Financial Statistical Journal 2025, 8(1), 10858.
https://doi.org/10.24294/fsj10858

Article

On the use of principal components analysis in index construction

Daniel Broby[1,*], William Smyth[2]

[1] Asian Institute of Management, Makati, Metro Manila 1229, Philippines
[2] Ulster University, Londonderry BT48 7JL, United Kingdom
*** Corresponding author:** Daniel Broby, dbroby@aim.edu

CITATION

Broby D, Smyth W. On the use of principal components analysis in index construction. Financial Statistical Journal. 2025; 8(1): 10858. https://doi.org/10.24294/fsj10858

ARTICLE INFO

Received: 12 September 2024
Accepted: 15 January 2025
Available online: 14 February 2025

Abstract: This paper introduces a novel application of principal component analysis (PCA) in constructing equity indices. While PCA is well-established in other fields, its use in financial index design remains underexplored. The proposed method addresses entropy concerns in nonlinear return time series. PCA is employed to determine equity weights, using factor loadings to guide its construction. This results in a factor model index (FMI) that identifies sub-sectors and assigns data-driven weights. The FMI framework is flexible, allowing adaptation to different asset sub-groups and facilitating synthetic replication of risk factors.

Keywords: principal component analysis; index construction; correlation matrix

1. Introduction

This paper explores the application of principal component analysis (PCA) in constructing investment indices and introduces a method for the construction of a factor model index (FMI). The study is theoretical and provides a detailed overview of the process. We explain and expand on the PCA method, focusing on its application in index construction. Our approach enables the selection of sub-groupings and clustering of asset proxies based on factor exposure. We argue this results in a more flexible factor-based weighting. The method leverages the dimension reduction properties of PCA, first identified by Pearson [1]. He highlighted PCA's broad range of applications. Here, we apply PCA specifically to one of these, namely index construction. We explain how the eigenfactor of the first principal component, within an equity asset sub-class, can be used to determine index weights.

Traditional equity indices, such as those constructed through market or equal-weighted capitalization, have long been criticized for their susceptibility to biases. These include over-concentration in certain sectors or stocks. Factor-based approaches offer a solution. Those in common usage, however, rely on pre-defined factors that may not capture all dimensions of market variance. In this context, PCA presents an alternative, enabling the extraction of orthogonal factors directly from observed market returns. We explore the applicability of PCA to construct equity indices that are both empirically robust and interpretable. While PCA's theoretical advantages in other fields are well documented, the specific equity index use case addressed by our paper remains under-articulated.

While finance academics frequently use PCA as a dimensionality reduction tool, its application to index construction remains uncommon. In contrast, other disciplines have applied PCA for similar purposes [2–5]. One prominent use of PCA is in the evaluation of environmental indices, particularly in assessing water quality [6]. PCA has also been utilized to construct indices that gauge competitiveness and soundness

[7]. The finance industry, meanwhile, has developed factor indices based on the Arbitrage Pricing Model, though PCA-derived factor models offer a distinct approach. PCA not only captures the diagonal elements of a covariance or correlation matrix but also accounts for off-diagonal terms, reflecting interdependencies between assets. This dual capability allows PCA to determine both asset characteristics and corresponding FMI weights.

PCA has several useful mathematical properties for indices. The most important is that the first principal component explains the largest portion of variance of the individual equities in an index (built using the FMI approach). This corresponds to the systemic risk factor of the capital asset pricing model. Malevergne et al. [8] argue that this property makes its use consistent with the self-consistency condition, namely that a market proxy should be composed of assets whose returns it aims to explain. Thus, PCA provides a systematic way to align index construction with both underlying market dynamics and finance theory.

There have been precedents in allied financial asset classes for the use of PCA methods to define index constituents (by their common attributes). Daniel et al. [9], for example, argue that its characteristics provide a better ex-ante forecast of the cross-sectional returns of futures markets. 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. [10] develop a PCA-based index that outperforms traditional commodity indices when applied to that asset class. For a detailed empirical example of the FMI approach, readers are encouraged to consult that paper.

The PCA method we present expands the range of approaches to equity index construction found in the literature. It is designed for effective performance attribution by virtue of being grouped into relevant sub-sectors. The techniques for achieving such sub-sector divisions are discussed in Meade and Salkin [11]. A practical example of that widely used approach is provided in the MSCI Methodology Booklet [12]. PCA can similarly be used to enhance index construction through systematic asset grouping.

Background

The use of PCA is well documented in disciplines other than finance. It is explained by Jolliffe [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. PCA is also used in time series analysis for tasks such as seasonal adjustment. Its potential in financial time series comes from its transformation of the original data into a set of orthogonal components. In the context of equity markets, this means that each equity can be represented as a linear weighted combination of the available instruments. This enables the resultant FMI to be constructed based on shared variance patterns.

PCA is therefore an established procedure in academic investigation. It 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 has been used to show correlated response and to identify predictor variables. PCA has, however, not been used previously to construct indices for equity assets. That said, it was used to index commodity prices by Barlett [13] and, as stated, by Broby et al. [14]. The former applied PCA to a time series of cotton prices over the period 1924–1938 in order to

better understand the nature of their returns. The latter applied it to commodity futures prices from 2008 to 2016.

In the literature there are several hierarchical models similar to PCA that are used to create optimal weights, as described by Polsen and Tew [15]. 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 infrequently priced asset classes, such as real estate. That said, most current methods, as explained, rely on representation rather than replication. Amenc et al. [16] explain that in the index replication stage, one should have two steps in the construction process, these being constituent and weighting scheme selection. This therefore has to be applied to the way we construct the FMI.

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 [17]. It also addresses issues with those indices that are constructed without attention to correlation, co-variance skew and kurtosis. Further, the use of FMI can help explain variability through insights on factors and correlation. The theorems behind PCA, matrix algebra and multivariate analysis are explained well by Rao [18], amongst others. It can be used on investment proxies, thereby filling an identified need in the literature as it relates to equity assets. We suggest that PCA can help define an appropriate index as a result of its robustness. 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 extracts the first principal component, which accounts for the greatest variance, followed by successive components that explain decreasing amounts of variance. This process allows PCA to be applied to equities, whose returns are driven by distinct factors. It enables a reevaluation of variances, covariances, and correlations.

In a portfolio context, Partovi et al. [19] demonstrate how PCA can reshape the efficient frontier by constructing portfolios from uncorrelated assets. While most assets exhibit some degree of correlation, their study shows that PCA simplifies portfolio structure and offers a more transparent framework for asset allocation. Similarly, Pasini [20] applies PCA to equity portfolio analysis, using it to assess how much a time series deviates from being a sequence of independent and identically distributed observations with finite mean and variance. Although this differs from the approach proposed in this paper, it provides valuable insight. Pasini finds that the first principal component typically represents the market factor, while the second principal component often captures most of the remaining risk.

In summary, the PCA approach constructs indices by identifying common components and assigning weights based on eigenvalue optimization. This contrasts with traditional index construction, which typically relies on market capitalization-

based weighting. PCA serves as a classification method by creating an index from factor weights, which can be evaluated for optimality. When applied to return time series, PCA enhances index robustness by reducing data complexity, enabling the creation of interpretable factors with assigned weights. This process is driven by an orthogonal transformation that isolates uncorrelated components, ensuring a systematic and transparent index construction framework.

2. Materials and methods—Factor model index (FMI)

It is easier to understand PCA visually. **Figure 1** presents a geometric representation based on two variables, X1 and X2. 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 that 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 of the two X1 and X2 variables.

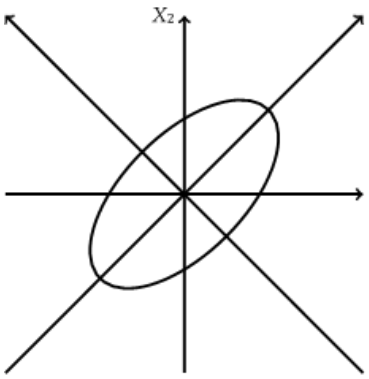

Figure 1. A representation of the first and second rotations.

Figure 1. 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 that 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.

$$\Sigma_{X_1,X_2} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (1)$$

a_{11} and a_{21} denote the weights from the first eigenvector of Σ; a_{12} and a_{22} are the weights from the second eigenvector. It can be represented by a 2 × 2 orthogonal (or rotation) matrix *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:

$$T = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \cos(90+\theta) \\ \cos(\theta-90) & \cos(\theta) \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \quad (2)$$

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.

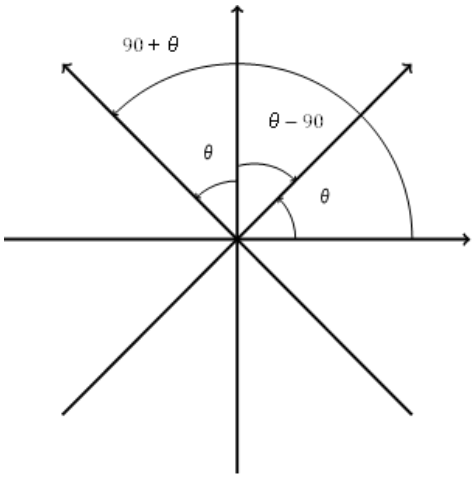

Figure 2. A PCA transformed axis showing the cosines of the angles on the horizontal and vertical axes.

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.

2.1. Orthogonal transformation

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 [21]. 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 is traditionally done in index construction through sampling rather than statistical technique.

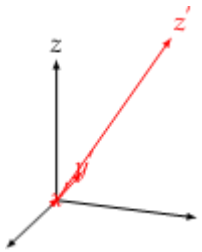

Figure 3. Non-orthogonal 3D coordinate systems.

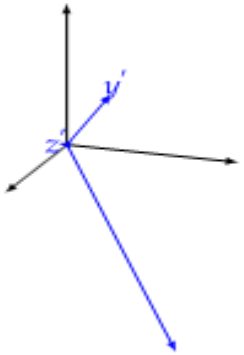

Figure 4. Vectors in 3d and non–orthogonal basis vectors.

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 (see [22] for formulas and proof). This is illustrated mathematically as where *w* equates to:

$$w_{(1)} = arg\max_{\|w\|=1}\left\{\sum_i \left(t_{1(i)}\right)^2\right\} = arg\max_{\|w\|=1}\left\{\sum_i \left(x_{(i)} \cdot w\right)^2\right\} \quad (3)$$

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

This is represented in matrix form as:

$$w_{(1)} = arg\max_{\|w\|=1}\{\|Xw\|^2\} = arg\max_{\|w\|=1}\{w^T X^T Xw\} \quad (4)$$

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.

$$\hat{X}_k = X - \sum_{s=1}^{k-1} X\, w_{(s)} ws\mathrm{T} \tag{5}$$

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:

$$w_{(k)} = arg\max_{\|w\|=1} \left\{ \left\| \hat{X}_k w \right\|^2 \right\} = arg\max \left\{ \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w} \right\} \tag{6}$$

where: $w_{(k)}$ = Weighting load of the kth factor.

The results can then be presented as a set of weights that 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 of the market model, which has a common systemic factor and various stock-specific factors. The creation of a common asset factor means that the PCA approach has a theoretical link to the market proxy. That proxy is derived from the market model and mean variance portfolio theory. It is 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.

2.2. Deriving factors from principal components

The properties of the PCA output mean that it is possible to derive investment factors. A factor is a measurable characteristic or attribute that explains variations in returns across securities. For example, in finance, systemic risk is often considered a common factor to all equities.

The process begins with a matrix representing the equity asset class opportunity set. PCA decomposes this matrix into principal components, where each component represents a factor explaining a portion of the total variance. Mathematically, this can be expressed using a stock matrix with multiple factor loadings, as illustrated in the equation below:

$$X = \mu + LF + \epsilon \tag{7}$$

where:

X: vector of the equity asset class returns.

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{\mathrm{n}} \end{pmatrix}.$$

μ: X is drawn from a universe of stocks with a mean vector.

$$\mu = \begin{pmatrix} -1 \\ -2 \\ \vdots \\ -n \end{pmatrix}.$$

L: $k \times n$ {matrix of factor loadings.} F =

$$\begin{pmatrix} l_{1,1} & l_{1,2} & \cdots & l_{1,5} \\ l_{2,1} & l_{2,2} & \cdots & l_{2,5} \\ \vdots & \vdots & \ddots & \vdots \\ l_{n,1} & l_{n,2} & \cdots & l_{n,5} \end{pmatrix}.$$

F: vector of common factors. f =

$$(f_1 \quad f_2 \quad \cdots \quad f_5).$$

ϵ: vector of errors (specific factors). ϵ =

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Using the PCA approach, the variance for the equity asset class ith is going to be equal to the total of the squared loadings and the variances of the instruments:

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

where: $\sum_{j=1}^{n} l_{i,j}^2$: communality of the equity asset class i,

$$\sum_{j=1}^{n} l_{i,j}^2.$$

ψ_i: specific factor for the equity asset class i.

$$X = \mu + \text{LTT}'F + \epsilon = \mu + L^*F^* + \epsilon.$$

For the orthogonal matrix T.

This approach was first presented in a generalized way [23]. It should be used in preference to the varimax orthogonal method, suggested by Kaiser [24]. In this way, the oblique solution is effectively obtained by trial and error, increasing the larger loads and reducing the smaller ones. A good knowledge of the time series of the equities in question is helpful. It helps to identify relevant factors accurately and interpret factor loadings within the context of financial market behavior.

The connection between factors and eigenvalues, which determines their use in index weighting, was explained by Roncalli [25]. His method applies PCA to the selected index universe, generating a risk factor from the covariance matrix. The first eigenvalue corresponds to the market risk factor, capturing the largest variance in the asset set. Subsequent eigenvectors represent additional common risk factors, each explaining a smaller portion of the variance. This hierarchical structure allows for systematic factor identification and weighting in index construction.

2.3. Linking to finance theory

We now turn to linking this to finance theory. In the context of the Capital Asset Pricing Model (CAPM), the first principal component derived from PCA serves as a proxy for systemic market risk. The creation of a common asset factor means that the PCA approach has a theoretical link to the market proxy. That proxy is derived from the market model and mean variance portfolio theory. It is 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. This component captures the largest share of variance across assets, representing the common risk that affects all securities. Subsequent principal components reflect specific risk factors unique to subsets of assets. These factors capture idiosyncratic risks not explained by the market factor. In this respect, the method is aligning with CAPM's framework where total risk is divided into market (systematic) and asset-specific (unsystematic) components. Thus, PCA provides a data-driven way to decompose asset returns into systematic and specific risks.

Building on this concept, Yang et al. [26] applied PCA to interpret the covariance matrix of asset returns. Interestingly, they found that even the last few principal components hold meaningful information, as they reveal instruments with nonlinear correlations. This finding is significant given the ongoing debate in finance about the number of factors needed to explain asset returns. It suggests that equity asset factors can be identified directly through PCA, without relying solely on traditional models.

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.

These properties mean that it is possible to use PCA and still create a mean variance optimal index. This can then generate a portfolio of assets representing an index that has been optimized to maximize expected returns while at the same time minimizing risk. This 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.

Achieving this outcome requires interpreting the covariance matrix (or historical covariance matrix) in the context of factor risk [27]. Alternatively, this can be done using shrinkage techniques applied to the sample covariance matrix or by employing common co-movement measures such as the Gerber statistic [28] and the modified Gerber statistic [29]. These methods enable the construction of indices using only the

return time series, without needing additional market information. This data-driven approach ensures more robust and adaptable index construction.

3. Results and discussion of method

We now focus on how to reorganize the results for index construction, so that a set of clear rules can be established. This is essential in financial markets for replication. The underlying instruments must first be transformed from their raw form using a variance reduction method. The process begins with applying PCA, followed by analyzing the equity asset returns using the variance-covariance matrix. Based on insights from the relevant literature, the equities should reveal several factors that explain the variation in returns.

Once the matrix is generated, its outputs are applied through factor analysis. This allows the creation of an index where securities are weighted according to their factor loadings. Additionally, a periodic ranking, either monthly or yearly, is then calculated. This enables a rebalancing process within a specified holding period. As a systematic approach, this ensures the index reflects changing market dynamics while maintaining diversification.

Once the common components have been established, it is possible to determine the factors present using an 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 then 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.

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. Portfolio and index are being used interchangeably in this context. At the end of the period, the FMI weights are updated and the portfolio re-balanced using the same procedure.

In the next stage, each equity receives a weighting equal to the nth ratio of its load relative to the sum of the loads contained in the sub-portfolio. The resulting group of sub-portfolios can then be 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 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 will result in groups of equity asset instruments as single factors. We stress that the underlying interactions between the factors should be thought of as similar to creating sectors in traditional index construction.

3.1. Data transformation and PCA application

This section provides an illustrative framework, supported by tables and equations to clarify each step. The process begins with the application of PCA to the equity asset return data, transforming raw instruments into principal components. The principal components are derived from the variance-covariance matrix, capturing common patterns in the returns. These components explain the variation in returns across the asset universe, allowing for dimensionality reduction.

The PCA output is analyzed to identify factors, with eigenvalues indicating the relative importance of each factor. Factor loadings (β) are calculated for each equity instrument, representing their correlation with the identified factors. A threshold of $\beta >$ 0.3 (shown as * in **Table 1**) was suggested by Chao and Wu [30] and is used to determine significant factor contributions. An example is shown in **Table 1**.

Table 1. Factor loadings.

Equity Instrument	Factor 1	Factor 2	Factor 3	Significant Factor
Security A	0.45*	0.10	0.05	Factor 1
Security B	0.20	0.35*	0.15	Factor 2
Security C	0.05	0.25	0.40*	Factor 3

Note: The highest factor loading in each column is highlighted with an *. This is the Significant Factor as shown in the final column.

Equity instruments are allocated into sub-portfolios based on their significant factor loadings. For example, a sub-portfolio for Factor 1 would include securities predominantly influenced by Factor 1, such as Security A, while a sub-portfolio for Factor 2 would consist of securities with substantial loadings on Factor 2, such as Security B. Each sub-portfolio represents a distinct underlying factor identified through PCA, ensuring that securities are grouped according to common risk exposures. The weights of individual securities within each sub-portfolio are then calculated as:

$$w_i = \frac{\text{Load}_\text{i}}{\sum Loads_{Sub-Portfolio}} \tag{8}$$

The sub-portfolios in **Table 2** are aggregated into a single portfolio. The weight of each sub-portfolio is determined by the variance explained by its associated factor:

Table 2. The relationship of variance explained to the various factors.

Factor	Variance Explained (%)	Cumulative Variance (%)
Factor 1	40%	40%
Factor 2	30%	70%
Factor 3	20%	90%

The process is repeated for each observation period, with weights updated and portfolios rebalanced. The rebalancing period (e.g., three or six months) is chosen to balance computation ease and transaction costs. The FMI weights are recalculated at each rebalancing date, ensuring the index reflects evolving market dynamics.

To refine the factor structure, an oblique rotation is applied. Oblique rotation allows factors to be correlated, reflecting the inherent interdependencies often observed in real-world data. This enhances the robustness of constructed indices by improving factor interpretability and aligning them with the underlying data structure. Additionally, oblique rotation minimizes cross-loadings, clarifying the distinctiveness of factors and ensuring each captures a specific dimension of variability. However, this approach introduces complexity due to correlated factors, risks of overfitting, and challenges in maintaining generalizability across datasets. This step allows for some correlation between factors, providing a clearer decomposition of variance and grouping equity instruments into cohesive sectors. This step is analogous to creating traditional index sectors but is data-driven based on the PCA output.

The resulting index in **Table 3** comprises aggregated sub-portfolios, each weighted by their explained variance. These could be, for example, sectors. The final portfolio reflects a systematic combination of securities optimally weighted by their factor contributions.

Table 3. Derived equity weights by sub portfolio and total portfolio.

Equity Instrument	Sub-Portfolio Weight	Portfolio Weight
Security A	50%	20%
Security B	30%	15%
Security C	20%	10%

This systematic approach ensures that the constructed index reflects the primary systemic and specific risks while maintaining diversification and minimizing rebalancing costs. The methodology provides a robust framework for data-driven index construction.

To demonstrate the practical application of PCA in equity index construction, we conducted an empirical analysis using a simulated equity dataset representing stocks from seven major GICS sectors over a five-year period. The dataset was based on 20 stocks and included Information Technology, Healthcare, Financials, Energy, Consumer Discretionary, Industrials, and Real Estate sectors. **Figure 5** below illustrates how the weights dynamically shift each year to reflect the changing factors. The PCA methodology was applied to the monthly return data to extract orthogonal factors, with the first principal component serving as the primary driver for the PCA-derived index.

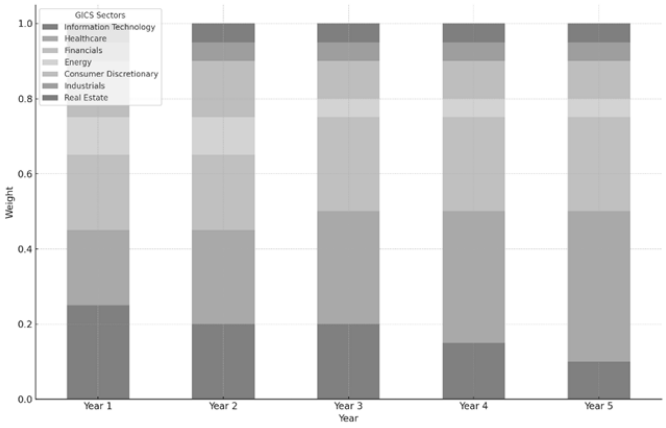

Figure 5 Sector weight changes of the simulated rebalanced portfolios.

Figure 5: This chart illustrates the evolving weights of major GICS sectors, including Information Technology, Healthcare, Financials, Energy, Consumer Discretionary, Industrials, and Real Estate, over a five-year period. The greyscale bars represent the proportion of each sector's contribution to the overall portfolio, showcasing shifts in market dynamics, such as the increasing dominance of healthcare and the relative stability of other sectors like Energy and Real Estate.

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 separate 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:

$$\mathrm{PCA}_{ri,t} = \beta_1 Fv_t + \beta_2 Fv_t + \cdots + \beta_n Fv_t \tag{9}$$

where: $\mathrm{PCA}_{ri,t}$ = FMI, the excess return of portfolio *i* in month *t*, *Fv* = Factor identified by eigenvalues.

To enhance the robustness of the index and manage rebalancing costs, we recommend using a multi-year observation period with a rebalancing frequency of three or six months. This approach helps to smooth out the effects of rebalancing and reduces transaction costs, ensuring the index remains both cost-efficient and reflective of underlying market dynamics.

In summary, the eigenvalues indicate the relative importance of each factor, while the equation uses factor loadings (β) as weights assigned to those factors. This approach allows for the stepwise construction of an index, aiming to identify the combination of variables that provides the best model fit based on a chosen metric.

3.2. Discussion

A significant contribution of the PCA-based approach to index construction lies in its ability to address the non-normal distribution of equity asset time series. Traditional index models often assume normally distributed returns, which can lead to inaccuracies in portfolio analysis. In contrast, the Factor Model Index (FMI) constructed using PCA accounts for skewed and non-Gaussian data structures. This makes the approach particularly relevant in the context of equity indices, where asset

returns frequently exhibit skewness and kurtosis. Hubert et al. [31] demonstrated that robust PCA methods remain effective even when faced with such data irregularities.

We summarise the differences between PCA based indices and traditional indices in **Table 4**. When testing for the appropriateness of an equity asset benchmark, a dialectic approach is best [32]. This avoids accepting statistical output at face value. In time series, correlations vary over time. This was addressed by Brown and Warner [33] 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.

Table 4. A comparison of PCA based indices with traditional and factor based indices.

Comparison Criteria	PCA-Based Indices	Traditional Market-Capitalization Indices	Other Factor-Based Indices
Construction Method	Based on dimensionality reduction of data using PCA.	Weighted by market capitalization of constituent stocks.	Based on pre-defined factors (e.g., value, growth).
Weighting Approach	Data-driven factor loadings derived from PCA.	Proportional to the market value of listed companies.	Factor exposure weights determined by specific metrics.
Flexibility	Highly flexible and adapts to changes in data correlations.	Limited to market cap adjustments and periodic updates.	Limited by pre-defined factors and their calculation rules.
Robustness to Noise	Handles noise well by focusing on dominant data patterns.	Sensitive to market volatility and extreme stock moves.	Sensitive to factor misestimation or market shifts.
Transparency	Complex and requires detailed knowledge of PCA interpretation.	Easy to understand due to straightforward weighting rules.	Moderate, it depends on the factor definitions.
Application	Suitable for analyzing complex, multi-dimensional datasets.	Commonly used for broad market tracking and benchmarking.	Widely used for targeted investment strategies.
Computational Requirements	High; requires advanced tools and processing power.	Low; straightforward calculations based on market data.	Moderate; depends on the complexity of factor calculations.

The theoretical justification for applying PCA stems from its capacity to represent complex data in reduced dimensions. While PCA inherently assumes that the data approximates a multivariate Gaussian distribution, this assumption serves primarily as a simplification for variance decomposition. In practice, PCA's use of eigenvalues and eigenvectors to capture key data patterns allows it to work well even when returns deviate from normality.

Another contribution of PCA to index construction is its potential for index replication without requiring direct investment in the underlying securities. By reducing the dimensionality of financial time series, PCA projects asset returns onto a lower-dimensional space while preserving essential information. This process helps eliminate data redundancy, making analysis and interpretation more efficient. The reduced representation enables the creation of synthetic indices that approximate the behavior of complex portfolios, supporting efficient market exposure with fewer underlying components. This capability has practical implications for financial product design, particularly in developing exchange-traded funds (ETFs) and other passive investment vehicles.

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 [34], 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 expert knowledge of the asset class to identify the factors.

A further limitation of the PCA approach, identified by Fralet and Raftery [35], 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 analyzed 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.

4. Conclusion

This paper presents a method for constructing financial market indices using Principal Component Analysis (PCA). The approach results in a Factor Model Index (FMI), where PCA assigns weights to individual equities based on factor loadings, enabling the aggregation of these equities into a portfolio. The derived FMI weights facilitate the identification and weighting of sub-sectors within the broader market, offering a data-driven approach to portfolio construction.

We demonstrate how PCA-derived indices can be constructed through eigenvector-based weighting, complemented by rules ensuring continuity, context, causality, and consistency. This approach represents a significant departure from conventional index construction methods, extending the theoretical framework of financial benchmarks. By incorporating PCA, the method enhances traditional benchmark theory and contributes to the broader literature on index design.

The steps involved in constructing a PCA-based index can be summarized as follows:

1) Identify the relevant equity instruments and obtain their historical return time series.

2) Standardize the return series and compute the covariance matrix to measure the relationships between asset returns.
3) Derive the eigenvalues and eigenvectors from the covariance matrix. The eigenvectors indicate the principal directions of data variance, while the eigenvalues quantify the amount of variance explained by each eigenvector.
4) Choose the top k eigenvectors based on the magnitude of their corresponding eigenvalues. These eigenvectors determine the dimensions retained in the index.
5) Assign weights to the original variables based on their contributions to the selected eigenvectors. Each variable's weight corresponds to its eigenvector element.
6) Create the index by calculating the weighted sum of the original variables using the assigned weights.

Future research could focus on conducting a sensitivity analysis to examine how different parameter choices, such as rebalancing frequency and the threshold for factor significance, impact the constructed index. This analysis would provide valuable insights into the stability and robustness of the proposed methodology, ensuring its reliability across various scenarios and applications.

We suggest that PCA-based indices offer a powerful tool for ESG investing. They enable the integration of often subjective metrics into a more objective index. The eigenvalues can be used to identify and emphasize the most influential ESG factors, moving away from the traditional negative screening common in such indices. We further suggest that this methodology can be applied across various asset sub-groupings, supporting the synthetic replication of risk factors.

In conclusion, the FMI framework addresses entropy issues commonly associated with non-linear return time series, improving the index's ability to approximate the market portfolio. This method fills a gap in the literature on index construction by offering a systematic process that does not rely on traditional proxies.

Author contributions: Conceptualization, DB; methodology, DB; software, DB; validation, DB and WS; formal analysis, DB and WS; investigation, DB; resources, DB and WS; data curation, DB; writing—original draft preparation, DB; writing—review and editing, WS; visualization, DB; supervision, DB; project administration, DB; funding acquisition, DB. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Pearson K. LIII. 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; 2(11): 559-572. doi: 10.1080/14786440109462720
2. Bucherie A, Hultquist C, Adamo S, et al. A comparison of social vulnerability indices specific to flooding in Ecuador: principal component analysis (PCA) and expert knowledge. International Journal of Disaster Risk Reduction. 2022; 73: 102897. doi: 10.1016/j.ijdrr.2022.102897
3. Hammoumi D, Al-Aizari HS, Alaraidh IA, et al. Seasonal Variations and Assessment of Surface Water Quality Using Water Quality Index (WQI) and Principal Component Analysis (PCA): A Case Study. Sustainability. 2024; 16(13): 5644. doi: 10.3390/su16135644

4. Dai PF, Xiong X, Zhou WX. A global economic policy uncertainty index from principal component analysis. Finance Research Letters. 2021; 40: 101686. doi: 10.1016/j.frl.2020.101686
5. Zheng C, Rahman MA, Hossain S, et al. Construction of a composite fintech index to measure financial inclusion for developing countries. Applied Economics. 2024; 56(52): 6498-6515. doi: 10.1080/00036846.2024.2313600
6. Ghazali N, Ali ZM. Principal Component Analysis Approach in Klang River Water Quality Index Modelling. Environment and Ecology Research. 2023; 11(1): 165-182. doi: 10.13189/eer.2023.110112
7. Xiao LP, Liu Y. Role of Principal Component Analysis (PCA) in the Evaluation of Competitiveness of Small Firms. Advanced Materials Research. 2014; 926-930: 3954-3957. doi: 10.4028/www.scientific.net/amr.926-930.3954
8. Malevergne Y, Sornette D. Self-consistent asset pricing models. Physica A: Statistical Mechanics and its Applications. 2007; 382(1): 149-171. doi: 10.1016/j.physa.2007.02.076
9. Daniel K, Grinblatt M, Titman S, et al. Measuring Mutual Fund Performance with Characteristic-Based Benchmarks. The Journal of Finance. 1997; 52(3): 1035-1058. doi: 10.1111/j.1540-6261.1997.tb02724.x
10. Broby D, McKenzie A, Bautheac O. Factor Model Index for Commodity Investment. The Journal of Index Investing. 2021; 12(3): 33-52. doi: 10.3905/jii.2021.1.110
11. Meade N, Salkin GR. Index Funds—Construction and Performance Measurement. Journal of the Operational Research Society. 1989; 40(10): 871-879. doi: 10.1057/jors.1989.155
12. MSCI. MSCI Global Investable Market Indexes Methodology. MSCI; 2018.
13. Bartlett MS. A Note on the Statistical Estimation of Supply and Demand Relations from Time Series. Econometrica. 1948; 16(4): 323. doi: 10.2307/1909273
14. Broby D. A guide to equity index construction, 2 ed. Risk Books; 2024.
15. Polson NG, Tew BV. Bayesian Portfolio Selection: An Empirical Analysis of the S&P 500 Index 1970-1996. Journal of Business & Economic Statistics. 2000; 18(2): 164. doi: 10.2307/1392554
16. Amenc N, Goltz F, Lodh A. Choose Your Betas: Benchmarking AlternativeEquity Index Strategies. The Journal of Portfolio Management. 2012; 39(1): 88-111. doi: 10.3905/jpm.2012.39.1.088
17. Bailey JV. Are Manager Unvierses Acceptable Performance Benchmarks? The Journal of Portfolio Management. 1992; 18(3): 9-13. doi: 10.3905/jpm.1992.9
18. Rao R. Separation theorems for singular values of matrices and their applications in multivariate analysis. Journal of Multivariate Analysis. 1979; 9(3): 362-377. doi: 10.1016/0047-259X(79)90094-0
19. Partovi H, Caputo M. Principal Portfolios: Recasting the Efficient Frontier. Technical report; 2004.
20. Pasini G. Principal component analysis for stock portfolio management. International Journal of Pure and Apllied Mathematics. 2017; 115(1). doi: 10.12732/ijpam.v115i1.12
21. Shlens, J. A Tutorial on Principal Component Analysis. Preprint arXiv; 2005.
22. Contributors W. Principal Component Analysis. Wikipedia; 2022.
23. Hendrickson AE, White PO. Promax: A quick method for rotation to oblique simple structure. British Journal of Statistical Psychology. 1964; 17(1): 65-70. doi: 10.1111/j.2044-8317.1964.tb00244.x
24. Kaiser HF. The Varimax Criterion for Analytic Rotation in Factor Analysis. Psychometrika. 1958; 23(3): 187-200. doi: 10.1007/bf02289233
25. Roncalli T, Teiletche J. An Alternative Approach to Alternative Beta. SSRN Electronic Journal; 2007.
26. Yang L, Rea W, Rea A. Identifying Highly Correlated Stocks Using the Last Few Principal Components. Technical report; 2015.
27. Markowitz HM. Foundations of Portfolio Theory. The Journal of Finance. 1991; 46(2): 469-477. doi: 10.1111/j.1540-6261.1991.tb02669.x
28. Gerber S, Markowitz HM, Ernst PA, et al. The Gerber Statistic: A Robust Co-Movement Measure for Portfolio Optimization. The Journal of Portfolio Management. 2021; 48(3): 87-102. doi: 10.3905/jpm.2021.1.316
29. Smyth W, Broby D. An enhanced Gerber statistic for portfolio optimization. Finance Research Letters. 2022; 49: 103229. doi: 10.1016/j.frl.2022.103229
30. Chao YS, Wu CJ. Principal component-based weighted indices and a framework to evaluate indices: Results from the Medical Expenditure Panel Survey 1996 to 2011. Podobnik B, ed. PLOS ONE. 2017; 12(9): e0183997. doi: 10.1371/journal.pone.0183997

31. Hubert M, Rousseeuw P, Verdonck T. Robust PCA for skewed data and its outlier map. Computational Statistics & Data Analysis. 2009; 53(6): 2264-2274. doi: 10.1016/j.csda.2008.05.027
32. Broby D. Equity Index Construction. The Journal of Index Investing. 2011; 2(2): 36-39. doi: 10.3905/jii.2011.2.2.036
33. Brown SJ, Warner JB. Measuring security price performance. Journal of Financial Economics. 1980; 8(3): 205-258. doi: 10.1016/0304-405X(80)90002-1
34. Wold S. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics. 1978; 20(4): 397-405. doi: 10.1080/00401706.1978.10489693
35. Fraley C, Raftery A. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal. 1998; 41(8): 578-588. doi: 10.1093/comjnl/41.8.578