

Post 2 provided, for Pearson's under the Gaussian identity matrix, an interactive spreadsheet that implements fully analytic p-values and confidence intervals

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

and combined with Post 3, both cover all but 2. in the list above. This Post 4 covers 2. above, expanding NAbC's range of application to ALL positive definite measures of dependence, with any values, under all real-world data conditions.

The objective of this work has been to provide a method that checks all of these boxes – 1. Through 7. – simultaneously, which is what is required for useful and useable portfolio analytics in real-world, non-textbook settings. 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.

With NAbC, we now have a powerful, applied approach enabling us to treat an extremely broad class of widely used dependence measures just like the other major parameters in our (finite sample) financial portfolio models. We can use NAbC in frameworks that identify, 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.

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.

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.

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.

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.

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

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

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.

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]

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

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.

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

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.

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

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

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.

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].

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.

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.

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.

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.

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.

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

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.