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 3 of 4: Pearson's Under ANY Values and Real-World Financial Data Conditions

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 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 third 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 3 expands NAbC's application from Pearson's under the Gaussian Identity Matrix to Pearson's under ANY correlation values and challenging, real-world financial data conditions. Post 4 will expand its range of application even further, beyond Pearson's to ANY positive definite dependence measure.

[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) 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, which is highly flexible, and allows for more precise attribution and intervention analyses.

POST 1: NAbC introduced.
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.

Review of Post 2: Correlations and Angles

To briefly review from Post 2, 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 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):

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

(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.

[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 1) 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]:

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

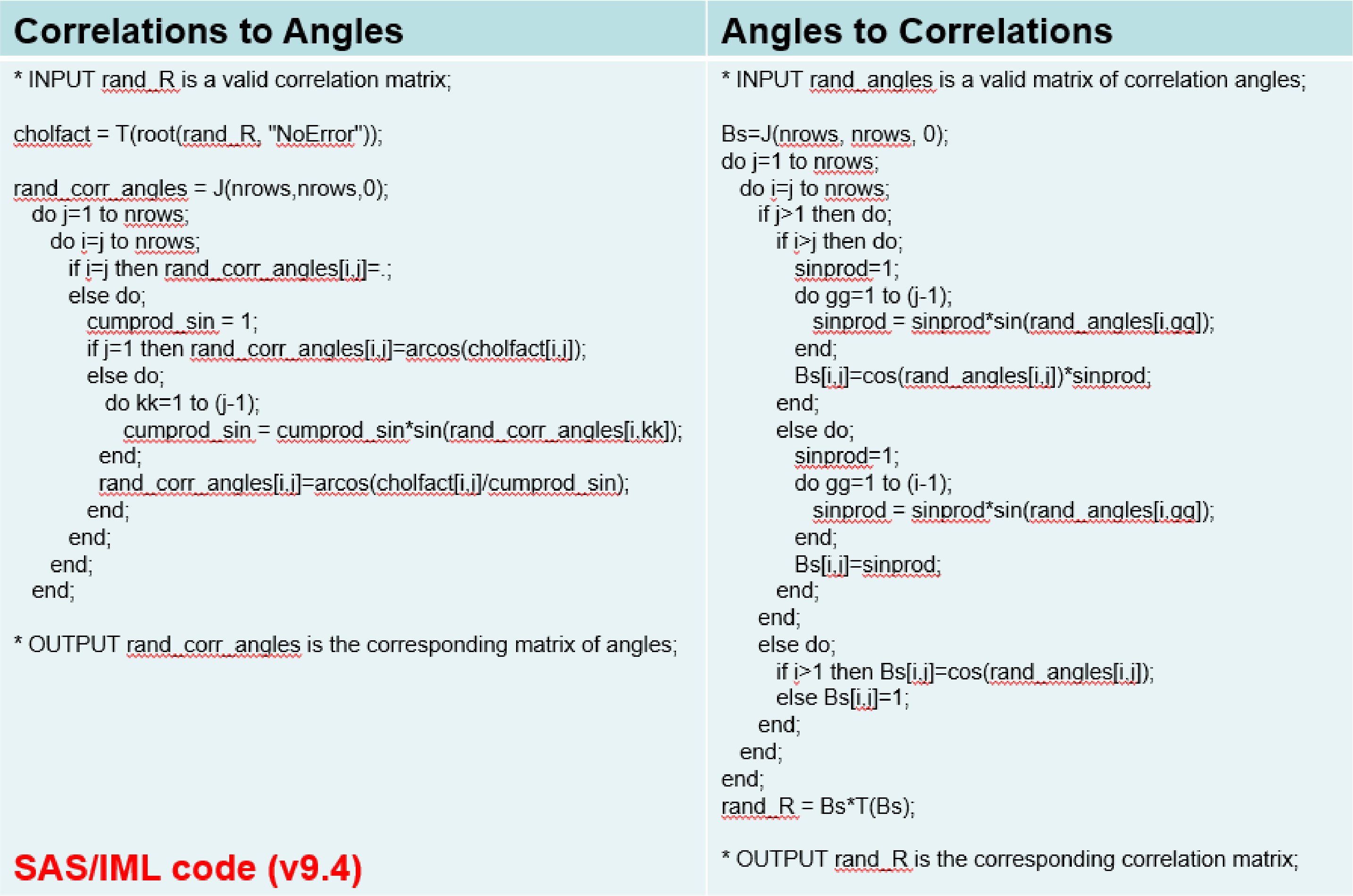
(2.b) To obtain an individual correlation, $r_{i,j}$, we have, simply from $R = BB^T$:

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

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:

[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;
```

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.

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.

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 below, I present some empirical examples of this in graphs below under real-world financial data conditions.

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

Beyond the Gaussian Identity Matrix: Pearson's Distribution for Any Values, Any Data

Recall from Post 2 that the only requirement for the bi-directional, one-to-one relationship between correlations and angles is that the correlation matrix be symmetric positive definite,[6] which, numerical issues aside,[7] is always the case for Pearson's product moment correlation matrix regardless of its values.[8] Consequently, we are not restricted to the Gaussian identity matrix if we want to use angles and their distributions to define the finite sample distribution of Pearson's matrix. Under the Gaussian identity matrix in Post 2, I first derived the straightforward, analytic distributions of the angles, presenting together their pdf's, cdf's, and quantile function (although the cdf and quantile function previously were claimed to be analytically intractable – see Makalic and Schmidt, 2018). Second, because of A. through D. above, I showed how it is straightforward to use these angles distributions to sample the correlation matrix and, more importantly, to define its finite sample distribution. This is used to obtain both cell-level and matrix-level p-values and confidence intervals that maintain analytic consistency across the two

[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 examples).

[6] See Pinheiro and Bates (1996), Rebonato and Jackel (2000), Rapisarda et al. (2007), Pouramadi and Wang (2015), and Cordoba et al. (2018). I discuss in Post 4 that this requirement of symmetric positive definiteness is true for any dependence measure, not just Pearson's.

[7] Below I discuss how angles distributions are more stable and robust than spectral distributions by several criteria, including numerically, especially as dependence matrices approach singularity, which arguably is the rule rather than the exception for non-small, real-world investment portfolios.

[8] Note that for Pearson's specifically, the first and second moments (mean and variance) of the distributions of the returns must exist.

levels.[9] The only difference between the Gaussian identity matrix and the more general case covered in this Post 3 – any correlation values under any real world financial data – is the angles distributions themselves: all other relationships (ie angles to correlations and correlations to angles) and conditions (A. – D.) hold. So all we need are the angles distributions under general conditions to obtain the distribution of Pearson's matrix under general correlation value and data conditions.

Angles Distributions for ALL Pearson's Values, for ANY Real-World Financial Data Conditions

Currently, the extant literature does not provide analytic forms for the angles distributions under general conditions. Deriving these appears to be a non-trivial problem. Spectral (eigenvalue) distributions, which many researchers turn to in this setting, have been shown to vary dramatically when data is characterized by different degrees of heavy-tailedness (see Burda et al., 2004, Burda et al., 2006, Akemann et al., 2009; Abul-Magd et al., 2009, Bouchaud & Potters, 2015, Martin & Mahoney, 2018; and Opdyke, 2024), as well as by different degrees of serial correlation (see Burda et al., 2004, 2011, and Opdyke, 2024), and the literature provides no general analytic form under general, real-world financial data conditions – certainly nothing that is analogous to convergence to the Marchenko-Pastur distribution under iid independence (Marchenko and Pastur, 1967).[10] If angles distributions are of similar complexity, deriving their general analytic form under general conditions, if possible, currently appears to be a non-trivial, unsolved problem.

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

Structurally: Aggregation level becomes relevant and important here. For a given correlation matrix *R* there are many more angles distributions than there are spectral distributions (i.e. p(p-1)/2 vs p, a factor of (p-1)/2 more). As a matrix approaches singularity (non-positive definiteness (NPD)), which arguably is the rule rather than the exception for non-small investment portfolios, a much smaller ***proportion*** of angles distributions will approach degeneracy than is true for eigenvalue distributions. Consequently, the overall construction of the correlation matrix via $R = BB^T$ generally will remain much more stable than one based on an eigen-decomposition of $R = V\Lambda V^{-1}$ where *V* is a matrix with column eigenvectors and Λ is a diagonal matrix of the corresponding eigenvalues.

[9] I describe below how this cell-level and matrix-level consistency is critical for attribution analyses specifically, not to mention correct inferences generally.

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

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

Distributionally: As shown graphically below under challenging, real-world financial data conditions, the angles distributions are relatively "well behaved," both in the general sense and relative to spectral distributions. They are relatively smooth and typically unimodal, and clearly bounded on $\theta \in (0, \pi)$. Spectral distributions, based on the same data, very often are spikey[11] and highly multimodal, and their unboundedness (in the general case) translates into larger variances and less tail accuracy. Simply put, they typically are much more complex and challenging to estimate precisely and accurately compared to individual angles distributions for a given correlation matrix *R* under real-world financial data.

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

So given the useful characteristics of the angles distributions (on both a general basis and relative to the alternative of spectral distributions), not to mention the fact that they remain at the right level of aggregation for granular analysis of the correlation matrix, we are able to proceed WITHOUT their analytic derivation: rather, we can use a time-tested nonparametric approach, such as kernel estimation, to reliably define them. All this requires is a single simulation (say, N=10,000) based on the known or well-estimated correlation matrix, and the known or well-estimated data generating mechanism. Then, after translating all N simulated correlation matrices to N matrices of angles, we fit a kernel to each empirical angle distribution, i.e. the empirical distribution of each angle for each cell of the matrix. We now have not only the densities of all the individual angles, but also the multivariate density of the matrix, which is

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

just the product of all the individual densities due to their independence per B. above. Note that this goes in both directions: we can perform 'look-ups' on the empirically defined distribution to obtain the cdf value(s) corresponding to particular angle value(s), or use cdf value(s) to 'look up' corresponding angle (quantile) value(s). This process is described step by step below.

1. Simulate samples (say, N=10k) based on the specified/known or well estimated correlation matrix and the specified/known or well estimated data generating mechanism.
2. Calculate the corresponding N correlation matrices, and their Cholesky factorizations, and transform each of these into a lower triangle matrix of angles (as described above in (2.a)).
3. Fit kernel densities to each of the p(p-1)/2 empirical angle distributions, each having N observations.
4. Generate samples based on the densities in 3.[12]
5. Convert the samples from 4. back to a re-parameterized Cholesky factorization, and then multiply by its transpose to obtain a set of N validly sampled correlation matrices (as described above in (2.b)). Positive definiteness is enforced automatically as the Cholesky factor places us on the **unit** hyper-hemisphere.

The distribution of correlation matrices from 5. is identical to that of 2., but after the kernel densities are fit once in 3., generating samples in 4. is orders of magnitude faster than relying on direct simulations in steps 1. and 2. And of course, using 3.-5. rather than 1. and 2. allows for correct probabilistic inference both at the cell level and at the matrix level, due to the independence of the angles distributions (remember the correlations themselves are NOT independent!) and subsequently, the proper transformation of the angles to correlations. This reliance on the angles, and their subsequent transformation to correlations, allows us to isolate specifically the distribution of the entire correlation matrix, for probabilistic inference, without touching any other distributional aspect of the data, which is the point of the methodology.

So this framework is essentially identical to that for the specific case of the Gaussian identity matrix derived in Post 2, the only difference being it is based on nonparametrically defined, as opposed to analytically defined, angles distributions. Before covering implementation details below, I show some examples of graphs of the angles distributions and the corresponding spectral distribution under real-world financial returns data. The multivariate returns distribution of the portfolio is generated based on the t-copula of Church (2012), with p=5 assets, varying degrees of heavy-tailedness (df=3, 4, 5, 6, 7), skewness (asym parm=1, 0.6, 0, -0.6, -1), non-stationarity (std dev=3σ, σ/3, σ; n/3 obs each), and serial correlation (AR1=-0.25, 0, 0.25, 0.50, 0.75), with a block correlation structure shown in (3) below and n=126 observations.[13] The spectral distribution is compared against Marchenko-Pastur as a baseline.

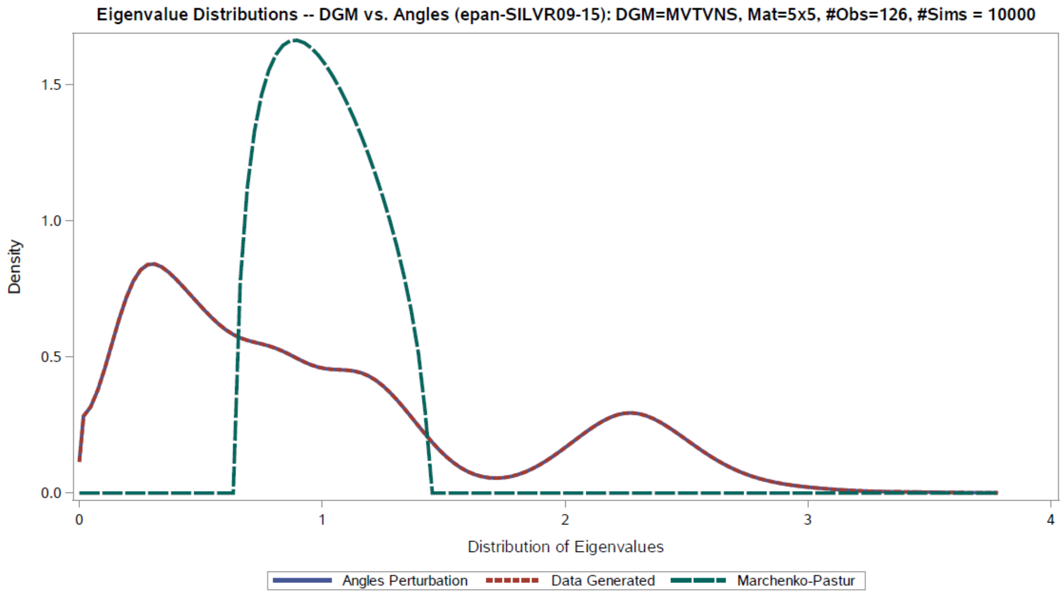
[12] Algorithms for sample generation from broadly used kernels (e.g. the Gaussian and Epanechnikov) are widely known. An example of the latter is simply the median of three uniform random variates (see Qin and Wei-Min, 2024).

[13] 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).

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

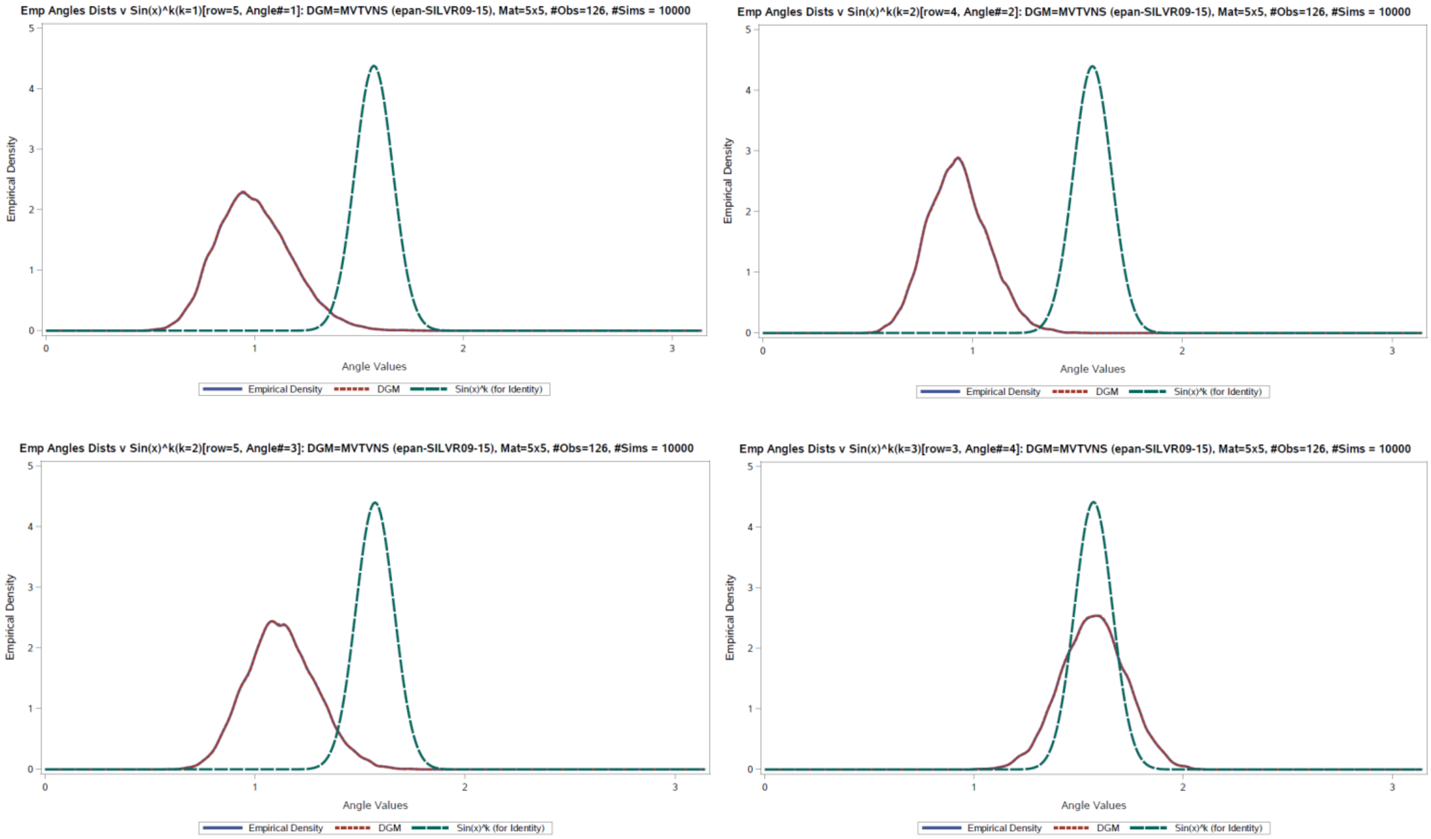
(3)

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



Eigenvalue Distributions -- DGM vs. Angles (epan-SILVR09-15): DGM=MVTVNS, Mat=5x5, #Obs=126, #Sims = 10000

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

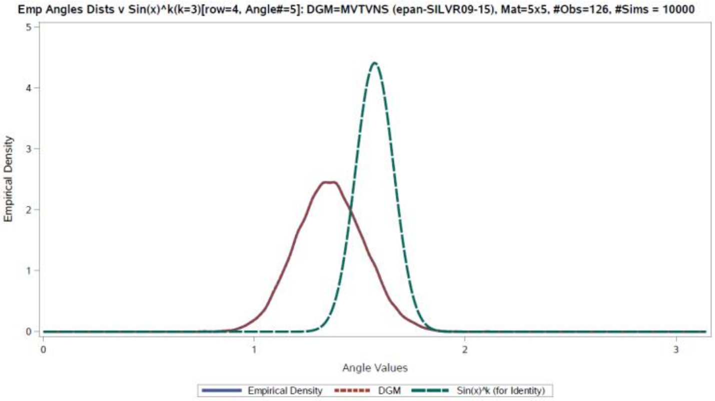


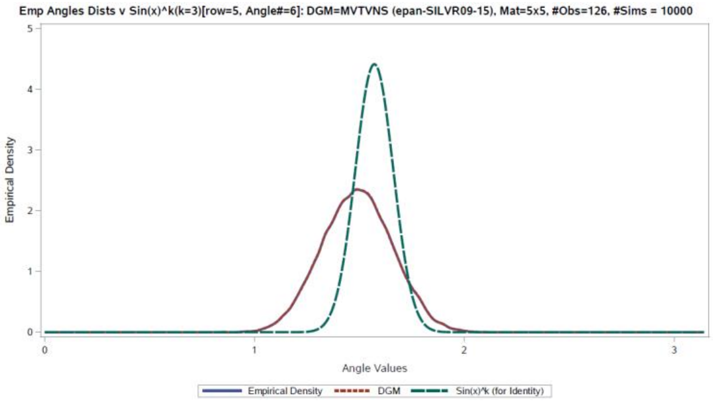
Emp Angles Dists v Sin(x)^k(k=1)[row=5, Angle#=1]: DGM=MVTVNS (epan-SILVR09-15), Mat=5x5, #Obs=126, #Sims = 10000

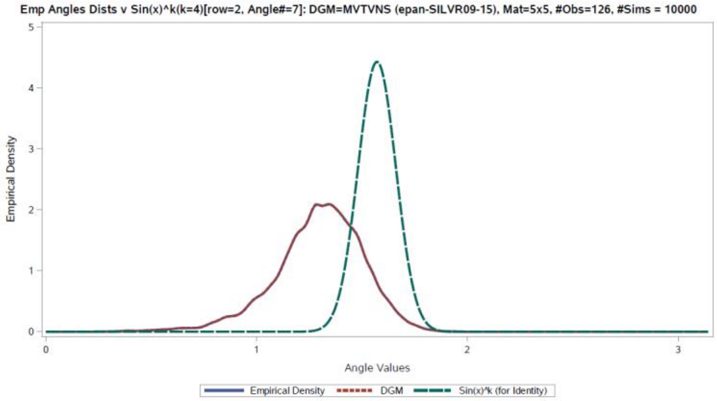
Emp Angles Dists v Sin(x)^k(k=2)[row=4, Angle#=2]: DGM=MVTVNS (epan-SILVR09-15), Mat=5x5, #Obs=126, #Sims = 10000

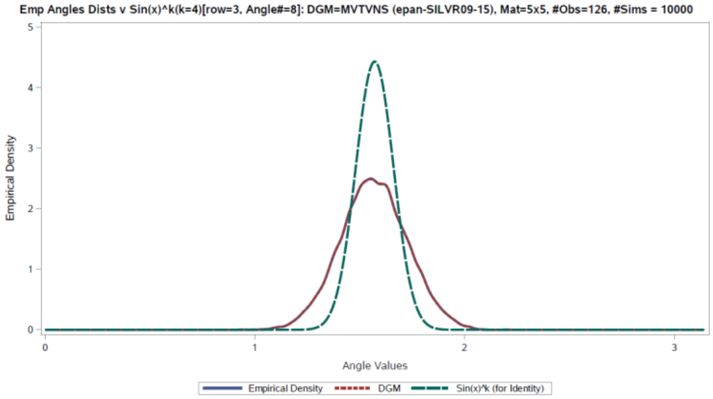
Emp Angles Dists v Sin(x)^k(k=2)[row=5, Angle#=3]: DGM=MVTVNS (epan-SILVR09-15), Mat=5x5, #Obs=126, #Sims = 10000

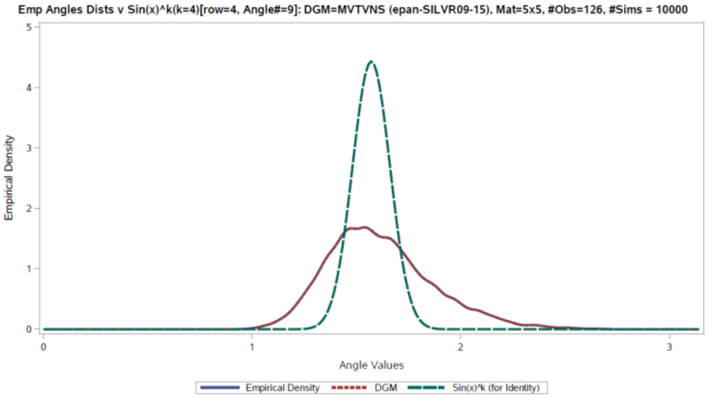
Emp Angles Dists v Sin(x)^k(k=3)[row=3, Angle#=4]: DGM=MVTVNS (epan-SILVR09-15), Mat=5x5, #Obs=126, #Sims = 10000

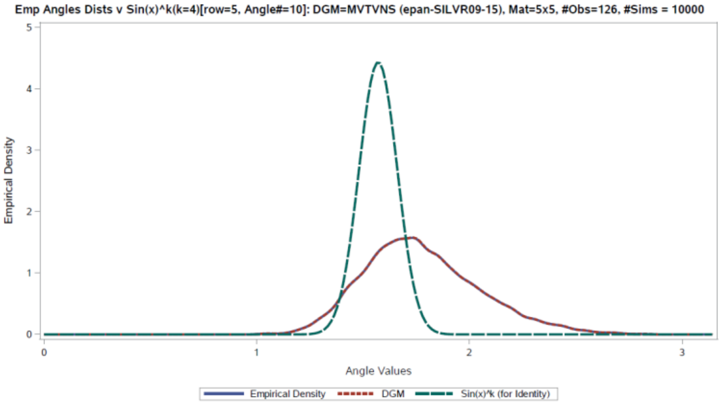












Several points are worth noting and reemphasizing from these graphs. First, the graphs of the angles distributions contain three densities: A. one based on angles perturbation (i.e. sampling from the fitted kernel) as described above in steps 3.-5., B. one based on direct data simulations (steps 1.-2.), and C. the analytical density under the (Gaussian) identity matrix as a comparative baseline. Note that the only reason I include B. is to demonstrate the validity of A, and as expected, the angles distributions from A. and B. are empirically identical (with A. being orders of magnitude faster and more computationally efficient). The spectral distributions based on the samples generated in both A. and B. also are identical, as are a wide range of additional analyses not presented herein (e.g. various norms, VaR-based economic capital, and 'generalized entropy' as described below). This empirically validates that the angles-perturbation approach is an efficient and correct method for isolating and generating the density of the correlation matrix, and unlike steps 1. and 2., one that preserves inferential capabilities. In other words, these results empirically validate that the angles contain all extant information regarding dependence structure here (see Fernandez-Duren & Gregorio-Dominguez, 2023, and Zhang & Songshan, 2023, as well as Opdyke, 2024).

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

Finally, as described above, note the multi-modal and unbounded nature of the spectral distribution for this portfolio compared to the angles distributions, where the biggest thing approaching an estimation challenge is a slight asymmetry. But this speaks only to estimation issues. More notable is the fact that on a cell-by-cell basis, the angles distributions deviate materially i. not only from central values of **π**/2, and less dramatically from perfect symmetry, when compared to their (analytic) distributions under the (Gaussian) identity matrix, but also ii. from each other! Each angle’s distribution can vary quite notably compared to the other angles’ distributions, especially under smaller sample sizes. There simply is no way that a single spectral density, even if perfectly estimated, will be able to capture and reflect all the richness of dependence structure reflected here at the granular level of the pairwise cells, for any useful purposes, including cell-level attribution analyses, granular scenario and reverse scenario analyses, cell-level intervention ‘what if’ analyses, and customized stress testing, let alone precise and correct inference at either the cell level OR the matrix level. I now stop beating this drum[14] and leave comparisons to spectral distributions behind to cover implementation issues below.

Nonparametric Kernel Implementation

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

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

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

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

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

$h=1.06\cdot\hat{\sigma}\cdot N^{-1/5}$, $h=0.79\cdot\text{IQR}\cdot N^{-1/5}$, $h=0.9\cdot\min\left(\text{IQR}/1.34,\hat{\sigma}\right)\cdot N^{-1/5}$, and $h=2.34\cdot\hat{\sigma}\cdot N^{-1/5}$ for Epanechnkov only, where $\hat{\sigma}=$ sample standard deviation and $\text{IQR}=$ sample interquartile range. As with virtually all kernel implementations, the choice of kernel matters less than the choice of bandwidth, although in this setting, across a broad range of data conditions and correlation values, the Epanechnikov kernel appears to perform slightly 'better' (i.e. with slightly less variance, thus providing slightly more statistical power) than the Gaussian, perhaps because its sharp bounds require reflection at the boundary less often than the Gaussian kernel. The bandwidth that appears to perform best across wide-ranging conditions is $h=0.9\cdot\min\left(\text{IQR}/1.34,\hat{\sigma}\right)\cdot N^{-1/5}$. Additionally, for larger matrices (e.g. p=100), bandwidths need to be tightened by multiplying *h* by a factor of 0.15. When there are many cells (e.g. for p=100, #cells= p(p-1)/2=4,950) this tightening avoids a slight drift in metrics that are aggregated across all the cells (e.g. correlation matrix norms, spectral distributions, and LNP (a type of 'generalized entropy' defined below)). Multiplying by this factor for smaller matrices does not adversely affect the density estimation in any way, so this factor always is used. For matrices much larger than p=100, a further tightening of this factor may be required, and this is readily determined by empirical testing of the aggregated metrics of interest.

This application of a nonparametric kernel for density estimation is straightforward and very well established in the literature, as are algorithms to generate samples from them (see Qin and Wei-Min, 2024). And as shown in the graphs above, when fitted to angles distributions and used as a basis for subsequent angle perturbation, it generates angles distributions, and corresponding correlation matrices and spectral distributions, that all are empirically identical to those based on direct data simulations, thus confirming the utility and appropriateness of this approach in this setting.

Analytically Consistent Cell-Level and Matrix-Level p-values and Confidence Intervals

Once the kernels have been estimated and the angles distributions generated by perturbing/sampling based on those kernels, the p-values and confidence intervals for both the individual correlation cells and the entire correlation matrix are the same as those derived in Post 2 for the Gaussian identity matrix. The only difference, aside from their now-nonparametric basis, is that the angles distributions are no longer symmetric by definition, as is true under the (Gaussian) identity matrix. This can be seen in the graphs of the angles distributions provided above. The p-value calculation, however, remains very straightforward, and it requires just a bit of care to properly account for asymmetry. The one-sided p-value remains simply (3):

(3) one-sided p-value = $F_X\left(x;k\right)$ or $1-F_X\left(x;k\right)$ for lower and upper tails, respectively,

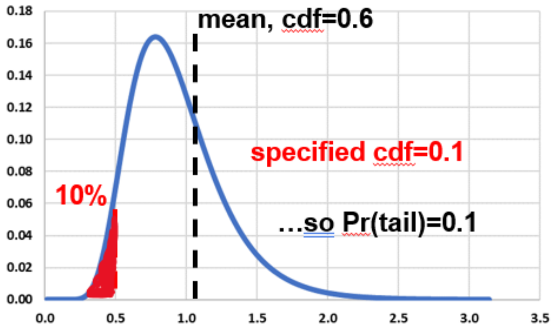
where k = n – column# – 2

However, due to possible (probable) asymmetry, the two-sided p-value is different, requiring first the calculation of the empirical mean correlation matrix from the simulations in step 2. above. This mean correlation matrix is then translated into a matrix of angles, and we obtain the empirical cdf's corresponding to these "mean angles" with a "look-up" on the entire angles distributions generated in step 4. These cdf's will be close to 0.5 when the angles distributions are close to symmetry, and they will deviate from 0.5 under asymmetry. The two-sided p-values are based on the distance between the cdf's of each of the angles of the specified correlation matrix being 'tested' and those of the "mean angles," where 'distance' is the integrated density (probability), not distance of the angle size on the x-axis. Specifically, the two-sided p-values are the sum of the probability in the tails BEYOND this distance.[15] Formulaically this is simply (4):

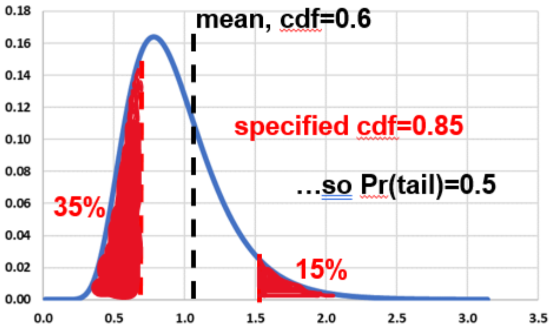
(4) two-sided p-value = max[0, Mcdf – d] + max[0, 1 – (Mcdf + d)], where
d = abs(Mcdf – cdf), Mcdf = mean angle cdf, cdf = cdf of specified angle

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

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



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



Note that while cdf=0.1 is hardly more 'extreme' than cdf=0.85 in absolute terms, relative to the mean angle cdf=0.6, it is twice as 'extreme,' i.e. twice as far probabilistically from the mean cdf=0.6, with a probabilistic distance of 0.5 for Graph 11, and 0.25 for Graph 12. Moreover, a value as extreme as the case of Graph 11 is associated with only 1/5 the probability of being observed compared to that of Graph 12 (compare the red shaded areas). This example demonstrates why asymmetry must be properly taken

[15] So this 'distance' is 0.5 for Graph 1 and 0.25 for Graph 2.

into account in this setting, but the two-sided p-value still remains a very straightforward calculation, and the "mean angles" matrix is used for additional, important purposes below, as discussed in the Scenarios section.

Cell-level confidence intervals simply are calculated as in (5), which automatically takes asymmetry into account. Asymmetry notwithstanding, this is identical to the same calculation under the (Gaussian) identity matrix.

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

k = n – column# – 2

The above describes p-values and confidence intervals at the cell level, i.e. for every cell in the correlation matrix, individually. The p-value and confidence interval(s) at the matrix level are based directly on these cell-level calculations and remain otherwise identical to those calculated in Post 2 under the (Gaussian) identity matrix. The matrix-level p-value, again, 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 (6).[16]

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

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 (6) because under independence, there is no dependence structure for them to exploit. So the straightforward calculation above in (6) is, by definition, the most powerful for FWER control.

Similarly, just as under the (Gaussian) identity matrix, calculation of the confidence interval for the entire matrix (7) 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.'

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

These critical alphas, when inserted as values in the empirically-based cdf 'look-up' 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 specified correlation matrix being 'tested,'

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

and is no longer strictly only the identity matrix. Independence of the angles distributions again makes these simultaneous confidence intervals very straightforward to calculate.

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

Flexible Scenarios, Reverse Scenarios, and Customized Stress Tests

NAbC is a 'bottom up' approach to defining the finite-sample distribution of the correlation matrix, based on the distributions of the individual correlation cells. In addition to analytic consistency, this provides a flexibility in scenario definition and scenario analytics that other approaches cannot match. Correlation (dependence) matrices under a tech market bubble (2000) vs those under a housing bubble (2008) vs those under Covid (2020) will change very different individual cells, and very different combinations of cells, in very different ways, often in terms of both direction and magnitude, while leaving many cells strongly affected under one upheaval completely unaffected under another. In other words, while correlation 'breakdowns' will occur under all of these extreme conditions, the granular nature of pairwise association matrices ensures that the fundamentally different nature of these breakdowns will be captured and reflected empirically in all related analyses. The only way to flexibly and realistically model this is at the most granular level – that of the individual correlation cells.

Fortunately, as described in detail in Post 2, NAbC allows for this, with full inferential powers under any definable scenario within the framework of pairwise associations defined by a correlation (dependence) matrix. Without repeating this description in detail, this is made possible by exploiting 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, which allows arbitrarily chosen cells to vary and the rest to remain constant/unaffected by the scenario, without violating positive definiteness. 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 effectively in this setting).

And while NAbC covers inference for a matrix of all pairwise associations, the same level of flexibility and sophistication exists in their estimation and simulation via, for example, vine copulas (see Czado and

Nagler, 2022).[17] So rather than imposing unrealistic restrictions, the framework of all pairwise associations is greatly liberating in its analytics, whether the focus is on inference, estimation, and/or (synthetic) scenario simulation. This all will be explored further in Post 4, which expands NAbC's range of application to dependence measures beyond Pearson's.

In this Post 3, however, one difference in scenario definition and implementation, compared to that of Post 2, is relevant for these scenario analytics: when allowing only selected cells of the correlation matrix to vary for a given scenario, while holding the remaining cells constant, we must insert angle values into the 'frozen' cells that not only hold the correlations constant, but also enforce positive definiteness. Where do we get those values? From the "mean angles" matrix defined above. As the mean of N=10,000 simulations, this is a stable and robust estimator of the correlation matrix under the (simulated) null hypothesis.[18] It will both hold constant the correlation values at their means, and it is itself positive definite, as it is based on a linear combination of positive definite matrices. We simply perform steps 1.-5. as usual, but after step 4. we overwrite values of the simulated angles in those cells to be held constant with the mean angle values, so that the 'frozen cells' in every simulation from 4. contain their respective (constant) mean values. The resulting scenario-restricted correlation matrices always will be positive definite, because the systematic row and column sorting for the scenario (described in detail in Post 2) in effect 'disables' systematic changes to these cells if we do not change the angle values, and their values will be 'frozen' at their mean correlation values. We did not need to do this under the (Gaussian) identity matrix because the mean values all were zero, by definition, and because simulation is not required in this case.

As mentioned above and described in detail in Post 2, this provides an unmatched degree of flexibility for scenario definition and scenario analytics: this approach works for ANY scenario restricted matrix within the framework of all pairwise associations, because the distributions of both the entire matrix and the individual cells are invariant to row and column order. Post 2 describes in detail how simply re-sorting the matrix allows for the selection of only specific cells that are meant to vary for a particular scenario, while maintaining proper inferential properties for both the individual cells and the entire matrix. Examples applying NAbC for inference on both unrestricted and scenario-restricted matrices are presented below.

[17] Even though vine copulas are a sophisticated and highly flexible method to define, estimate, and simulate bivariate dependence structures, NAbC (from Post 2) replaces earlier attempts to define and sample the gaussian identity correlation matrix using these methods, as they are unnecessarily computationally demanding and complex for this purpose (see Lewandowski et al., 2009 and Kurowicka, 2014) when a fully analytic solution exists, as shown in Post 2.

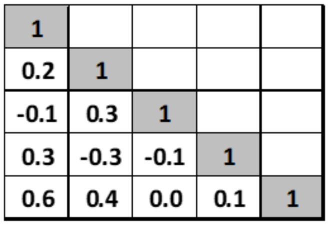
[18] The degree of empirical accuracy attained when using these means is based directly on the number of simulations. This can be seen in the scenario-restricted results (specifically, in the 'frozen' cells of the scenario-restricted matrices) in the next section.

NAbC Applied: Unrestricted and Scenario-Restricted p-values and Confidence Intervals

Below I apply NAbC to obtain both p-values and confidence intervals 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.

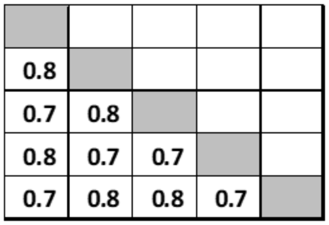
UNRESTRICTED CASE: Given a specified or well-estimated correlation matrix [A], and its specified or well-estimated data generating mechanism:

[A]



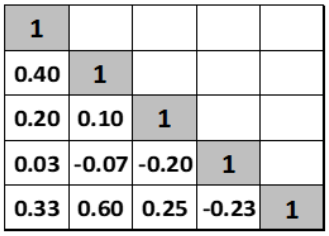
1				
0.2	1			
-0.1	0.3	1		
0.3	-0.3	-0.1	1	
0.6	0.4	0.0	0.1	1

[B]



0.8				
0.7	0.8			
0.8	0.7	0.7		
0.7	0.8	0.8	0.7	

[C]



1				
0.40	1			
0.20	0.10	1		
0.03	-0.07	-0.20	1	
0.33	0.60	0.25	-0.23	1

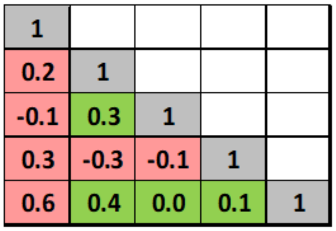
Q1. **Confidence Intervals**: What are the two correlation matrices that correspond to the lower– and upper–bounds of the 95% confidence interval for [A]? What are, simultaneously, the individual 95% confidence intervals for each and every cell of [A]?

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

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

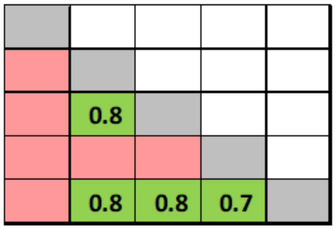
SCENARIO-RESTRICTED CASE: Under a specific scenario only selected pairwise correlation cells of [A] will vary (green), while the rest (red) are held constant, unaffected by the scenario (e.g. COVID). This is matrix [D].

[D]



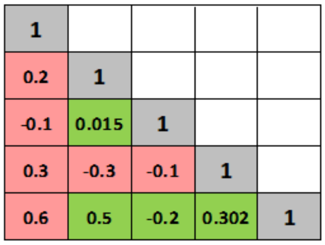
1				
0.2	1			
-0.1	0.3	1		
0.3	-0.3	-0.1	1	
0.6	0.4	0.0	0.1	1

[E]



🟥				
🟥	0.8			
🟥	🟥	🟥		
🟥	0.8	0.8	0.7	

[F]



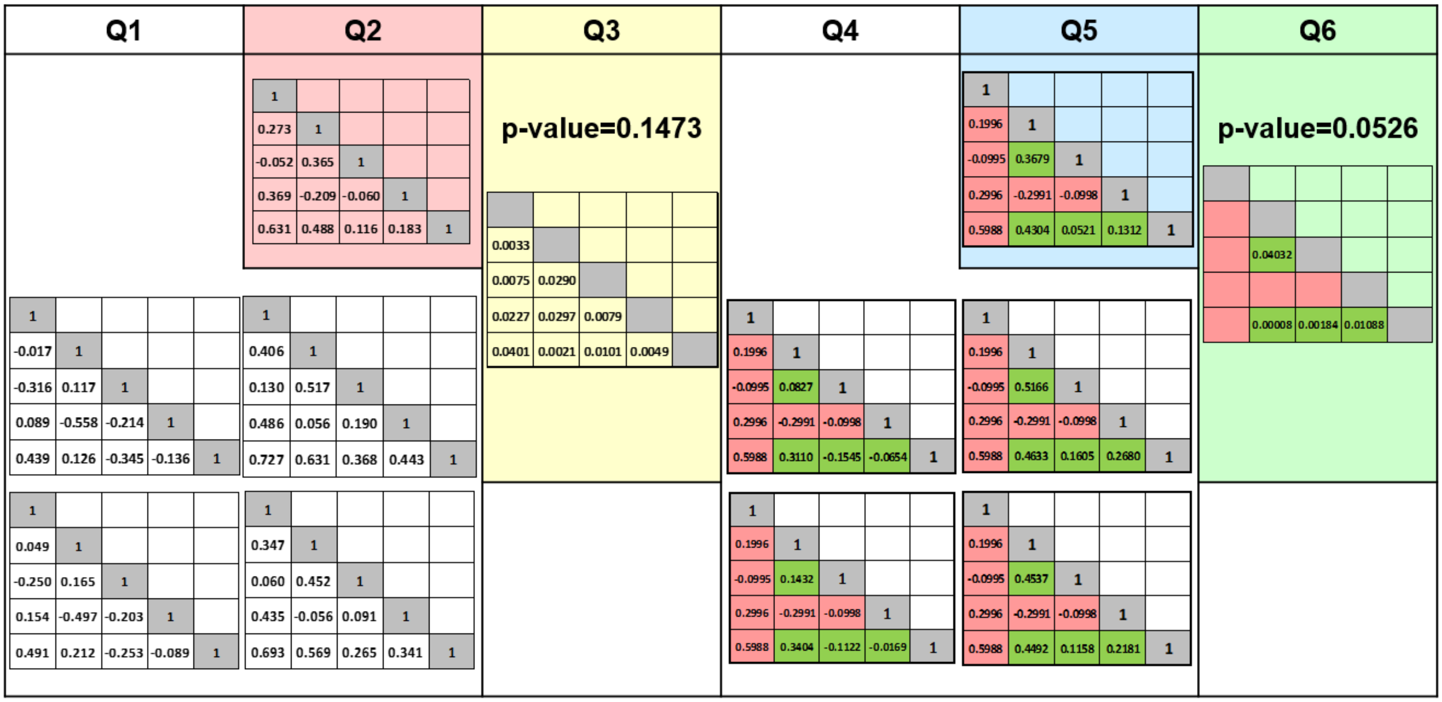
1				
0.2	1			
-0.1	0.015	1		
0.3	-0.3	-0.1	1	
0.6	0.5	-0.2	0.302	1

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

Q2

1				
0.273	1			
-0.052	0.365	1		
0.369	-0.209	-0.060	1	
0.631	0.488	0.116	0.183	1

Q3

p-value=0.1473

0.0033				
0.0075	0.0290			
0.0227	0.0297	0.0079		
0.0401	0.0021	0.0101	0.0049	

Q4

Q5

1				
0.1996	1			
-0.0995	0.3679	1		
0.2996	-0.2991	-0.0998	1	
0.5988	0.4304	0.0521	0.1312	1

Q6

p-value=0.0526

	0.04032			
	0.00008	0.00184	0.01088	

1				
-0.017	1			
-0.316	0.117	1		
0.089	-0.558	-0.214	1	
0.439	0.126	-0.345	-0.136	1

1				
0.406	1			
0.130	0.517	1		
0.486	0.056	0.190	1	
0.727	0.631	0.368	0.443	1

1				
0.049	1			
-0.250	0.165	1		
0.154	-0.497	-0.203	1	
0.491	0.212	-0.253	-0.089	1

1				
0.347	1			
0.060	0.452	1		
0.435	-0.056	0.091	1	
0.693	0.569	0.265	0.341	1

1				
0.1996	1			
-0.0995	0.0827	1		
0.2996	-0.2991	-0.0998	1	
0.5988	0.3110	-0.1545	-0.0654	1

1				
0.1996	1			
-0.0995	0.5166	1		
0.2996	-0.2991	-0.0998	1	
0.5988	0.4633	0.1605	0.2680	1

1				
0.1996	1			
-0.0995	0.1432	1		
0.2996	-0.2991	-0.0998	1	
0.5988	0.3404	-0.1122	-0.0169	1

1				
0.1996	1			
-0.0995	0.4537	1		
0.2996	-0.2991	-0.0998	1	
0.5988	0.4492	0.1158	0.2181	1

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 FWER. Note also that the green cells of Q5 differ from the corresponding cells in Q2: even though the (green) angles distributions themselves remain unaffected by scenario restrictions, the ultimate correlation values of those cells ARE affected due to

$$R = BB^T \; .$$

NAbC Remains "Estimator Agnostic"

Another important and useful characteristic of NAbC, under both unrestricted and scenario-restricted cases, is that it remains "estimator agnostic," that is, valid for use with any reasonable estimator of 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. In Post 4 I address how NAbC's estimator-agnostic nature applies beyond Pearson's correlation, to any positive definite measure of dependence, thus adding one more flexible aspect of NAbC that further expands its already wide range of use and utility.

LNP: a Measure of Generalized Entropy

As I did in Post 2, it is worth taking an arguably minor digression here to examine further the meaning and implications of the cell-level (two-sided) p-values shown above in (4). The (two-sided) p-value provides what can be viewed as a competitor to distance metrics that has some advantages over traditional distance metrics, such as norms. Some commonly used norms in this setting for measuring correlation 'distances' are listed below in (8).

(8)
$$\|x\| = \left(\sum_{i=1}^{d} |x_i|^m \right)^{1/m}$$

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 (4) is the natural log of the product of the p-values, dubbed 'LNP' in (9) below:

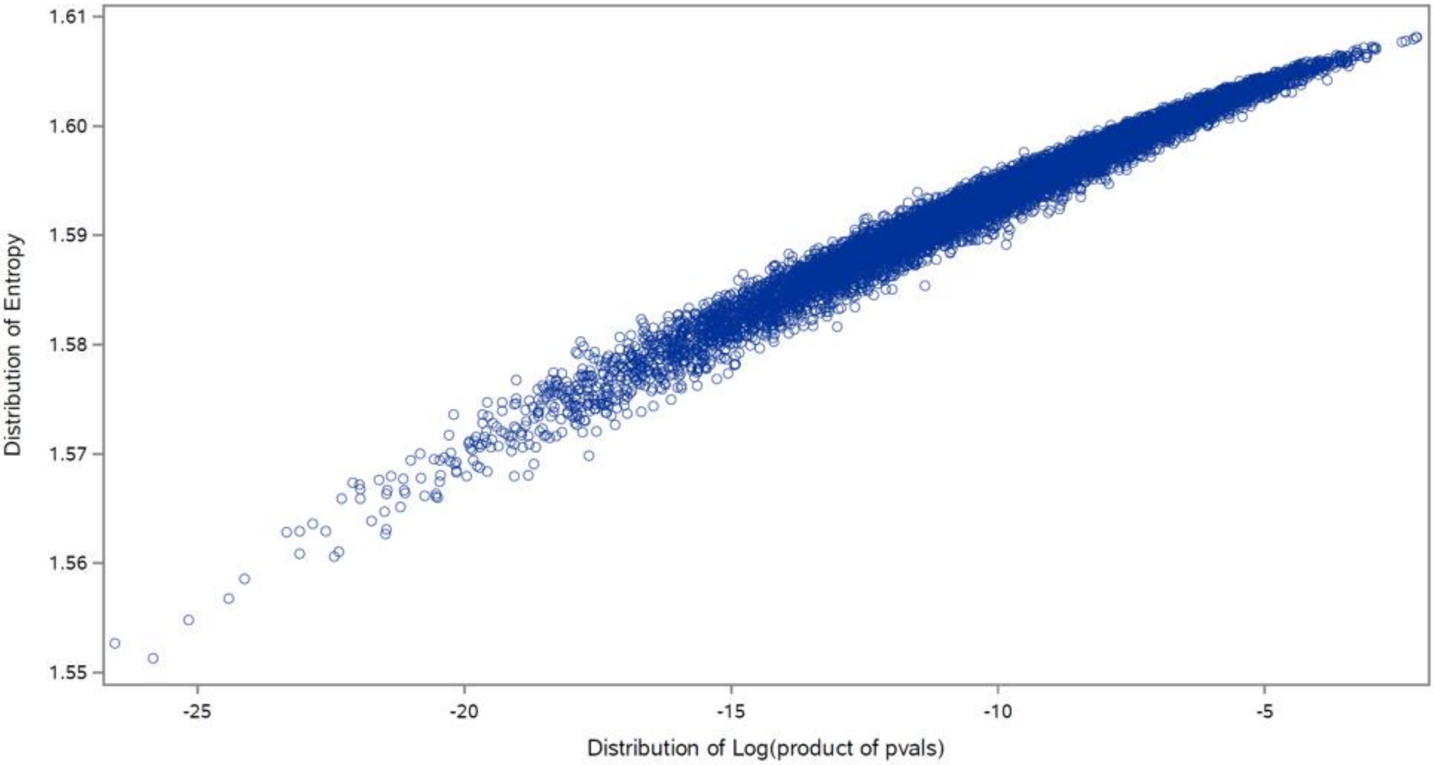
(9) $$\text{"LNP"} = \ln\left(\prod_{i=1}^{q} p\text{-}value_i \right) = \sum_{i=1}^{q} \ln\left[p\text{-}value_i \right] \text{ where } q = p\left(p-1\right)/2$$ and $p\text{-}value_i$ is 2-sided.

This was shown in Post 2, 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 (10) below:

$$\text{Entropy} = Ent\left(R/p\right) = -\sum_{j=1}^{p} \lambda_j \ln\left(\lambda_j\right)$$
(10)

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 (10), like NAbC, is valid for ANY positive definite measure of dependence, not just Pearson's, as will be discussed in Post 4). Graph 13 below 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 13: Identity Matrix Simulations -- LNP v Entropy



It is important to note, however, that entropy here is limited to being calculated relative to the case of independence, which for Pearson's corresponds only with the identity matrix.[19] In contrast, LNP can be calculated and retains its meaning in all cases, based on ANY values of Pearson's matrix, not just the identity matrix. Yet the correspondence of LNP to entropy under the specific case of the identity matrix 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 correlation 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

[19] Recall, 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.

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 measures such as various norms, has a solid and meaningful probabilistic foundation. Its calculation applies not only beyond the identify matrix for Pearson's, and the independence case generally, but also to ALL positive definite measures of dependence, regardless of their values, as discussed further in Post 4. 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 and 2 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 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 covered 4, 5, 6, and 7 above, with NAbC providing a fully analytic solution for the finite sample distribution of Pearson's under the Gaussian identity matrix. This Post 3 expanded NAbC's range of application to cover 1 and 3 as well, providing the solution for Pearson's under ALL matrix values and ALL real-world financial data conditions, using exactly the same angles-based framework. The only difference was the nonparametric rather than the analytic basis for defining the angles distributions, but all other components of the framework remain the same. Post 2 provided 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

while this Post 3 provides the same answers in an example, above, under much more general conditions. Post 4 will continue to expand NAbC's range of application to characteristic 2. above, providing the finite sample distribution for cases beyond Pearson's, including ALL positive definite measures of dependence.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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.

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

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.

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.

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

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

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

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

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

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.

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.

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