Big Data in Portfolio Allocation

By Irene Aldridge

Irene@AbleMarkets.com

AbleMarkets.com
BigDataFinance.org

This Draft: March 17, 2018

ABSTRACT

In the classic portfolio management theory, the weights of the optimized portfolios are directly proportional to the inverse of the asset correlation matrix. We show that, from the Big Data perspective, the inverse of the correlation matrix adds more value to optimal portfolio selection than the correlation matrix itself. We further show the empirical results of portfolio reallocation under different common portfolio composition scenarios, and outperform traditional portfolio allocation techniques out-of-sample, delivering nearly 400% improvement over the equally-weighted allocation over a 20-year investment period.

I. INTRODUCTION

According to DeMiguel, Garlappi and Uppal (2009, p. 1915), the idea of diversifying one's financial portfolio dates back at least to the fourth century AD, when Rabbi Issac bar Aha documented a rule for asset allocation in the Babylonian Talmud (Tractate Baba Mezi'a, folio 42a): "One should always divide his wealth into three parts: a third in land, a third in merchandise, and a third ready to hand."

Modern portfolio theory originated from Markowitz (1952, 1954) and these papers suggested not only diversifying assets and asset classes, but also finessing portfolio composition by taking into account mutual co-movement of returns. Investments, the theory goes, should be diversified so that if or when one investment head south, the others rise or at least they counterbalance the total value of the portfolio. Co-movement of returns is often proxied by correlation matrices. The optimal portfolio weights are computed to be directly proportional to the correlation matrix inverse.

When the number of positions is relatively small and stable, the classic Markowitz framework may work well. For larger portfolios, such as mutual funds and hedge funds with assets valued in the billions of USD, diversification suffers with unstable variance-covariance matrices, costly reallocation requirements and with some illiquid positions. Exchange-Traded Funds (ETFs) further complicate the situation by providing a low-cost universe of potentially redundant securities that did not exist during Markowitz era. The correlation matrices become very large. Big Data techniques become necessary to intelligently reduce the size of the correlation matrices, to select the key drivers in portfolios and to remove redundant securities. Doing so helps portfolio managers improve transaction costs, stability of portfolio weights and liquidity. With the advent of MiFID II and streamlined potentially flat transaction fees per financial instrument, the smaller universe of financial instruments traded may be particularly beneficial to institutional investors.

Another benefit of reducing portfolio selection is the shortened history required for a robust performance estimation. As illustrated by DeMiguel, Garlappi and Uppal (2009), increasing the number of instruments in the portfolio requires a significant increase in the length of historical data. Specifically, DeMiguel, Garlappi and Uppal (2009) find that a portfolio of 25 assets with monthly re-allocation requires a 250 year estimation window (across all of the positions) to reliably outperform the equally-weighted strategy. This is a difficult requirement to fulfil considering reliable daily records have been kept for less than 70 years. DeMiguel, Garlappi and Uppal (2009) also show out that the required estimating window scales linearly with the number of assets in the portfolio. Thus, the portfolio with 5 assets requires only 50 years of monthly data for reliable estimation.

Several techniques have been proposed over the years to mitigate the issues surrounding the Markowitz model. At the core of portfolio management is the question “Which instruments should be removed and which ones kept”? The decision is hardly trivial. Big data techniques do help to pinpoint the keeper`s in a reasonable time.

Traditional, not-big-data solutions to the problem of optimal portfolio allocation fall roughly into two categories: Bayesian and non-Bayesian. Bayesian approaches include statistical, diffuse-priors, “shrinkage” estimators and asset-pricing model priors. The Diffuse-priors approach was pioneered by Barry (1974) and Bawa, Brown and Klein (1979). The original shrinkage estimators date back to Jobson, Korkie and Ratti (1979), Jobson and Korkie (1980), Jorion (1985) and Jorion (1986). The original asset-pricing models for establishing a prior were discussed by Pastor (2000) and Pastor and Stambaugh (2000). More recently, Brandt, Goyal, Santa-Clara and Stroud (2005). They develop, for example, a simulation-based approach using recursion of approximations to the portfolio policy. Garlappi and Skoulakis (2008) simulate optimal portfolio choices using recursion of approximations to the portfolio value function.

Non-Bayesian non-big-data approaches to minimizing estimation errors are similarly numerous. Goldfarb and Iyengar (2003) and Garlappi, Uppal and Wang (2007) propose “robust” portfolio optimization to deal with estimation errors using uncertainty structures and confidence intervals, respectively. Kan and Zhou (2007) introduce a market portfolio into a classic tangent and riskless asset portfolio model as a way to mitigate estimation errors. MacKinlay and Pastor (2000) restrict the moments of returns by imposing factor dependencies. Best and Grauer (1992), Chan, Karceski and Lakonishok (1999), Ledoit and Wolf (2004a) and Ledoit and Wolf (2004b) propose methods for reducing the errors in the estimation of variance-covariance matrices. Frost and Savarino (1988), Chopra (1993) as well as Jagannathan and Ma (2003) introduce short-selling constraints.

A separate stream of literature considers different portfolio optimization frameworks that depend on the concurrent market regime (i.e bull versus bear market). For example, Ang and Bekaert (2002) use Markov Regime Switching Model to show that regime-switching strategies that rely on macro factors as “states” outperform static portfolio allocation strategies out-of-sample.

Optimization problems from other disciplines with similarities to portfolio management and optimal asset allocation have been successfully studied in great detail in the field of Big Data analytics. And Big Data has been making inroads in portfolio management. Partovi and Caputo (2004) were the first to apply PCA to the portfolio choice problem to decompose “principal portfolios” uncorrelated by construction. Meucci (2009) followed up on the idea with the creation of “maximum entropy portfolios”. Garlappi and Skoulakis (2008) apply Singular Value Decomposition (SVD) to solving several portfolio optimization problems in the context of the investor utility maximization. To do so, they deploy SVD to decompose state variables into fundamental “drivers” and “shocks”. The highest singular values or eigenvalues portray the drivers, while the lowest identify the shocks. Garlappi and Skoulakis (2008) apply the technique to solving the classic portfolio choice problem first proposed by Samuelson (1970) and extended by Hakansson (1971) and, later, Loistl (1976), Pulley (1981, 1983), Kroll, Levy and Markowitz (1984) and Markowitz (1991), among others. Allez and Bouchaud (2012) study eigenvalue evolution in covariance matrices and attempt to find a time-based pattern of covariance evolution. They find that the covariance eigenvalues evolve in time, as expected.

This paper is the first to study the Big Data properties of the inverse of the correlation matrix and shows that the inverse is much more informative than the correlation matrix itself, from the Big Data perspective. Subsequently, the paper proposes Big Data approaches to harness the correlation inverse and to deliver superior-out-of-sample returns.

The key advantages of the method we are proposing are:

1) Conceptual simplicity
2) Analytically-tractable performance improvements
3) Empirically-verified portfolio gains

II. BIG DATA OVERVIEW

Many Big Data techniques, such as spectral decomposition, first appeared in the 18th century when researchers grappled with solutions to differential equations in the context of wave mechanics and vibration physics. Fourier has furthered the field of eigenvalue applications extensively with partial differential equations and other work.

At the heart of many Big Data models is the idea that the properties every data set can be uniquely summarized by a set of values, called eigenvalues. An eigenvalue is a total amount of variance in the dataset explained by the common factor. The bigger the eigenvalue, the higher proportion of the data set dynamics that eigenvalue captures.

Eigenvalues are obtained via one of the techniques: Principal Component Analysis (PCA) or Singular Value Decomposition (SVD), discussed below. The eigenvalues and related eigenvectors describe and optimize the composition of the data set, perhaps best illustrated with an example of an image.

Consider the black-and-white image shown in Figure 1. It is a set of data points, “pixels” in computer lingo, whereby each data point describes the color of that point on a 0-255 scale, where 0 corresponds to pure black, 255 to pure white, and all other shades of gray lie in between. This particular image contains 454 rows and 366 columns.

Figure 1. Original sample image.

To perform spectral decomposition on the image, we utilize Singular Value Decomposition (SVD), a technique originally developed by Beltrami (1873). For detailed history of SVD, please see Stewart (1993). Principal component analysis (PCA) is a related technique that produces eigenvalues and eigenvectors identical to those produced by SVD, when PCA eigenvalues are normalized. Raw, non-normalized, PCA eigenvalues can be negative as well as positive and do not equal the singular values produced by SVD. For the purposes of the analysis presented here, we assume that all the eigenvalues are normalized, equal to singular values, and we'll use the terms singular values and eigenvalues interchangeably throughout this paper, as the results presented can be developed using SVD as well as PCA techniques.

In SVD, a matrix X is decomposed into three matrices: *U, S and V*

$$X = USV' \tag{1}$$

where
X is the original n x m matrix;
S is an m x m diagonal matrix of singular values or eigenvalues sorted from the highest to the lowest on the diagonal;
V' is a transpose of the m x m matrix of so-called singular vectors, sorted accordingly to the sorting of S; and

U is an n x n "user" matrix containing characteristics of rows vis-a-vis singular values.

SVD delivers singular values sorted from largest to the smallest. The plot of the singular values corresponding to the image in Figure 1 is shown in Figure 2. The plot of singular values is known as a "scree plot" since it resembles a real-life scree, a rocky mountain slope.

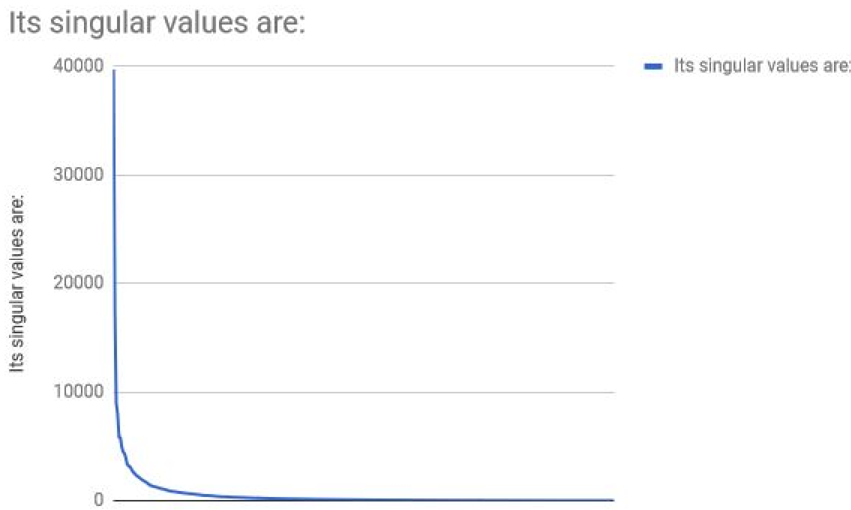

Figure 2. Scree plot corresponding to the image in Figure 1.

A Scree Plot is a simple line segment plot that shows: the fraction of total variance in the data as explained or represented by each singular value (eigenvalue). The singular values are ordered, are assigned a number label, by decreasing order of contribution to total variance.

To reduce the dimensionality of a data set, we select k singular values. If we were to use the most significant of the singular values, typically containing macro information common to the data set, we would select the first k values. However, in applications involving idiosyncratic data details. we may be interested in the last k values, for example, when we need to evaluate the noise in the system. A rule of thumb dictates to break the eigenvalues into sets before the "elbow" and after the elbow sets in the scree plot.

What is the perfect number of singular values to keep in the image of Figure 1? An experiment presented in Figures 3-9 shows the evolution of the data with varying number of eigenvalues included. The eigenvalues and the corresponding eigenvectors comprised of linear combinations of the original data create new "dimensions" of data. As the Figures show, as few as 10 eigenvalues allow a human eye to identify the content of the image, effectively reducing dimensionality of the image from 366 columns to 10.

To create the reduced data set, we restrict the number of columns in the S and V matrices to *k* by selecting *k* first elements. The resulting matrix $X_{reduced}$ has dimensions *n* rows and *k* columns, where

$$X_{reduced,\, n\, x\, k} = U_{n\, x\, k}\, S_{k\, x\, k}\, V^{T}_{k\, x\, k} \tag{2}$$

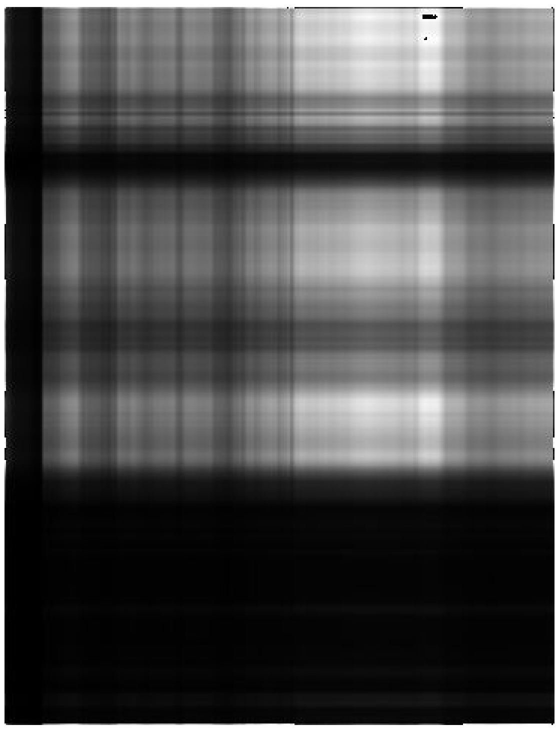

Figure 3. Reconstruction of the image of Figure 1 with just the first eigenvalue.

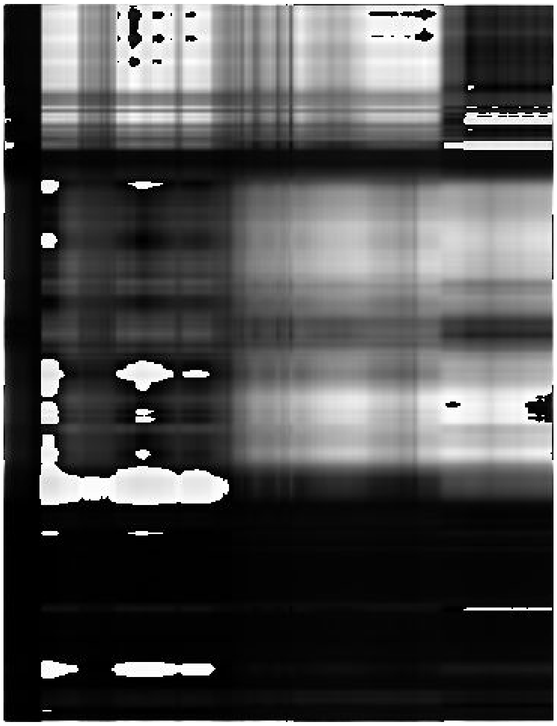

Figure 4. Reconstruction of the image in Figure 1 with the first two eigenvalues.

Figure 5. Reconstruction of the image of Figure 1 with the first 4 eigenvalues. The outlines of the figure are beginning to appear.

Figure 6. Reconstruction of the image in Figure 1 with the first 10 eigenvalues.

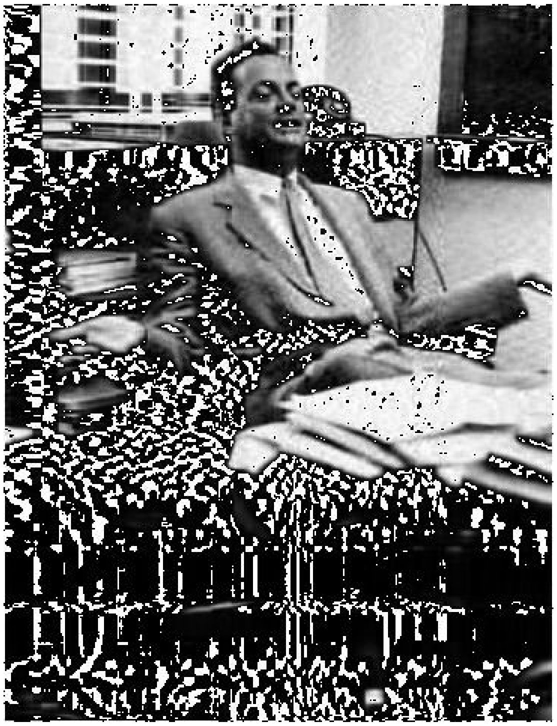

Figure 7. Reconstruction of the image in Figure 1 with the first 50 eigenvalues.

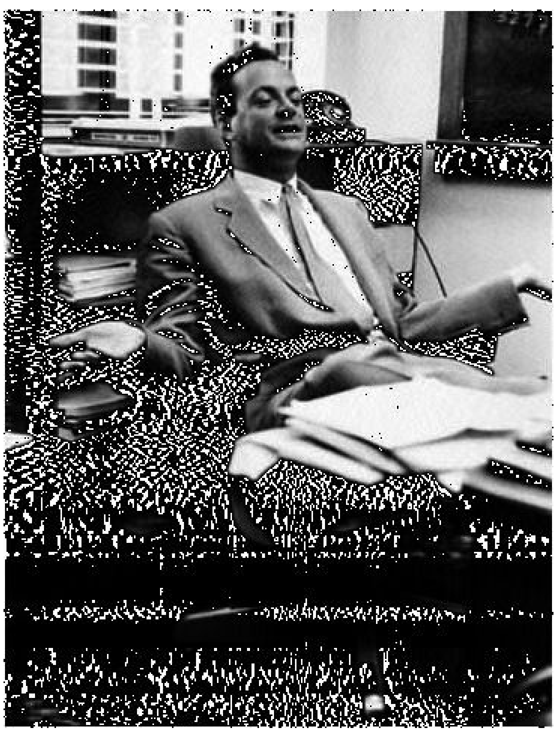

Figure 8. Reconstruction of the image in Figure 1 with the first 100 eigenvalues.

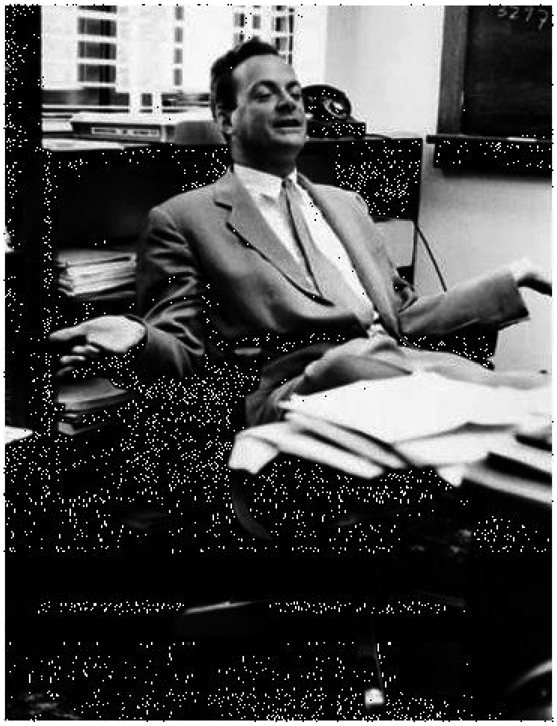

Figure 9. Reconstruction of the image in Figure 1 with the first 200 eigenvalues.

III. BIG DATA IN PORTFOLIO OPTIMIZATION: THEORY AND SIMULATION

3.1. Traditional portfolio overview

Markowitz-style portfolio optimization is often known as “mean-variance optimization” as it seeks to increase mean returns while simultaneously decrease variance in portfolios. Denoting the beginning prices of each asset *i* $X_i,\ i = 0, 1, ..., n$, we can express the investment portfolio as:

$$w_0X_0 + w_1X_1 + \ldots + w_nX_n \quad (3)$$

where $w_i,\ i = 0, 1, ..., n$ are portfolio weights: the proportion of the total portfolio wealth that is invested in the asset *i*. The sum of the weights of the portfolio assets is then equal to 1, and $w_0 + w_1 + \ldots + w_n = 1$. The asset with $i = 0$ is often assumed to be the prevailing risk-free rate, denoted r_0 .

Denoting risk aversion as γ, we now express the traditional mean-variance optimization as follows:

$$max_{w,w_0} (w_0 r_0 + w'\mu - \gamma w'\Sigma w) \qquad s.t.\ w_0 + w'1 = 1 \tag{4}$$

where Σ represents the variance-covariance matrix of the returns of the *n* assets under consideration.

Subtracting the risk-free rate, the maximization problem can be rewritten as follows:

$$max_{w,w_0} (w'(\mu - r_0 1) - \gamma w'\Sigma w) \qquad s.t.\ w_0 + w'1 = 1 \tag{5}$$

Equation (2) then leads to the following optimal solution:

$$w = \tfrac{1}{2\gamma}\Sigma^{-1}(\mu - r_0 1) \tag{6}$$

While the vector of returns is typically assumed to be the long-running average of returns on assets under consideration (see, for example, Jegadeesh and Titman, 1993), the covariance matrix presents several challenges to researchers and practitioners. Specifically, the covariance matrix can in turn be decomposed into variance and correlation matrices. While variances tend to be “sticky” and reasonably predictable by techniques like GARCH (Engle 1982, Bollerslev 1986, Andersen et al. 2006, Wong 2014), correlations of asset returns are notoriously volatile (see Davis and Mikosch 1998, Gourieroux 1997 and Cont 2001). It is the properties of correlation matrices that induce two key problems portfolio managers encounter when implementing mean-variance optimization:

1) Possibly extreme positions in selected assets (i.e. large proportion of the portfolio) resulting in liquidity constraints and violating the economic equilibrium of the portfolio allocation. To solve the issue, Black and Litterman (1993) and others propose a blended solution between economic equilibrium and mean-variance optimization.
2) Possibly extreme changes in portfolio weights from one investment period to the next, resulting in large transaction costs. Bertsimas and Lo (1998), Liu and Mei (2004), Muthuraman and Kumar (2006), Lynch and Tan (2008), DeMiguel and Nogales (2016), for example, propose penalizing the mean-variance optimization function with transaction costs as the remedy to the problem. However, such methods often tend to be opaque in practice.

Spectral decomposition of the inverse of the correlation matrix appears to be more promising and robust. The eigenvectors of an invertible matrix are also the eigenvectors of the matrix’ inverse. To show this, consider an invertible matrix *A*. Matrix *A* is invertible if and only if its determinant is not zero (Lipschutz 1991, p. 45), which in turn implies that matrix A columns are linearly independent, further implying that its eigenvalue λ is not zero. Suppose that matrix *A* has eigenvectors ***v***. By definition of eigenvectors, $A\boldsymbol{v} = \lambda \boldsymbol{v}$. Multiplying by A^{-1} from the left, we obtain:

$$\boldsymbol{v} = A^{-1}\lambda \boldsymbol{v} \tag{7}$$
$$A^{-1}\boldsymbol{v} = (1/\lambda)\boldsymbol{v} \tag{8}$$

Setting eigenvalues to zero prior to matrix inversion renders matrices singular, and, therefore, non-invertible. In other words, reducing the spectral dimensionality of the correlation matrices and subsequent inversion "blow up" the outcome. To overcome the issue, researchers often use "whitening" -- replacing set-to-zero eigenvalues with white noise N(0,1) to allow matrix invertibility. The process introduces noise into the system and may not be optimal.

Another solution is to exploit the fact that singular values of a matrix may be found and the dimensions reduced after the inversion with equal success and without sacrificing data precision.

$$(AB)^{-1} = B^{-1}A^{-1} \tag{9}$$

More generally,

$$\left(\prod_{k=0}^{N} A_k\right)^{-1} = \prod_{k=0}^{N} A_{N-k}^{-1} \tag{10}$$

So, for Singular Value Decomposition

$$A(p) = U(p)S(p)V'(p) \tag{11}$$

the inverse becomes

$$(A(p))^{-1} = (V')^{-1}S^{-1}U^{-1} \tag{12}$$

Singular Value Decomposition of the inverse of the correlation matrix is, therefore, much more precise as no data is lost due to the poorly-specified input to the inversion process that occurs with Whitening methodology.

Accordingly, the Singular Value Decomposition in the case of the matrix inversion can be performed as follows: the spectral decomposition can be performed after the matrix inversion without sacrificing results.

If SVD decomposes a correlation matrix C into $C = USV^T$, then the inverse of the matrix *C* can be written as $C^{-1} = \left(V^T\right)^{-1} S^{-1} U^{-1}$, where S^{-1} is the inverse of the diagonal matrix *S*:

$$S = \begin{matrix} {}_1 & 0 & 0 & \dots & 0 \\ 0 & {}_2 & 0 & \dots & 0 \\ & \dots & & & \\ 0 & 0 & \dots & {}_{n-1} & 0 \\ 0 & 0 & 0 & \dots & {}_n \end{matrix}$$

$$S^{-1} = \begin{matrix} 1/_n & 0 & 0 & \dots & 0 \\ 0 & 1/_{n-1} & 0 & \dots & 0 \\ & \dots & & & \\ 0 & 0 & 1/_2 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1/_1 \end{matrix}$$

Inverting the correlation matrix first, and then spectrally decomposing it to retrieve eigenvalues $\{\}$, therefore, allows the researchers to retain a lot more precision. Instead of replacing the irrelevant eigenvalues with noise to allow inversion, the proposed process is to replace the eigenvalues directly with 0 post-inversion.

3.2. Reducing the inverse of the correlation matrix

Which eigenvalues should you keep or discard? This, once again, is a non-trivial question. Spectral decomposition of the original, non-inverted, correlation matrix results in principal components or portfolios sorted according to their universality vis-a-vis all assets considered. Thus the largest component often represents the “global macro” portfolio factor driving most of the performance and typically reflecting the broad market movement. Several of the following eigenvalues deliver portfolios that induce synchronized fluctuations of groups of stocks; these can be factors driving industries, etc. The remaining small components are idiosyncratic in nature. Spectral decomposition of the inverted correlation matrix produces the eigenvalues sorted in the opposite order: from smallest to the largest.

Numerous big data techniques have been developed to help us understand the information content of the matrix under consideration, in our case, the inverse of the correlation matrix. Here, we develop and prove a conjecture that the top eigenvalue information content of the inverse of the correlation matrix always exceeds that of the correlation matrix itself. As a result, the big data analysis pertaining to the optimal portfolio allocation should be carried out on the correlation matrix inverse, not on the correlation matrix as it is done at present.

Correlation and covariance matrices contain real values, are symmetric with $a_{ij} = a_{ji}\ \forall i,j$, and are, therefore, called real symmetric, also known as Hermitian. A Hermitian $n \times n$ matrix has long been shown to have *n* real eigenvalues, all of which are non-negative and non-trivial (see, for example, Deift and Gioev, 2009). Under the SVD, these eigenvalues are

$$0 \le \lambda_n \le \ldots \le \lambda_1 \quad (16)$$

As shown by equation (12), the inverse of the correlation matrix has eigenvalues:

$$0 \le 1/\lambda_1 \le \ldots \le 1/\lambda_n \quad (17)$$

We are interested in the most descriptive eigenvalues and eigenvectors of the correlation matrix and its inverse. From the pure correlation matrix point of view, we should be retaining the largest eigenvalue in the correlation matrix, λ_1 . However, from the perspective of the inverse of the correlation matrix we may consider retaining the largest eigenvalue of the inverse, $1/\lambda_n$, instead..

A stream of literature following Borodin and Forrester (2003) develop a detailed Central Limit Theorem derivation for the largest and the smallest eigenvalues for Gaussian matrices. Tao and Vu (2012) and Wang (2012), among others, extend the results to covariance matrices, which are a special case of correlation matrices.

Specifically, Tao and Vu (2012) develop results for covariance matrices of any two random matrices *M* satisfying the following conditions: the matrix elements are i.i.d., normally distributed with mean 0 and variance 1, and have finite moments. Such conditions clearly satisfy the properties of normalized lognormal returns of financial instruments. The Tao and Vu (2012) results, therefore, apply to the financial returns covariances and correlation matrices as the special subset of the covariance structures. According to these results, the variances of the largest and the smallest eigenvalues are bounded by $n^{-4/3}$, where *n* is the rank of the matrix and the number of the eigenvalues. The Tao and Vu (2012) result, however, does not let us identify the bounds of the largest and the smallest eigenvalues in a sufficient enough detail: both the correlation matrix and its inverse have the same rank, making it impossible to compare the eigenvalues of the correlation matrix and those of the inverse.

Instead, we follow Wolkowicz and Styan (1980), who prove that for any matrix *A* with real eigenvalues $\lambda(A)$, the following result holds:

$$m - s(n-1)^{1/2} \le \lambda_{min} \le m - s/(n-1)^{1/2} \quad (18)$$

$$m + s/(n-1)^{1/2} \le \lambda_{max} \le m + s(n-1)^{1/2} \quad (19)$$

where $m = tr(A)/n$, $s^2 = tr(A^2)/n - m^2$

The trace of the correlation matrix, $tr(A)$, is defined as the sum of the diagonal elements and is always known to us in advance:

$tr(A) = n$ (20)

since all the diagonal elements of a correlation matrix are identities, 1.0. Thence, *m* in equations (18) and (19) becomes 1.

Similarly, the trace of A^2, $tr(A^2)$, is easy to find, since trace is also always equal to the sum of the eigenvalues of the matrix. For A^2,

$$tr(A^2) = \sum_{i=1}^{n} \lambda_i^{\,2} \tag{21}$$

Then, equations (18) and (19) can be rewritten as follows with $m = 1$, $s^2 = \sum_{i=1}^{n} \lambda_i^{\,2}/n - 1$:

$$1 - s(n-1)^{1/2} \leq \lambda_{min} \leq 1 - s/(n-1)^{1/2} \tag{22}$$

$$1 + s/(n-1)^{1/2} \leq \lambda_{max} \leq 1 + s(n-1)^{1/2} \tag{23}$$

Ultimately, we are interested in bounds of the smallest and the largest eigenvalues of the inverse of the correlation matrix, $\lambda'_{min} = 1/\lambda_{max}$ and $\lambda'_{max} = 1/\lambda_{max}$, respectively. Since, by equation (17), $0 \leq 1/\lambda_{max} \leq \ldots \leq 1/\lambda_{min}$, the so-called "bulk" of the eigenvalues will fall into place between the bounds on the edge values.

From equation (23), $1 + s/(n-1)^{1/2} \leq \lambda_{max} \leq 1 + s(n-1)^{1/2}$, we obtain:

$$\lambda'_{min} = 1/\lambda_{max} \geq 1/(1 + (\sum_{i=1}^{n} \lambda_i^{\,2}/n - 1)(n-1)^{1/2}) \tag{24}$$

$$\lambda'_{min} = 1/\lambda_{max} \leq 1/(1 + (\sum_{i=1}^{n} \lambda_i^{\,2}/n - 1)/(n-1)^{1/2}) = (n-1)^{1/2}/((n-1)^{1/2} + s) \tag{25}$$

Similarly, from equation (22),

$$\lambda'_{max} = 1/\lambda_{min} \leq 1/(1 - s(n-1)^{1/2}) \tag{26}$$

$$\lambda'_{max} = 1/\lambda_{min} \geq 1/(1 - s/(n-1)^{1/2}) = (n-1)^{1/2}/((n-1)^{1/2} - s) \tag{27}$$

The results of equations (24)-(27) hold for any matrix, real or complex, and the eigenvalues may be negative. Correlation matrices happen to be real and symmetric, known as Hermitian, and their eigenvalues are bounded by 0 from below. We would like to compare the range of eigenvalues for the correlation and its inverse. Hence, we are interested in establishing the lower bounds for the maximum eigenvalues, λ'_{max}, and the upper bounds for the minimum eigenvalues, λ'_{min}, for the correlation inverse, and comparing the two between correlation matrices and their inverses.

From (22)-(27), we obtain:

$$\lambda_{min} \leq 1 - s/(n-1)^{1/2} \tag{28}$$

$$\lambda'_{min} \leq (n-1)^{1/2}/((n-1)^{1/2} + s) \tag{29}$$

$$\lambda_{max} \geq 1 + s/(n-1)^{1/2} = (n-1)^{1/2}/((n-1)^{1/2} + s) \tag{30}$$

$$\lambda'_{max} \geq 1/(1 - s/(n-1)^{1/2}) = (n-1)^{1/2}/((n-1)^{1/2} - s) \tag{31}$$

From equations (28)-(31), we can conclude that while the lower bounds for the smallest eigenvalues of the correlation matrix and its inverse may be comparable, the upper bounds for the largest eigenvalues are clearly higher for the largest eigenvalue of the inverse.

$$\lambda'_{max} \geq \lambda_{max} \tag{32}$$

Furthermore, the largest eigenvalue of the correlation matrix is less or equal to 1, while the largest eigenvalue of the correlation inverse definitely exceeds 1.

THEOREM 1:

The largest eigenvalue of the inverse of the correlation matrix is always larger than the largest eigenvalue of the correlation matrix itself.

PROOF: by previous.

The obtained results are independent of the underlying distribution of returns. Indeed, the result accommodate Gaussian, leptokurtic and other distributions with equal effect, making the strategy robust to a variety of financial return models.

3.3. Simulation

To ascertain the validity of our conjecture, we perform 10,000 experiments of the following nature:

1. We create a random symmetric 100x100 matrix $\{A_{ij}\}$ simulating real-life correlation structure: all the values on the diagonal are set to 1.0, and all other values for $i \neq j$ range in the interval $[-1.0,\ 1.0]$, with entries $a_{ij} = a_{ji}\ \forall\ i, j$.
2. We compute and document the eigenvalues of the correlation matrix and its inverse.

As the results presented in Table 5 illustrate, the top eigenvalue of the inverse is considerably higher than the top eigenvalue of the correlation matrix itself.

TABLE 1. Summary statistics for eigenvalues of 10,000 simulated correlation matrices and their inverses.

	Top 1	Bottom 1	Top 1 Inverse	Bottom 1 Inverse
mean	11.60537822	0.05523877	**312.46715013**	0.08622607

stdev	0.30493585	0.04221966	**8,508.60273172**	0.00225207
skew	0.30964001	1.00654776	**50.50362873**	-0.14982966
kurt	0.17120805	0.88971901	**2,776.65596581**	0.02559518
max	12.83546300	0.25704200	500,000.000000	0.09435499
99%	12.39773146	0.18056210	1,545.25835255	0.09108325
95%	12.12671910	0.13857045	262.28504037	0.08983642
90%	12.00535900	0.11442080	121.92152061	0.08909438
75%	11.80207150	0.08023350	46.19737847	0.08777734
50%	11.59161850	0.04672950	21.39976314	0.08626923
25%	11.39246225	0.02164625	12.46362244	0.08473089
10%	11.22405270	0.00820200	8.73966974	0.08329613
5%	11.13134355	0.00381265	7.21654767	0.08246254
1%	10.97896710	0.00064715	5.53826083	0.08065992
min	10.59827400	0.00000200	3.89041480	0.07790915

III. OUT-OF-SAMPLE APPLICATIONS TO FINANCIAL DATA

We next test investment strategies on the historical financial data. The test utilizes daily closing price data for the S&P500 constituents for the 20-year period spanning 1998-2017 obtained from Yahoo!. We assume monthly portfolio reallocation and test the following strategies on the S&P 500 data: Equally-Weighted (EW), vanilla Mean-Variance Optimization (MVO), Principal Component Analysis (PCA) with the top eigenvalues retained, and Inverse Principal Component Analysis (Inverse_PCA) with the *bottom* eigenvalues of the inverse taken into the account, and the bottom eigenvalues discarded.

To compute strategy performance, we first determine the lognormal daily returns from the price data:

$$r_t = log(P_t) - log(P_{t-1}) \qquad (33)$$

We next compute monthly correlation matrices using the returns falling into each calendar month in the 1998-2017 span. Each correlation matrix then serves as an input to the strategy evaluation over the following month. For example, the correlation matrix computed on January 30, 1998, serves as the input for portfolio selection for February 1998.

Monthly performance of the strategies is next measured using the strategy weights computed on the last day of the previous month. The performance evaluation applies the weights to the returns observed on the last trading day of the following month vis-a-vis the price levels observed on the last day of the portfolio creation month. Thus, the performance of portfolios created on January 30, 1998, is tested by returns observed from the closing price on January 30, 1998, to the closing price observed on February 27, 1998.

Figures 10-13 document performance of the monthly reallocation of the strategies. As the Figures show, the PCA_Inverse strategy outperforms the other strategy when the number of selected eigenvalues is small, like the Top 1 eigenvalue selected in PCA_Inverse strategy shown in Figures 1 (with outliers) and 2 (outliers removed for clarity). As the number of retained eigenvalues increases, the PCA_Inverse strategy loses its power and eventually yields to the EW strategy.

Table 2 shows the Sharpe ratios from the obtained strategies. As Table 2 shows, the PCA_inverse strategy consistently outperforms other portfolio management strategies, particularly when the outliers, such as extreme one-time returns, are discarded from the data. Table 3 presents average monthly returns for each strategy computed over the 1998-2017 period. As shown in Table 3 once again, the PCA_inverse strategy delivers superior results when a concentrated number of eigenvalues is deployed to create an optimal portfolio allocation.

The results of the analysis so far show that just the top eigenvalue of the inverse of the correlation matrix contains enough portfolio information to outperform the other strategies. Just how many instruments does such a strategy contain? Figure 14 and Table 4 help answer this question. The number of positions with the absolute value greater or equal to 2% of the total portfolio value varied throughout the 20-year period, the number of stocks was significantly smaller than that of other strategies, pointing to a “smart diversification” portfolio selection of the PCA_Inverse strategy.

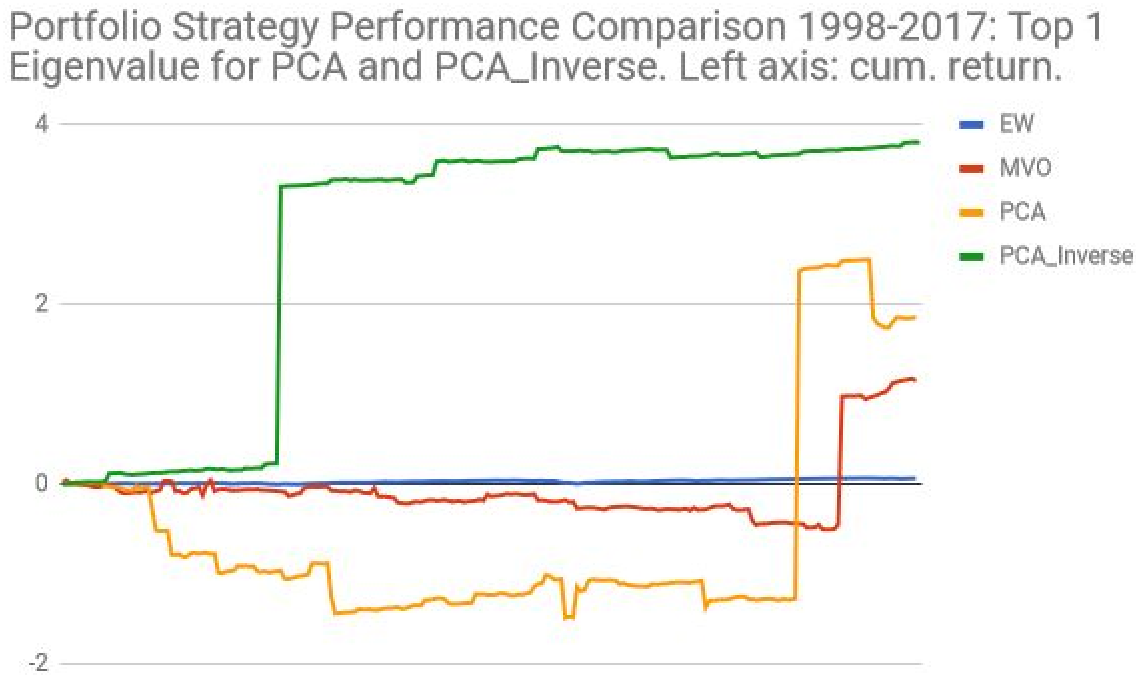

FIGURE 10. S&P 500 Strategy, monthly reallocations: keep the top 1 eigenvalue in the correlation matrix (bottom 1 eigenvalue in the correlation matrix inverse). Gross cumulative annualized returns of Equally-weighted (EW), standard mean-variance optimized (MVO), inverse correlation largest eigenvalue deciles (Inverse Largest), and inverse correlation smallest eigenvalue decile (Inverse Smallest) portfolios.

FIGURE 11. S&P 500 Strategy, monthly reallocations: keep the top 1 eigenvalue in the correlation matrix (bottom 1 eigenvalue in the correlation matrix inverse), outliers removed.

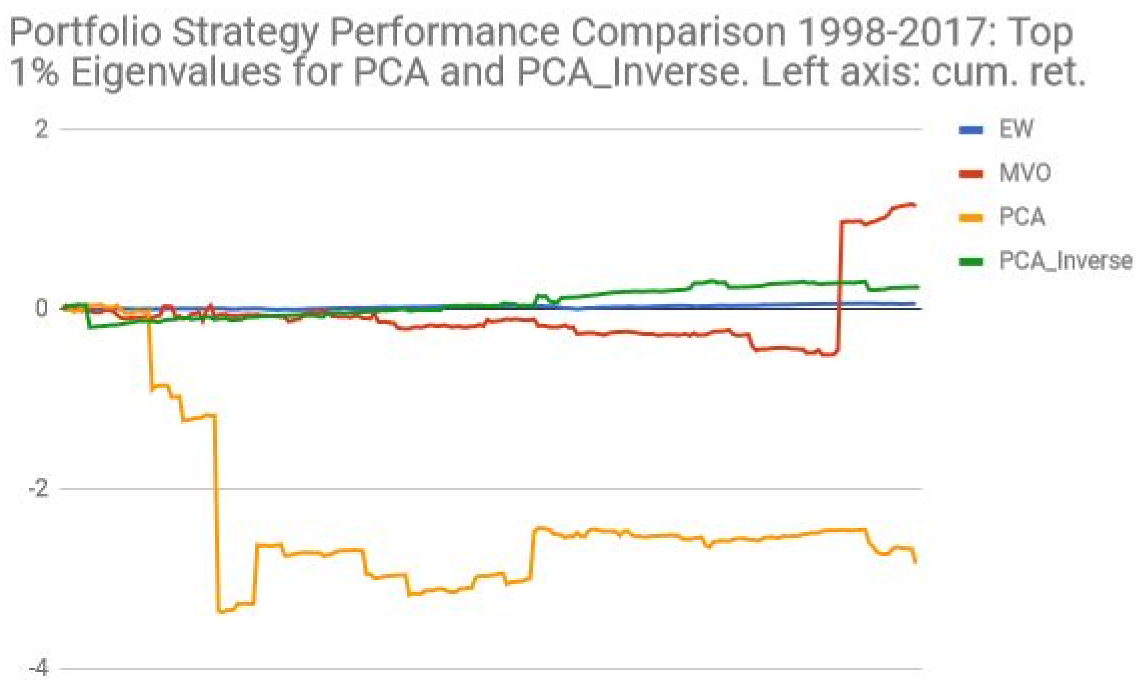

FIGURE 12. S&P 500 Strategy, monthly reallocations: keep the top 1% of eigenvalues in the correlation matrix (bottom 1% of eigenvalues in the correlation matrix inverse).

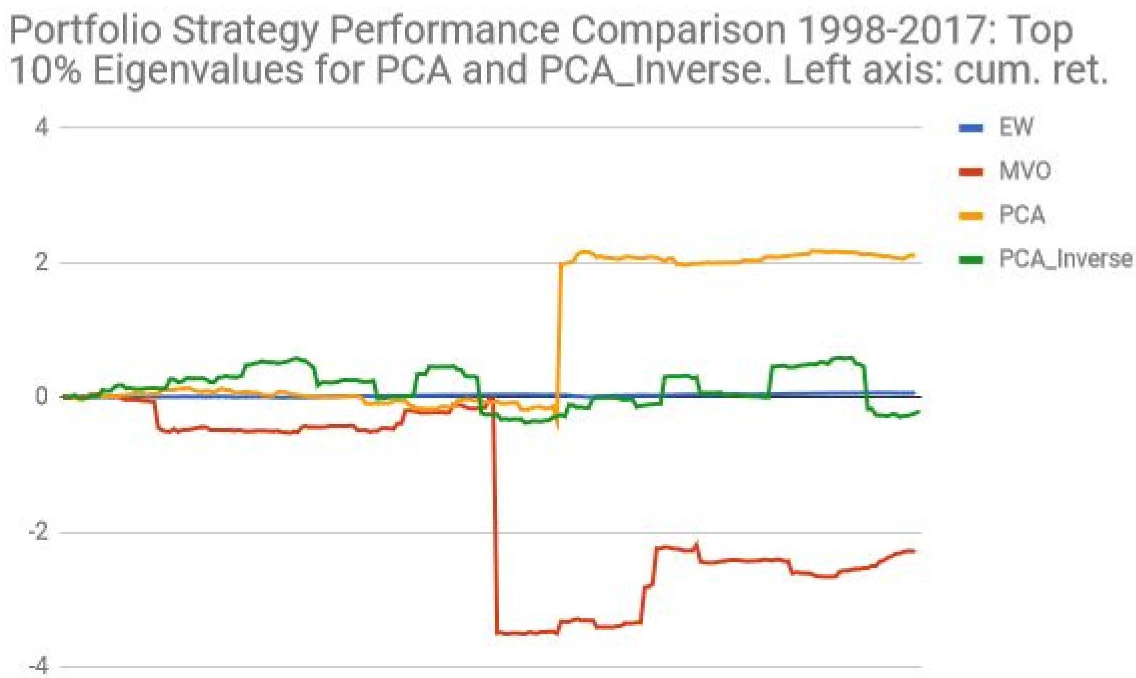

Figure 13. S&P 500 Gross cumulative annualized returns of Equally-weighted (EW), standard mean-variance optimized (MVO), inverse correlation largest eigenvalue deciles (Inverse Largest), and inverse correlation smallest eigenvalue decile (Inverse Smallest) portfolios.

TABLE 2. Sharpe ratios on strategy performance, S&P 500, 1998-2017, monthly reallocation

	EW	MVO	PCA	PCA_Inverse
10% Eigenvalues	0.4398529652	0.1660338977	0.2175831572	-0.05572290167
1% Eigenvalues	0.4398529652	0.1660338977	-0.2620669154	0.1854018356
1 Eigenvalue, with outliers	0.4398529652	0.1660338977	0.1121639833	0.2838331832
1 Eigenvalue, outliers removed	0.4398529652	-0.1695154284	-0.3799479568	0.6117174285

TABLE 3. Average monthly returns per strategy, S&P 500, 1998-2017, monthly reallocation

	EW	MVO	PCA	PCA_Inverse
10% Eigenvalues	0.0002994329004	0.004832313043	0.01287782073	-0.001001351801
1% Eigenvalues	0.0002994329004	0.004832313043	-0.01278558696	0.001204575758
1 Eigenvalue, with outliers	0.0002994329004	0.004832313043	0.007869265217	0.01649431169
1 Eigenvalue, outliers removed	0.0002994329004	-0.001353022026	-0.008363651982	0.003141938326

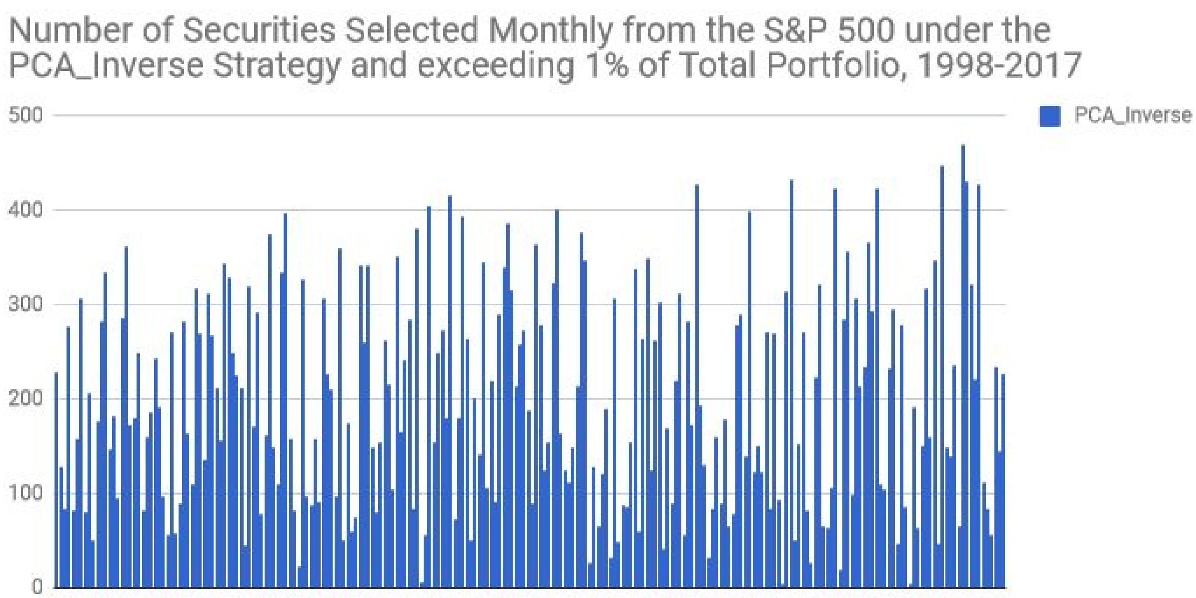

FIGURE 14: Number of securities selected each month from the S&P 500 by Inverse PCA method, 1998-2017, using only the top eigenvalue of the inverse. It is common for the algorithm to deliver single-digit number of names under this portfolio construction.

TABLE 4. Mean and Standard deviation (in parentheses) for the number of equities from the S&P 500 with absolute values of weights exceeding 1% or 2% of the entire portfolio selected monthly by vanilla Mean-Variance Optimization (MVO), Principal Component Analysis (PCA)

and PCA_Inverse method for different eigenvalue cutoffs. Data: 1998-2017, monthly portfolio rebalancing.

		MVO	PCA	PCA_Inverse
Top 1 Eigenvalue	|weight|>1%	375.49 (82.16)	375.15 (82.94)	192.44 (117.07)
	|weight|>2%	343.57 (88.33)	343.40 (89.77)	112.64 (116.94)
Top 1% of Eigenvalues	|weight|>1%	375.49 (82.16)	376.49 (84.30)	190.90 (115.95)
	|weight|>2%	343.57 (88.33)	346.27 (91.65)	109.98 (115.37)
Top 10% of Eigenvalues	|weight|>1%	375.69 (84.82)	377.83 (85.41)	372.13 (82.73)
	|weight|>2%	341.54 (93.03)	343.97 (93.17)	335.50 (89.24)

CONCLUSIONS

We show that the inverse of the correlation matrix carries inherently higher information content than the correlation matrix itself, affecting the Markowitz portfolio allocation strategies. To harness the power of Big Data analytics to capitalize on this information content, we propose a Big Data refinement to portfolio selection and test it on the S&P 500 index from 1998 through 2017. Our methodology appears to consistently outperform other common methods, such as equally-weighted portfolio allocation, plain mean-variance optimization and previously suggested big data portfolio optimization methodologies.

REFERENCES

Allez, Romain and Jean-Philippe Bouchaud, 2012. “Eigenvector dynamics: General theory and some applications,” *Physical Review E*, pp. 46-202.

Aldridge, Irene, 2009. “Systematic Funds Outperform Discretionary Funds.” Working paper, available at SSRN: https://ssrn.com/abstract=1516020

Aldridge, Irene and Steven Krawciw, 2017. Real-Time Risk: What Investors Should Know About Fintech, High-Frequency Trading and Flash Crashes. Wiley: Hoboken, NJ.

Ang, A., & Bekaert, G., 2002. “International Asset Allocation with Regime Shifts.” *Review of Financial Studies*, 15(4), pp. 1137-1187.

Avellaneda, Marco and Jeong-Hyun Lee, 2008. “Statistical Arbitrage in the U.S. Equities Market.” Working paper, Courant Institute, New York University. Available at: https://www.math.nyu.edu/faculty/avellane/AvellanedaLeeStatArb071108.pdf

Bauer, F. L. and C. T. Fike, 1960. “Norms and Exclusion Theorems.” *Numerische Matematik,* 2: 137-141.

Beltrami, E., 1873. “Sulle funzioni bilineari.” *Giornale di Matematiche ad Uso degli Studenti Delle Universita*, 11, pp. 98-106.

Bertsimas, D. and A. W. Lo, 1998. “Optimal Control of Execution Costs.” *Journal of Financial Markets,* 1 (1), 1-50.

Black, Litterman (1993) “Global Portfolio Optimization”, *Financial Analysts Journal*, 48, pp. 28-43.

Borodin, Alexei and Peter Forrester, 2003. “Increasing subsequences and the soft-to-hard edge transition in matrix ensembles,” Journal of Physics A 36, 2963-2981.

Brandt, M. W., A. Goyal, P. Santa-Clara, and J. R. Stroud, 2005, “A simulation approach to dynamic portfolio choice with an application to learning about return predictability,” *Review of Financial Studies*, 18, 831–873.

Cont, Rama, 2001. “Empirical properties of asset returns: stylized facts and statistical issues.” Quantitative Finance Volume 1 (2001) 223-236.

Dallaporta, S., 2012. “Eigenvalue Bounds for Wiegner and Covariance Random Matrices.” Available online at: https://arxiv.org/pdf/1203.1597.pdf

Davis R A and Mikosch T., 1998. “Limit theory for the sample ACF of stationary process with heavy tails with applications to ARCH.” *Ann. Statistics* 26, 2049–80.

Dieft, Percy and Dimitry Gioev, 2009. Random Matrix Theory: Invariant Ensembles and Universality. American Mathematical Society, Courant Institute of Mathematical Sciences: New York, NY.

DeMiguel, Victor, Lorenzo Garlappi, and Raman Uppal, 2009, “Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?”, *Review of Financial Studies*, 22, 1915-1953.

Feldmann, Michael, Dietrich Franken and Wolfgang Koch, 2011. “Tracking of Extended Objects and Group Targets Using Random Matrices”, *IEEE Transactions on Signal Processing*, 59, 1409-1420.

Garlappi, Lorenzo and Georgios Skoulakis, 2008. “A State Variable Decomposition Approach for Solving Portfolio Choice Problems.” *Review of Economic Studies,* 23, 3346–3400.

Gourieroux, Christian, 1997, ARCH Models and Financial Applications (Berlin: Springer).

Gupta, A. K. and D. K. Nagar, 1999. Matrix Variate Distributions. Chapman & Hall.

Hakansson, N. H., 1971, “Capital growth and the mean-variance approach to portfolio selection,” *Journal of Financial and Quantitative Analysis*, 6, 517–557.

Hilbert, D., 1904. “Grundzüge einer allgemeinen Theorie der linearen Integralgleichungen, Erste Mittelung”, *Nachr. Akad. Gott. Wiss. Math. Phys*., pp. 49-91

Jegadeesh, Narasimhan and Sheridan Titman, 1993. “Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.” *Journal of Finance,* Vol. 48, No. 1, pp. 65-91.

Kroll, Y., H. Levy, and H.M. Markowitz, 1984, “Mean-variance versus direct utility maximization,” *Journal of Finance*, 39, 47–75.

Liu, H., 2004. “Optimal Consumption and Investment with Transaction Costs and Multiple Risky Assets.” *The Journal of Finance 59*(1), 289-338.

Loistl, O., 1976, “The erroneous approximation of expected utility by means of a Taylor’s series expansion: analytic and computational results,” *The American Economic Review*, 66, 904–910.

Lynch, A.W. and S. Tan, 2008. “Multiple Risky Assets, Transaction Costs and Return Predictability: Allocation Rules and Implications for Us Investors.” *Journal of Financial and Quantitative Analysis* 45 (4), 1015.

Markowitz, H. M., 1991, “Foundations of portfolio theory,” *Journal of Finance*, 46, 469–477.

Mei, Xiaoling, Victor DeMiguel and Francisco Nogales, 2016. “Multiperiod portfolio optimization with multiple risky assets and general transaction costs”, *Journal of Banking and Finance 69*, pp. 108-120.

Meucci, Attilio, 2009. “Managing Diversification.” Risk, pp. 74-79.

Murthy, D.V. and R.T. Haftka, 1988. “Derivatives of eigenvalues and eigenvectors of a general complex matrix.” *International Journal For Numerical Methods in Engineering*, 26:293–311.

Muthuraman, K. and S. Kumar, 2006. “Multidimensional Portfolio Optimization with Proportional Transaction Costs.” *Mathematical Finance 16* (2), 301-335.
R.B. Nelson, 1976. “Simplified calculation of eigenvector derivatives.” *AIAA Journal*, 14(9):1201–
1205.

Partovi, M. H. and Caputo, M. (2004). “Principal portfolios: recasting the efficient frontier.” Economics Bulletin, 7(3):1–10.

Peche, Sandrine, 2006. “The largest eigenvalue of small rank perturbations of Hermitian random matrices,” *Probability Theory and Related Fields*, 134, pp. 127-134.

Pulley, L. B., 1981, “A general mean-variance approximation to expected utility for short holding periods,” *Journal of Financial and Quantitative Analysis*, 16, 361–373.

Pulley, L. B., 1983, “Mean-variance approximations to expected logarithmic utility,” *Operations Research*, 31, 685–696.

Samuelson, P. A., 1970. “The fundamental approximation theorem of portfolio analysis in terms of means, variances and moments.” *Review of Economic Studies,* 37, 537-542.

Skaf, Joelle and Stephen Boyd, 2009. “Multi-Period Portfolio Optimization with Constraints and Transaction Costs.”

Stewart, G.W., 1993. “On the Early History of the Singular Value Decomposition.” *SIAM Review*, Volume 35, Issue 4, 551-566.

Tao, Terence and Van Vu, 2012. “Random covariance matrices: universality of local statistics,” *The Annals of Probability,* Vol. 40, 1285–1315.

Van Der Aa, Nico, H.G. Ter Morsche and R.R.M. Matteij, 2007. “Computation of eigenvalue and eigenvector derivatives for a general complex-valued eigensystem.” *The electronic journal of linear algebra ELA*, September 2007.

Wang, Ke, 2011. “Random covariance matrices: universality of local statistics of eigenvalues up to the edge.”

H. Wolkowicz and G. P. H. Styan, 1980. “Bounds for eigenvalues using traces,” Linear Algebra Appl. 29, pp. 471-506.