On the use of Principal Components Analysis in Index Construction

Daniel Broby and William Smyth

Contact information of the corresponding author

{Correspondence: d.broby@ulster.ac.uk }

Abstract

This methodological paper presents a principal component analysis (PCA) derived way to construct equity indices. Although established in other disciplines, it represents a new application in finance. It is postulated that it can be useful to address entropy issues with non linear return time series that could potentially impact an index's ability to be a proxy for the market portfolio. In this approach, PCA is used to assign weights to individual equities, whilst the procedures to aggregate those equities are based on the PCA loadings. The method creates a factor model index (FMI) derived from PCA that delivers identifiable sub-sectors and weightings. The resultant portfolio recasts the efficient frontier and its weights can then be used to construct an index. The FMI can potentially be used for a number of asset sub-groupings. The approach can also be used to facilitate synthetic replication of risk factors.

Article

On the use of Principal Components Analysis in Index Construction

Abstract: This methodological paper presents a principal component analysis (PCA) derived way to construct equity indices. Although established in other disciplines, it represents a new application in finance. It is postulated that it can be useful to address entropy issues with non linear return time series that could potentially impact an index's ability to be a proxy for the market portfolio. In this approach, PCA is used to assign weights to individual equities, whilst the procedures to aggregate those equities are based on the PCA loadings. The method creates a factor model index (FMI) derived from PCA that delivers identifiable sub-sectors and weightings. The resultant portfolio recasts the efficient frontier and its weights can then be used to construct an index. The FMI can potentially be used for a number of asset sub-groupings. The approach can also be used to facilitate synthetic replication of risk factors.

Keywords: Principal Component Analysis; Index construction; Correlation Matrix.

1. Introduction

In this methodological paper, we present how to apply principal component analysis (PCA) to the construction of investment indices. The method creates a factor model index (FMI). Our paper is theoretical in nature and aims to provide a comprehensive overview of the PCA index approach.

Finance academics regularly use PCA as a reduction dimension technique, but it is not commonly used to construct indices. Other disciplines have, however, used them for this purpose. There have also been factor indices used in the finance industry derived from the Arbitrage Pricing Model [1]. That said, factor models derived from PCA represent a distinct method. Not only does PCA explain the diagonal terms of a covariance matrix and/or a correlation matrix, but it also explains the off-diagonal terms. The PCA method therefore incorporates both the characteristics and the assigned coefficients. These in turn can be used in to determine equity instrument weights and construct an index.

PCA has several useful mathematical properties for indices. The most important is that the index obtained from the first principal component explains the largest portion of variance of the individual equities in that index. This corresponds to the systemic risk factor of the capital asset pricing model. Malevergne et al (2007) [2] argue that this property makes its use consistent with the the self-consistency condition, namely that the market proxy is constructed of assets whose returns it is designed explain.

There have been precedents in the use of PCA methods to define index constituents by their common attributes. Daniel et al (1997) [3], for example, argue that such characteristics provide a better ex-ante forecast of the cross-sectional future returns. As such, they argue characteristic identification is a superior way of matching the likely realized returns of an asset class against a benchmark. Broby et al (2021) [4], meanwhile, derive a PCA factor model index (FMI) which proves more effective than existing commodity indicies, when used to construct an index for that asset class. The reader is referred to their paper for an empirical example of the method proposed.

We explain and expand on the PCA method used by Broby et al, (2021). It allows for the selection of sub-groupings and the grouping of asset proxies. The FMI method builds on the dimension reduction concept proposed by Pearson (1901) [5]. He documents

Received:
Accepted:
Published:

a multitude of uses of PCA. This papers presents its use purely from the perspective of index construction. In this respect, PCA is used in the determination of weights in indices through the eigenfactor of the first principal component in an asset sub-class.

The PCA method we illustrate adds to the number of different approaches to equity index construction method found in the literature. To be appropriate for its stated objective and performance attribution, a PCA index construction method should result in a combination of securities and/or asset classes that co-vary and can be grouped into appropriate sub-sectors. The various additional ways that this can be done is described in Meade and Salkin (1989) [6]. An example of one common method can be found in the MSCI Methodology Booklet (2018) [7].

2. Background

The use of PCA is well documented in disciplines other than finance. It is explained by Jolliffe (2002) [8] in his textbook on the method. As a statistical tool it is used in a number of fields where data is investigated in an exploratory manner. It is used in time series to seasonally adjust data, for example in the analysis of whether PCA can be extended for use in financial time series. PCA is a method to create uncorrelated indices. Specifically, it can be applied where each equity instrument is a linear weighted combination of the set of available instruments.

PCA is an established procedure in academic investigation but has only recently started to be used as a method in finance as a response to over-fitting in traditional multivariate regressions. In economics, it is used to show correlated response and predictor variables and is used in their statistical analysis. PCA has not been used previously to construct indices for equity assets. That said, it was used in a price context in commodities by Barlett (1948) [9]. He applied PCA to the time series of cotton over the period 1924 - 1938 in order to understand the nature of their returns.

In the literature there are several hierarchical models similar to PCA that are used to create optimal weights, as described by Polsen and Tew (2000) [10]. They show how they can be used to construct portfolios that can in turn be used as benchmark indices. They detail how Bayesian methods can be incorporated to treat parameter uncertainty, such as missing return data. This approach is useful for indices focused on in infrequently priced asset classes, such as real estate. That said, most current methods, as explained, rely on representation rather than replication. Amenc et al (2012) [11] explain that in the index replication stage that one should have two steps in the construction process, these being constituent and weighting scheme selection.

The advantage of PCA usage in an equity universe application is that clusters are easily identified. It overcomes the problems with peer indices identified by Bailey (1992)[12] and those constructed without attention to correlation, co-variance skew and kurtosis. It also explains variability, and when adapted using factor analysis, correlation. The theorems behind PCA, matrix algebra and multivariate analysis are explained well by Rao (1979) [13], amongst others. It can be used on investment proxies, thereby filling an identified need in the literature as relates to equity assets.

We suggest that PCA can help define an appropriate index as a result of its robustness. A good explanation of how it can be applied is given by Jolliffe (2002) [8]. He points out it is a particularly useful method if there is a large amount of data and one wants to view the various sub-groups visually in two dimensional space. As there are a number of sub-groups in equity asset classes that are very different, this is deemed appropriate. For example, gold mining stocks are very different from coal mining stocks.

The PCA has similarities to a regression model. In this respect, it creates an orthogonal transformation of the individual instruments, thereby better explaining the way they group together. In technical terms it results in a linear transformation of the data at the same time as preserving the statistical symmetry. It does this by taking data representing the first principal component and data representing the second greatest variance from that,

and then regroups them. As a result, it can be used on the time series of equity assets to re-evaluate the variances, co-variances and correlations.

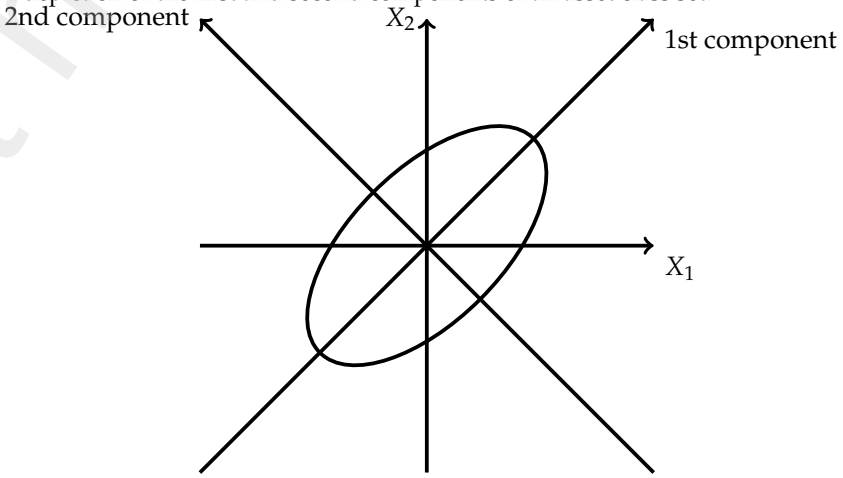
PCA is also used in a portfolio context. Partovi et al (2004) [14] demonstrate how it is possible to recast the efficient frontier using PCA from a set of uncorrelated assets. It should be noted, however, that most assets have some degree of correlation. That said, they illustrate how PCA simplifies portfolio structure and delivers a conceptually more transparent solution. Meanwhile, PCA is also applied to the analysis of equity portfolios by Pasini (2017) [15]. He proposes that the method can be used to determine an index to test how much a time series departs from being a sequence of independent and identically distributed random observations with finite mean and variance. This is a different application from the one proposed in this paper, but delivers an interesting insight. He identifies that the first principal component is typically equivalent to the market factor, and that the second principal component is typically sufficient to represent the remaining risk.

In summary, the PCA approach is a method of constructing indices that involves identifying common components and assigning weights based on the optimization of eigenvalue results. This is different from traditional index instrument sampling, which is typically based on an optimization of market capitalization. PCA can be used as a classification method by constructing an index based on the weights of identified factors, which can then be tested to determine if the weights are optimal. The use of PCA in time series of returns helps to create robust indices, as it is a data reduction technique that allows the index constructor to create interpretable factors and assign weights to them. The orthogonal transformation is the process by which this occurs.

3. PCA explained

It is easier to understand PCA visually. Figure 1 presents a geometric representation based on two variables, X_1 and X_2. These are centered on their respective means. The ellipse illustrates the scatter of sample points. The line that transects the first principal component is derived from the widest point. The second component is the line which is at right angles to this first principal component. The initial reference point is used and a rigid transformation is applied around the origin. This results in a new set of axes. The origin is given by the sample mean average on the two X_1 and X_2 variables.

Figure 1. A depiction of the first and second components of an asset class set.



A geometric representation based on two asset variables, X_1 and X_2, showing the first component and second component rotations. In the case of equity assets these could be the first component in the direction along which the asset instruments have the largest variance. The second principal component is the direction which maximizes variance in those instruments from all directions orthogonal to the first component.

Meanwhile, figure 2 shows the transformed axis. The components in it can be explained algebraically based on the two variables, X_1 and X_2, with the following variance-covariance matrix

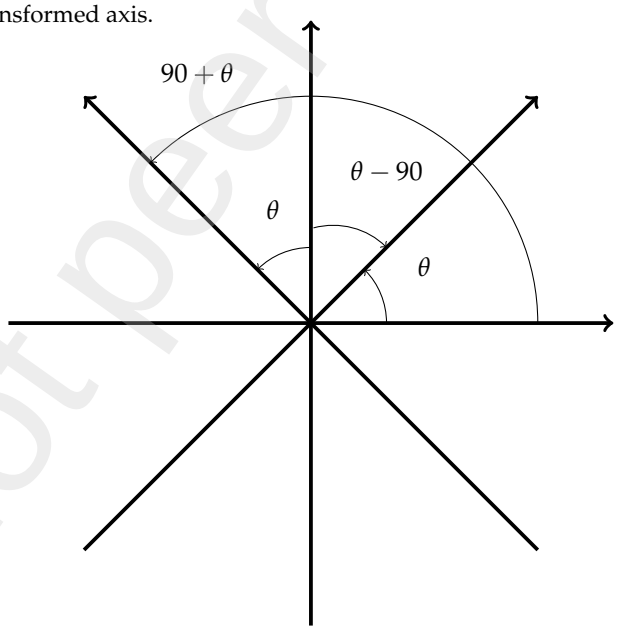
$$\mathbf{\Sigma} \boldsymbol{X_1, X_2} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

a_{11} and a_{21} denote the weights from the first eigenvector of $\mathbf{\Sigma}$; a_{12} and a_{22} are the weights from the second eigenvector. If can be represented by a 2×2 orthogonal (or rotation) matrix $\mathbf{T}$, with the first column containing the first eigenvector weights and the second column the second eigenvector weights. This then allows the calculation of the direction cosines of the new axes based on the following:

$$\mathbf{T} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \cos(90+\theta) \\ \cos(\theta - 90) & \cos(\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.$$

The cosines of the angles are based on the positive (horizontal and vertical) axes. The orientation of the transformed axis can therefore be found by multiplication of the relevant eigenvector values by -1, as depicted in figure 4.1.

Figure 2. A PCA transformed axis.



A transformed axis showing the cosines of the angles on the horizontal and vertical axes. The case of two-dimensional rotations can be extended to three or more dimensions by using the appropriate matrix of the direction cosines. In this way, one can build multi-factor models from which to build indices. The axis shows the direction of maximum spread. This is the principal axis. With this it is possible to subtract the variance to obtain the remaining variance. The same procedure is applied to find the next principal axis from the residual variance. The principal axis must be orthogonal to any other principal axes. The transformed data become the principal components.

4. Orthogonal transformation

In order to understand how PCA can be used as a sampling method to construct an index it is necessary to specify the process. The technique is primarily a data analytic technique, so its use in indices is not widely appreciated. A tutorial is given by Shlens (2005) [16]. It uses linear algebra to obtain transformations of the data. These are orthogonal in nature and help with identifying how the data is grouped. The non-orthogonal vectors are depicted in figures 3 and 4. In index construction, this results in a linear transformation that preserves the integrity of the relationships between the various asset instruments. This

allows for weights to be assigned. This traditionally is done in index construction through sampling rather than statistical technique.

Figure 3. Non-orthogonal 3D coordinate systems.

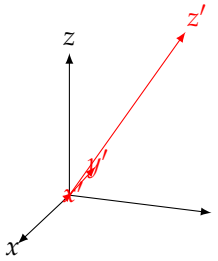


Figure 4. Vectors in 3d and non–orthogonal basis vectors.

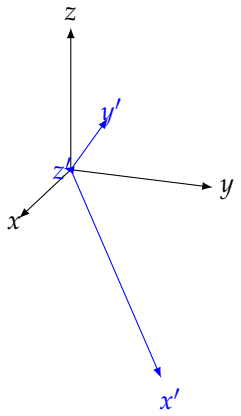


Figure 5. Vectors in 3d and non–orthogonal basis vectors.

Non–orthogonal 3D coordinate systems. Orthogonal is a term used to mean normal . In Euclidean space, two vectors are orthogonal if they make an angle of 90 degrees, or one of the vectors is zero. This figure represents the transformation that a set of asset instruments would go through when PCA is applied.

As a result of transforming the first loading vector in the way depicted in the diagram, the variance of the individual asset instruments is maximized. The total variance remains the same. It results in a redistribution of the new equity asset instruments on a different dimension. The outcome is determined by the most "unequal" result. In this way, the first equity asset not only explains the most variance among the new assets, but the largest variance of any single instrument. This is illustrated mathematically as where w equates to (see [17] for formulas and proof):

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i (t_1)^2_{(i)} \right\} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \sum_i \left(\mathbf{x}_{(i)} \cdot \mathbf{w} \right)^2 \right\} \tag{1}$$

Where:

- $w_{(1)}$ = Weighting load factor one.

This is represented in matrix form as:

$$\mathbf{w}_{(1)} = \arg\max_{\|\mathbf{w}\|=1} \{ \|\mathbf{X}\mathbf{w}\|^2 \} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right\} \tag{2}$$

Where:

- $w_{(1)}$ = Weighting load factor one.

When the transformation has been made, the next step is to extend the statistical input by the calculation of an additional factor component. This kth component is found by subtracting the result from the first component. In effect, another rotation is made. This has

the effect of splitting out different types of asset groupings (similar to equity sub-sectors). Think of it as potentially isolating different investment characteristics. The equation below shows how this is presented algebraically, highlighting the weighting of the respective identified factor.

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^{\mathrm{T}} \tag{3}$$

Where:

- $w_{(k)}$ = Weighting load of the Kth factor.

Once the weighting has been identified, the loading factor vector should then be calculated. This is the point of the maximum variance from the new data matrix. It is shown algebraically thus:

$$\mathbf{w}_{(k)} = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\hat{\mathbf{X}}_k\mathbf{w}\|^2 \right\} = \arg\ \max \left\{ \frac{\mathbf{w}^T\hat{\mathbf{X}}_k^T\hat{\mathbf{X}}_k\mathbf{w}}{\mathbf{w}^T\mathbf{w}} \right\} \tag{4}$$

The results can then be presented as a set of weights which can be used in an index. These are mathematically expressed as P dimensions. In the construction of an index using this method, the random vector of returns is found from the universe of the relevant equity assets. This is done with a mean vector where the vector is the *common asset factors* and the matrix of factor loadings are the *specific factors*. Note that this is similar to the output to the market model, which has a common market factor and various stock specific factors. It is not necessarily necessary to re-estimate the model every day when using PCA in index construction. However, the model may need to be updated quarterly or annually to account for changes in the data or to incorporate new information.

The creation of a common asset factor mean that the PCA approach has a theoretical link to the market proxy, as found in finance theory. That proxy is derived from the market model and mean variance portfolio theory. It can be used to justify broad market indices. The output shows that the variance for the asset equals the sum of the squared outputs for that equity asset.

Using this approach, the structure of equity assets generates an estimate of the relevant factors from their eigenvectors. That is, it identifies those factors associated with the largest eigenvalues of the matrix output. It is these that form the basis of the weight of the contender equity asset index, as shall be further explained.

5. Deriving factors from principal components

The properties of the PCA output means that it is possible to further derive investment factors. This is done in the same way that one can deduce that systemic risk is an important factor through principal components. One does this by starting with a matrix of the equity asset class opportunity set. This can be expressed mathematically, for example on a stock matrix with five factor loadings. This as illustrated in the equation below:

$$X = \mu + LF + \epsilon \tag{5}$$

Where:

- X: vector of the equity asset class returns. $X = \begin{pmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \dots \\ \mathbf{X_n} \end{pmatrix}$
- μ: X is drawn from a univserse of stocks with mean vector $\mu = \begin{pmatrix} \bar{}_{\mathbf{1}} \\ \bar{}_{\mathbf{2}} \\ \dots \\ \bar{}_{\mathbf{n}} \end{pmatrix}$

- L: $k \times n$ matrix of factor loadings. $F = \begin{pmatrix} l_{1,1} & l_{1,2} & ... & l_{1,5} \\ l_{2,1} & l_{2,2} & ... & l_{2,5} \\ ... & ... & ... & ... \\ l_{n,1} & l_{n,2} & ... & l_{n,5} \end{pmatrix}$
- F: vector of common factors. $F = \begin{pmatrix} \mathbf{f_1} & \mathbf{f_2} & ... & \mathbf{f_5} \end{pmatrix}$
- ϵ: vector of errors (specific factors). $\epsilon = \begin{pmatrix} \mathbf{''_1} \\ \mathbf{''_2} \\ ... \\ \mathbf{''_n} \end{pmatrix}$

Using the PCA approach, the variance for the equity asset class i^{th} is going to be equal to the total of the squared loadings and the variances of the instruments:

$$var(X_i) = \sum_{j=1}^{n} l_{i,j}^2 + \psi_i \tag{6}$$

Where:

- $\sum_{j=1}^{n} l_{i,j}^2$: communality of the equity asset class i
- ψ_i: specific factor for the equity asset class i.

The PCA approach is effectively an estimation using the the maximum likelihood statistical method. Although the original loadings are ambiguous, and therefore the instrument factors are not easy to determine, the rotations give a better understanding of how the asset class is broken down. The model that shows this is as follows:

$$X = \mu + LTT'F + \epsilon = \mu + L^*F^* + \epsilon \tag{7}$$

for the orthogonal matrix $\mathbf{T}$.

This approach was first presented in a generalised way by Hendrikson (1964) [18]. He argues the method should be used in preference to the varimax orthogonal method, suggested by Kaiser (1958) [19]. In this way, the oblique solution is effectively obtained by trial and error, increasing the larger loads and reducing the smaller ones. As suggested earlier, a good knowledge of the time series of the asset in question is helpful to an index constructor.

The link with factors and eignevalues, which determines how they can be used in index weights, was explained by Ronacalli (2007) [20]. The method is applied to the identified index universe and results in a risk factor being generated in the co-variance matrix in the first instance. This is the same as the market risk factor. The eigenvectors that follow are, in this universe of assets, common risk factors.

5.1. Interpreting the risk factors

The most important principal component is a proxy for systemic market risk, and the subsequent components contain useful information about financial time series (specific risk factors). Yand et al (2015) [21] used PCA to interpret a co-variance matrix of asset returns. Interestingly, the last few common components are found to be meaningful, as they identify instruments with non linear correlations. This is relevant because there is an on-going debate in finance about the number of factors that explain return. It implies that with equity assets one does not have to resort to Capital Asset Pricing Models to identify common factors.

In linear algebra, an eigenfactor is a scalar value that, when multiplied by a given matrix, produces a new matrix that is a scalar multiple of the original matrix. The scalar value is known as the eigenvalue of the matrix, and the process of finding it is known as finding the eigenvalues of the matrix. As far as index construction goes, the second eigenvector is a combination of asset weights orthogonal to the first eigenvector and so on.

In this way, the factors identify the variance not explained by the first eigenvector. This can be critiqued as difficult to use to identify a specific asset class group, as it means there is no real way of determining the number of eigenvectors without knowing the original number of sub-groupings that the equity asset class exhibits. Financial industry experience, however, can be used to manually identify these but for the purpose of index creation the first eigenvector is sufficient.

It is possible to use PCA to create a mean variance optimal index.[**?**] This is portfolio of assets representing an index that has been optimized to maximize expected returns while minimizing risk. It is calculated using the mean and variance of the returns of the assets in the portfolio, with the goal of finding the optimal balance between risk and return. This approach to portfolio construction is based on the idea that investors are risk-averse and willing to trade off higher expected returns for lower levels of risk.

To do this it is necessary to interpret the co-variance matrix (or historical co-variance matrix) in the context of the factor risk ([22]. Alternatively, using those outputs generated using shrinkage of the sample covariance matrix; or with the two common co-movement measures, the Gerber statistic [23] and the modified Gerber statistic [24]. Such an approach allows for indices to be constructed without knowing anything other than the return time series.

6. Re-grouping PCA results into index components

To create an index, the underlying instruments must be regrouped from their raw form. This is done through a variance reduction method. The first step is the application of PCA. The next step is to analyze the data using the variance-covariance matrix for equity asset returns. The aforementioned literature suggest the results, if applied to equity asset classes, should have a number of factors that determine the variation of the data. The matrix has been created, it is then necessary to use the output for subsequent index construction by using factor analysis. In this way, it is possible to create an index. A yearly or monthly ranking can be calculated and a re-balancing holding period can be applied.

Once the common components have been established, it is possible to determine the factors present using the associated dimension reduction technique. This is a method for modeling observed variables and their co-variance structure for a small number of underlying un-observable latent factors. It can be considered as an inversion of the PCA. The next step is to create linear combinations of the observed variables. To do this the FMI weights are derived from a factor analysis implemented through a variance-covariance matrix of the returns of equity asset instrument sets. This is repeated on each date of the new reconciliation.

The results deliver a variance fraction for each of the identified factors. With these results, for each identified factor, the formation of a sub-portfolio is possible. This is based on only instruments with a significant loading to the identified factor. A loading factor has to be determined. This is a statistical measure that represents the strength of the relationship between a particular observed variable and an underlying latent factor. In factor analysis, the observed variables are believed to be influenced by a smaller number of unobserved, underlying factors. The loading factors are used to quantify the extent to which each observed variable is related to each latent factor. The loading factor for a particular observed variable and latent factor is calculated as the correlation between the observed variable and the latent factor. A factor of greater than 0.3 was recommended by Chao and Wu (2017).[25]

It is suggested that the FMI weights be derived from data observed over annual observation periods. This is for ease of computation. That said, the method can be used to construct equity asset portfolios held over a shorter re-balancing period. Each portfolio that is created in this way is essentially the index at this time. At the end of this period, any FMI weights are updated and the portfolio re-balanced using the same procedure.

In the next stage of the PCA approach, each equity investment vehicle receives a weighting equal to the *nth* ratio of its load relative to the sum of the loads of the commodities

Table 1. Illustrative factor output

This table illustrates how the means and standard deviations for individual equities and sector loadings are presented based on the first five factors corresponding to the rolling factor analysis used for FMI construction.

	Factor 1		**Factor 2**		**Factor 3**		**Factor 4**		**Factor 5**	
	Mean	σ	**Mean**	σ	**Mean**	σ	**Mean**	σ	**Mean**	σ
Security	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%
Security	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	00.00	0.00%
Security	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%
Sub-sector grouping	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%	0.00	0.00%

contained in the sub-portfolio. The resulting group of sub-portfolios is then aggregated into an overall portfolio in which each sub-portfolio receives a weight equal to the ratio of the variance component. This is explained by the factor resulting from the total variance explained by the factors determined.

The experimental factor results can then be put into an oblique rotation. This allows for some correlation between the underlying factors and provides a clearer picture of the variance decomposition. It should show groups of equity asset instruments as single factors. It stresses the underlying interactions between the factors and should be thought of as similar to creating sectors in traditional index construction.

6.1. Deriving index weights

To recap, the FMI is derived through the constituent weights for each period using factor analysis implemented using oblique rotations. This is a transformation of coordinate axes in which the new axes are not perpendicular to one another, thereby producing seperate factor outputs. The output would appear as in Table 1. The PCA index return is therefore a weighted average of the returns of the derived equity asset portfolio constituents. The resultant factor model can be described as such:

$$PCA_r i, t = \beta 1 Fv_t + \beta 2 Fv_t + ...\beta_n Fv_t \tag{8}$$

Where:

- $PCA_r i, t$ = FMI, the excess return of portfolio i in month t,
- Fv = Factor identified by eigenvalues

This model can be used to construct an index in a stepwise fashion. The goal is to find the combination of variables that results in the best model, in terms of some measure of model fit. We recommend a multi-year observation period along with a three or six-month rebalancing frequency conducted at each rebalancing date. This is to smooth out the effect of rebalancing and to minimize the impact of rebalancing costs.

7. Discussion

A key contribution of the PCA approach is that it explicitly addresses the time series of equity assets as being non-normal in their distribution. This fills a gap in the literature as this was not previously addressed by scholars. Hubert et al (2009) [26] demonstrated how robust PCA can be in the face of skewed data. The method is justified in this context because, in order to represent the data in lower dimensions, PCA assumes the joint distribution of data follows a non multivariate normal distribution. Indeed, the only distribution that can represent the time series in a compact form is Gaussian distribution. As a result, the PCA makes an implicit assumption that data should follows Gaussian distribution.

The other contribution made by the application of the PCA method to index construction is its ability to allow replication without necessarily requiring direct investment in the underlying instruments. Because PCA reduces the dimensionality of a financial time series by projecting the data onto a lower-dimensional space while preserving as much of the

original information as possible, it can be useful in index construction because it can help to eliminate redundancy in the data and make it easier to analyze and interpret.

One advantage of using PCA in index construction is that it can be more accurate in tracing the performance of the component stocks. By reducing the dimensionality of the data, PCA can help to eliminate noise and capture the most important patterns and trends in the data. This can be especially useful when working with large and complex datasets, as it can help to simplify the analysis and make it more interpretable.

Another advantage of using PCA in index construction is that it can be less computationally intensive. Because PCA reduces the dimensionality of the data, it requires fewer calculations and can be faster to run than other techniques that might be used to analyze the data. This can be especially useful when working with real-time data or when the index needs to be updated frequently.

One potential disadvantage of using PCA in index construction is that it can be sensitive to the scaling of the data. If the data are not properly scaled, the results of the PCA analysis may be distorted. In addition, PCA is a linear technique, which means that it can only capture linear relationships in the data. This means that it may not be suitable for data that exhibits more complex patterns or trends. Other identified limitations of PCA include domain shape dependence, lack of stability, and the presence of sampling errors. Additionally, as the number of factors approaches the smaller of the dimensions, spurious correlations may occur, which may lead to miss-classification of smaller equity asset class instruments, according to Wold (1978) [27] points out that as the number of factors approaches the smaller of the dimensions, spurious correlations may occur. This may mean the smaller equity asset class instruments might get miss-classified. The use of PCA also requires an expert knowledge of the asset class to identify the factors.

A limitation of the PCA approach, identified by Fralet and Raftery [28], relates to computing requirements that grow at a nonlinear rate relative to the size of the groupings. This can limit the size of the data set being analysed when the researcher does not have adequate computing power. As equity assets have a large number of instruments, this is relevant. The index construction method cannot realistically be done without the relevant software.

When testing for the appropriateness of an equity asset benchmark, a dialectic approach is best.[29] This avoids accepting statistical output at face value. In time series, correlations vary over time. This was addressed by Brown and Warner (1980) [30] who showed that, when events are not clustered in time, the differences between the various methodologies are quite small. As a result, there is no evidence that existing equity construction methodologies convey any benefit over and above the FMI. We therefore consider that the PCA method is equally valid as a method in the index construction process as any other.

8. Conclusion

To conclude, we illustrate a method for using PCA to construct indices for financial markets. This approach creates a FMI in which PCA is used to assign weights to individual equities, and then using these weights to aggregate the equities into a portfolio. The FMI weights are derived and used to identify sub-sectors and their weightings within the market.

We show how PCA derived indicies can be constructed using eigenvector based weighting combined with rules that allow for changes in continuity, context, causality, and consistency. The introduction of the PCA approach extends the literature behind benchmarks, both direct and indirect. The PCA approach is a departure from common practice but represents a contribution to benchmark theory. It is an enhancement of the methodology and an important contribution to knowledge that supports investment benchmarks. In simplified form, the steps are as follows:

1. Select the equity constituents and return time series.
2. Calculate the covariance matrix of the standardized variables.

3. Compute the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors represent the directions in which the data vary the most, and the eigenvalues represent the amount of variation in the data along each eigenvector.
4. Select the top k eigenvectors, where k is the number of dimensions (variables) that you want to retain in the derived index. These eigenvectors will be used to form the derived index.
5. Assign weights to the original variables based on their contributions to the selected eigenvectors. The weight of a variable is simply the element of the corresponding eigenvector.
6. The derived FMI is then created by taking the weighted sum of the original variables. The weights are the coefficients that multiply each variable in the sum.

We suggest that the FMI methodology can be used for various asset sub-groupings and could potentially be used to replicate risk factors synthetically. The FMI is intended to address entropy issues with non-linear return time series that may impact the ability of an index to be a good proxy for the market portfolio. We fill a gap in the literature on index construction methods. Our contribution is in illustrating how equity assets can be regrouped without proxies. The use of PCA derived indices allows for synthetic replication of factor risk exposures in equity asset classes, and therefor synthetic index construction.

Conflicts of Interest: The authors declare no conflict of interest. The manuscript is a conversion of one of the chapters in the PHD thesis of Daniel Broby, a collection of essays that explore current theory and practice in respect of benchmarks for alternative asset classes.

Abbreviations

The following abbreviations are used in this manuscript:

PCA	Principal Component Analysis
FMI	Factor Model Index

References

1. Broby, D. *A guide to equity index construction*; Risk Books, 2007.
2. Malevergne, Y.; Sornette, D. Self-consistent asset pricing models. *Physica A: Statistical Mechanics and its Applications* **2007**, *382*, 149–171.
3. Daniel, K.; Grinblatt, M.; Titman, S.; Wermers, R. Measuring Mutual Fund Performance with Benchmarks. *the Journal of Finance* **1997**, *52*, 1035–1058. https://doi.org/10.1111/j.1540-6261.1997.tb02724.x.
4. Broby, D.; McKenzie, A.; Bautheac, O. Factor Model Index for Commodity Investment. *The Journal of Index Investing* **2021**, *12*, 33–52.
5. Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**. https://doi.org/10.1080/14786440109462720.
6. Meade, N.; Salkin, G. Index funds-construction and performance measurement. *Journal of the Operational Research Society* **1989**, *pp.871-879*.
7. MSCI Inc.. MSCI Global Investable Market Indexes Methodology. *Indexes* **2018**.
8. Jolliffe, A. *Principal Component Analysis*; Springer Series in Statistics, Springer-Verlag: New York, 2002. https://doi.org/10.1007/b98835.
9. Barlett, M. A note on the Statistical estimation of supply and demand relations from time series. *Econometrica (pre-1986); Evanston* **1948**.
10. Polson, N.; Tew, B. Bayesian portfolio selection: An empirical analysis of the S&P 500 index 1970–1996. *Journal of Business & Economic Statistics* **2000**, *18(2), pp.*
11. Amenc, N.; Goltz, F.; Lodh, A. Choose Your Betas: Benchmarking Alternative Equity Index Strategies. *The Journal of Portfolio Management* **2012**, *39*, 89–111. https://doi.org/10.3905/jpm.2012.39.1.088.
12. Bailey, J. Are Manager Universes Acceptable Performance Benchmarks? *The Journal of Portfolio Management* **1992**, pp. 9–13.
13. Rao, R. Separation theorems for singular values of matrices and their applications in multivariate analysis. *Journal of Multivariate Analysis* **1979**, *9*, 362–377. https://doi.org/10.1016/0047-259X(79)90094-0.
14. Partovi, H.; Caputo, M. Principal Portfolios: Recasting the Efficient Frontier. Technical report, 2004.
15. Pasini, G. A Principal Component Analysis for stock portfolio management. *International Journal of Pure and Applied Mathematics* **2017**, *115*, 153–167. https://doi.org/10.12732/ijpam.v115i1.12.
16. Shlens, J. A Tutorial on Principal Component Analysis. *Preprint arXiv:1404.1100* **2005**, *51*, 52. https://doi.org/10.1.1.115.3503.
17. Contributors, W. Principal Component Analysis. *[online] Wikipedia* **2022**.
18. Hendrickson, A.; White, P. Promax: a quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology* **1964**, *17*, 65–70.
19. Kaiser, H. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* **1958**, *23*, 187–200.
20. Roncalli, T. An Alternative Approach to Alternative Beta. *Unpublished paper* **2007**.
21. Yang, L.; Rea, W.; Rea, A. Identifying Highly Correlated Stocks Using the Last Few Principal Components. Technical report, 2015.
22. Markowitz, H.M. Foundations of portfolio theory. *The journal of finance* **1991**, *46*, 469–477.
23. Gerber, S.; Markowitz, H.M.; Ernst, P.A.; Miao, Y.; Javid, B.; Sargen, P. The Gerber statistic: a robust co-movement measure for portfolio optimization. *The Journal of Portfolio Management* **2022**, *48*, 87–102.
24. Smyth, W.; Broby, D. An enhanced Gerber statistic for portfolio optimization. *Finance Research Letters* **2022**.
25. Chao, Y.S.; Wu, C.J. Principal component-based weighted indices and a framework to evaluate indices: Results from the Medical Expenditure Panel Survey 1996 to 2011. *PloS One* **2017**, *12*, e0183997.
26. Hubert, M.; Rousseeuw, P.; Verdonck, T. Robust PCA for skewed data and its outlier map. *Computational Statistics and Data Analysis* **2009**. https://doi.org/10.1016/j.csda.2008.05.027.
27. Wold, S. *Cross-validatory estimation of the number of components in factor and principal components models*; Technometrics: - [s. l.], 1978; pp. 397–405.
28. Fraley, C.; Raftery, A. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. Technical report.
29. Broby, D. Equity index construction. *The Journal of Index Investing* **2011**, *2*, 36–39.
30. Brown, S.; Warner, J. Measuring security price performance. *Journal of Financial Economics* **1980**.