
Crypto project

Name: Yihang Yan, Tomas Gudmundsson, Nathaniel Stein

Collaborators: None

1

(a)

2 - Principal Components Analysis

(a) Introduction

Principal components analysis (PCA) is a multivariate technique that analyzes a data in which observations are described by several inter-correlated quantitative dependent variables. This technique extracts the important information from the data, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations.

Factor models that offer explanations of stock returns and correlations have been very popular in finance. PCA is unlike traditional factor models because the factors it creates do not usually have an economic interpretation. Rather, the components(factors) constructed in PCA are built to have special statistical characteristics:

- Each component accounts for as much variation in the underlying data as possible.
- Each component is uncorrelated with every other factor.
- Principal components elucidate the dominant combinations of variables within the covariance structure of the data.

Mathematically, pca depends upon the eigen-decomposition of positive semi-definite matrices and upon the singular value decomposition (svd) of rectangular matrices.

(b) Mathematics of Principal Components

Singular Value Decomposition (SVD)

Any real symmetric $m \times m$ matrix A has a spectral decomposition of the form,

$$A = U \Delta U^T \quad (1)$$

where U is an orthonormal matrix (matrix of orthogonal unit vectors: $U^T U = I$ or $\sum_k U_{ki} U_{kj} = \delta_{ij}$) and Δ is a diagonal matrix. The columns of U are the eigenvectors of matrix A and the diagonal elements of Δ are the eigenvalues. If A is positive-definite, the eigenvalues will all be positive. Multiplying with U , equation 1 can be re-written to,

$$AU = U \Delta U^T U = U \Delta$$

This can be written as a normal eigenvalue equation by defining the i th column of U as u_i and the eigenvalues as $\lambda_i = \Delta_{ii}$:

$$A u_i = \lambda_i u_i$$

Let's look at more general case. An unsymmetrical ($n \times m$) matrix, where $n \geq m$ B has the decomposition,

$$X = U \Delta V^T$$

where U is a $n \times m$ matrix with orthonormal columns ($U^T U = I$), while V is a $m \times m$ orthonormal matrix ($V^T V = I$), and Δ is a $m \times m$ diagonal matrix with positive or zero elements, called the singular values.

From B we can construct two positive-definite symmetric matrices, BB_T and $B_T B$, each of which we can decompose

$$\begin{aligned} BB_T &= U \Delta V_T V \Delta U_T = U \Delta^2 U_T \\ B_T B &= V \Delta^2 V_T \end{aligned}$$

We can now show that BB_T which is $n \times n$ and $B_T B$ which is $m \times m$ will share m eigenvalues and the remaining $n - m$ eigenvalues of BB_T will be zero.

Using the decomposition above, we can identify the eigenvectors and eigenvalues for BB_T as the columns of U and the squared diagonal elements of Δ , respectively. Denoting one such eigenvector by v and the diagonal element by γ , we have:

$$\begin{aligned} B_T B v &= \gamma^2 v \\ BB_T B v &= \gamma^2 B v \end{aligned}$$

This means that we have an eigenvector $u = B v$ and eigenvalue γ^2 for BB_T as well, since:

$$(BB_T) B v = \gamma^2 B v$$

We have now shown that B_TB and BB_T share m eigenvalues.

In order to prove that the remaining $n - m$ eigenvalues of BB_T is zero. We need to consider an eigenvector for BB_T , u_\perp : $BB_T u_\perp = \beta_\perp u_\perp$ which is orthogonal to the m eigenvectors u_i already determined, i.e. $U_T u_\perp = 0$. Using the decomposition $BB_T = U\Delta^2 U_T$, we immediately see that the eigenvalues β_\perp must all be zero,

$$BB_T u_\perp = U\Delta^2 U_T u_\perp = 0u_\perp$$

Principal component analysis (PCA) by SVD

We denote the matrix of eigenvectors sorted according to eigenvalue by \hat{U} and we can then PCA transformation of the data as $Y = \hat{U}^T X$. The eigenvectors are called the principal components. By selecting the first d rows of Y , we can project the data from n down to d dimensions.

We decompose X using SVD, i.e.

$$X = U\Delta V^T$$

and find that we can write the covariance matrix as

$$C = \frac{1}{n} X X^T = \frac{1}{n} U \Delta^2 U^T$$

Following from the fact that SVD routine order the singular values in descending order we know that, if $n < m$, the first n columns in U corresponds to the sorted eigenvalues of C and if $m \geq n$, the first m corresponds to the sorted non-zero eigenvalues of C . The transformed data can thus be written as:

$$Y = \hat{U}^T X = \hat{U}^T U \Delta V^T$$

where $\hat{U}^T U$ is a simple $n \times m$ matrix which is one on the diagonal and zero everywhere else. So we can write the transformed data in terms of the SVD decomposition of X .

Finding the components

In pca, the components are obtained from the singular value decomposition of the dataset X . Specifically, with $X = U\Delta V^T$ (equation 1), the matrix of factor scores, denoted F is obtained as

$$F = U\Delta \tag{2}$$

The matrix V gives the coefficients of the linear combinations used to compute the factors scores. This matrix can also be interpreted as a projection matrix because multiplying X by V gives the values of the projections of the observations on the principal components. This can be shown as:

$$F = U\Delta = U\Delta V V^T = X V \tag{3}$$

The components can be represented geometrically by the rotation of the original axes. Each of these components will be linear combinations of the observed variables we have in our data, and will be orthogonal to each other. That is, each components is independent of each other, and variation in one is unrelated to variance in another.

Contribution of an observation to a component

The eigenvalue associated to a component is equal to the sum of the squared factor scores for this component. Therefore, the importance of an observation for a component can be obtained by the ratio of the squared factor score of this observation by the eigenvalue associated with that component. This ratio is called the contribution of the observation to the component. Formally, the contribution of observation i to component l is denoted $ctr_{i,l}$, it is obtained as:

$$ctr_{i,l} = \frac{f_{i,l}^2}{\sum f_{i,l}^2} = \frac{f_{i,l}^2}{\lambda_l}$$

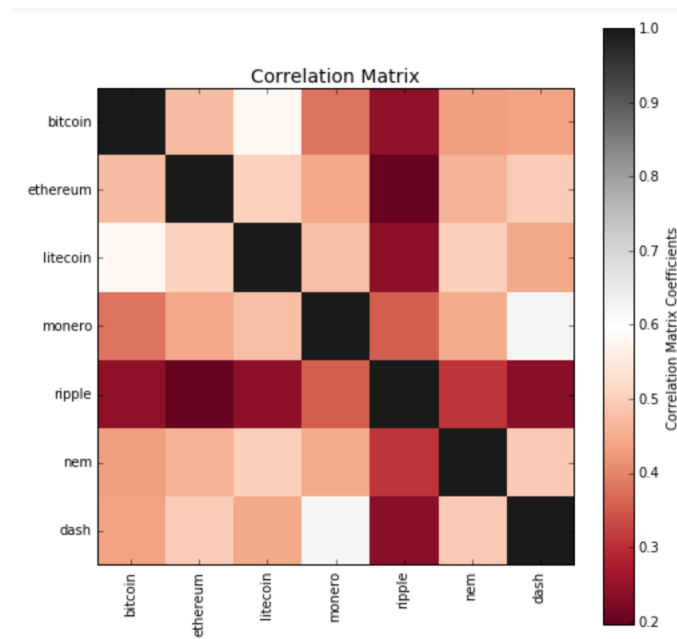
where λ_l is the eigenvalue of the l th component. The value of a contribution is between 0 and 1 and, for a given component, the sum of the contributions of all observations is equal to 1. The larger the value of the contribution, the more the observation contributes to the component.

(d) Implementation

Our dataset contains all the daily details of the crypto-markets as they close for top seven crypto currencies in terms of market cap and their tokens listed on CoinMarketCaps historical tables. Daily return for each type of currencies are calculated based on its closing prices.

we normalized time-series data of daily return of seven major crypto-currencies and the correlation matrix of normalized dataset serves as an input for PCA. The correlation matrix typically used instead of the covariance matrix. However, the eigendecomposition of the covariance matrix (if the input data was standardized) yields the same results as a eigendecomposition on the correlation matrix, since the correlation matrix can be understood as the normalized covariance matrix.

- correlation matrix:



[Interpretation of correlation diagram...]

Computing eigenvectors and corresponding eigenvalues

We will use Jacobi method to find eigenvalues and eigenvectors of correlation matrix since this method is a fairly robust way to extract all of the eigenvalues and eigenvectors of a symmetric matrix. The method is based on a series of rotations, called Jacobi or Givens rotations, which are chosen to eliminate off-diagonal elements while preserving the eigenvalues. Details of Jacobi method will be covered in Section xxx.

[report eigen values and vectors result]

We can check if the eigenvector-eigenvalue calculation is correct by using the equation:

$$M_{cov}u_i = \lambda_i u_i$$

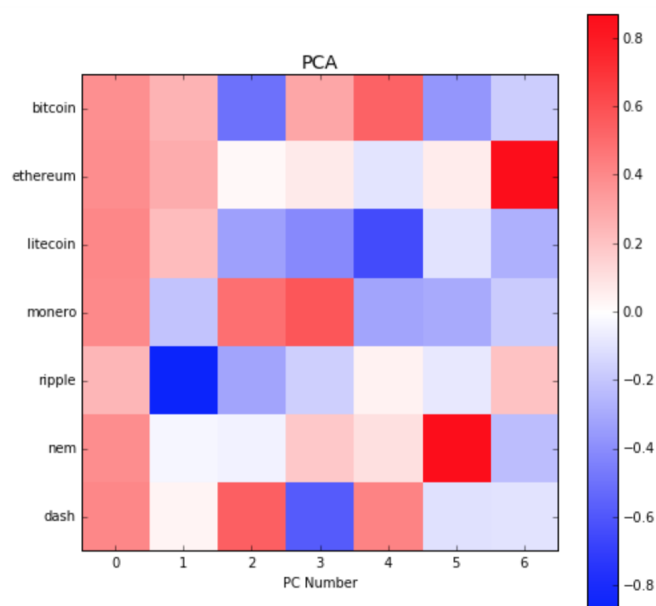
where

M_{cov} = covariance matrix,

u_i = the i th column of eigenvector matrix,

λ_i = eigenvalue associated with u_i

and generate Principal Component Loading diagram:



[Interpretation of PCA diagram ...]

Sorting Eigenpairs & Explained Variance

In order to decide which eigenvector(s) can be dropped without losing too much information for the construction of lower-dimensional subspace, we need to inspect the corresponding eigenvalues: The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones that can be dropped. In order to do so, the common approach is to rank the eigenvalues from highest to lowest in order to choose the top k eigenvectors.

After sorting the eigenpairs, the next question is how many principal components are we going to choose for our new feature subspace? A useful measure is the proportion of variance, which can be calculated from the eigenvalues. The explained variance tells us how much information (variance) can be attributed to each of the principal components.

Table XX displays the results of a principal-component analysis of the cryptocurrencies daily returns: the single-strongest factor only explains 51% of the variation of crypto-currency returns. Moreover, each subsequent factor is providing only slowly declining additional information content, so that XX factors are needed in order to account for 90% of the variation from these 7 cryptocurrencies. The eigenvalues, proportion of variance, and cumulative variance for each component are shown in the Table xxx below:

Principal Component	Eigenvalue	Proportion of Variance	Cumulative Variance
1	3.56566572	Example text	50.86896635
2	0.89534151	Example text	63.64220445
3	0.72106461	Example text	73.92915057
4	0.54715135	Example text	81.73499268
5	0.51625754	Example text	89.10009347
6	0.43768556	Example text	95.3442604
7	0.32634457	Example text	100

?????? ADD (General price trend in cryptocurrencies is not very strong-maybe we compare with S&P500 and this is the initial indicator of the fact that the patterns of the cryptos do not tend to move in parallel)

Component Loadings

In multivariate space, the correlation between the principal component and the original variables (cryptocurrencies) is called the component loadings. Based on loadings, we can tell how much of the variation in a variable is explained by the component.

Loadings = Orthonormal Eigenvectors * $\sqrt{\text{Absolute Eigenvalues}}$

Principal Component	bitcoin	ethereum	litecoin	monero	ripple	nem	dash
1	3.						
2	0.						
3	0.						
4	0.						
5	0.						
6	0.						
7	0.						

The first principal component is required to have the largest possible variance. The second component is computed under the constraint of being orthogonal to the first component and to have the largest possible inertia. The other components are computed likewise.

The first principal component is strongly correlated with six out of seven currencies. The first principal component increases with increasing 1,2,3,4,5,6 scores. **This suggests that these six currencies have high chance to vary together. If one increases, then the remaining ones tend to as well.** This component can be viewed as a measure of, Furthermore, we see that the first principal component correlates **most strongly** with the xxx. While ETH shows strong representation in the second component. This observation is supported by the low correlation of the two crypto-currencies.

3 - XXX**(a) Introduction**