Advanced Scientific Computing: Numerical Methods AM205
Harvard University
Professor Chris Rycroft

December 1, 2017
Crypto project

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 m matrix A has a spectral decomposition of the form,

$$A = U \triangle U^T \tag{1}$$

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 $\triangle$ is a diagonal matrix. The columns of U are the eigenvectors of matrix A and the diagonal elements of $\triangle$ are the eigenvalues. If A is positive-definite, the eigenvalues will all be positive. Multiplying with U, equation 1 can be re-written to,

$$AU = U \triangle U^T U = UA$$

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

$$Au_i = \lambda_i u_i$$

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

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

where U is a n x m matrix with orthonormal columns ($U^T U = I$), while V is a m x m orthonormal matrix ($V_T V = I$), and $\triangle$ is a m 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

$$BB_T = U \triangle V_T V \triangle U_T = U \triangle^2 U_T$$

$$B_T B = V \triangle^2 V_T$$

We can now show that BB_T which is n x n and $B_T B$ which is m 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 V and the squared diagonal elements of $\triangle$, respectively. Denoting one such eigenvector by v and the diagonal element by γ, we have:

$$B_T B v = \gamma^2 v$$

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

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

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

We have now shown that $B_T B$ 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 $BB_T = U\triangle^2 U_T$, we immediately see that the eigenvalues $\beta_\perp$ must all be zero,

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

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.

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

and find that we can write the covariance matrix as

$$C = \frac{1}{n} XX^T = \frac{1}{n} U\triangle^2 U^T$$

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:

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

where $\hat{U}^T U$ is a simple n x 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 $X = U\triangle V^T$ (equation 1), the matrix of factor scores, denoted F is obtained as

$$F = U\triangle$$

(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:

$$F = U\triangle = U\triangle VV^T = XV$$

(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:

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

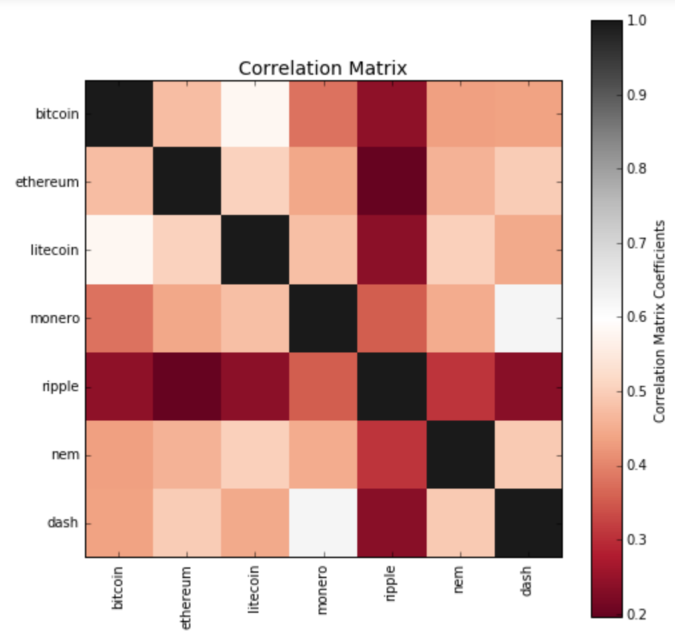
where λ_l is the eigenvalue of the lth component. The value of a contribution is between 0 and 1 and, for a given component, the sum of the contributions of all observations is equal to 1. The larger the value of the contribution, the more the observation contributes to the component.

(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.

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 ith colummn of eigenvector matrix,
λ_i = eigenvalue associated with u_i

Sorting Eigenpairs & Explained Variance

In order to decide which eigenvector(s) can 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 can be dropped. In order to do so, the common approach is to rank the eigenvalues from highest to lowest in order choose the top k eigenvectors. 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.

[report eigen values and vectors result from next section]

Principal Component	Eigenvalue	Eigenvector
1	3.561	array([0.382, 0.387, 0.408, 0.400, 0.244, 0.392, 0.4056])
2	0.894	array([-0.252, -0.279, -0.227, 0.209, 0.873, 0.036, -0.034])
3	0.720	array([0.503, -0.021, 0.333, -0.488, 0.319 , 0.046, -0.542])
4	0.546	array([-0.364, 0.064, -0.102, -0.298, -0.079, 0.864, -0.109])
5	0.516	array([-0.178, 0.868, -0.276, -0.184 , 0.203, -0.229, -0.103])
6	0.437	array([0.532, -0.096, -0.643, -0.318, 0.044, 0.104, 0.424])
7	0.326	array([0.299, 0.071, -0.414, 0.578, -0.167, 0.184, -0.581])

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, proportio of variance, and cumulative variance for each component are shown in the Table xxx below:

Principal Component	Eigenvalue	Proportion of Variance	Cumulative Variance
1	3.561	50.869	50.869
2	0.894	12.773	63.642
3	0.720	10.287	73.929
4	0.546	7.806	81.735
5	0.516	7.365	89.100
6	0.437	6.244	95.344
7	0.326	4.656	100

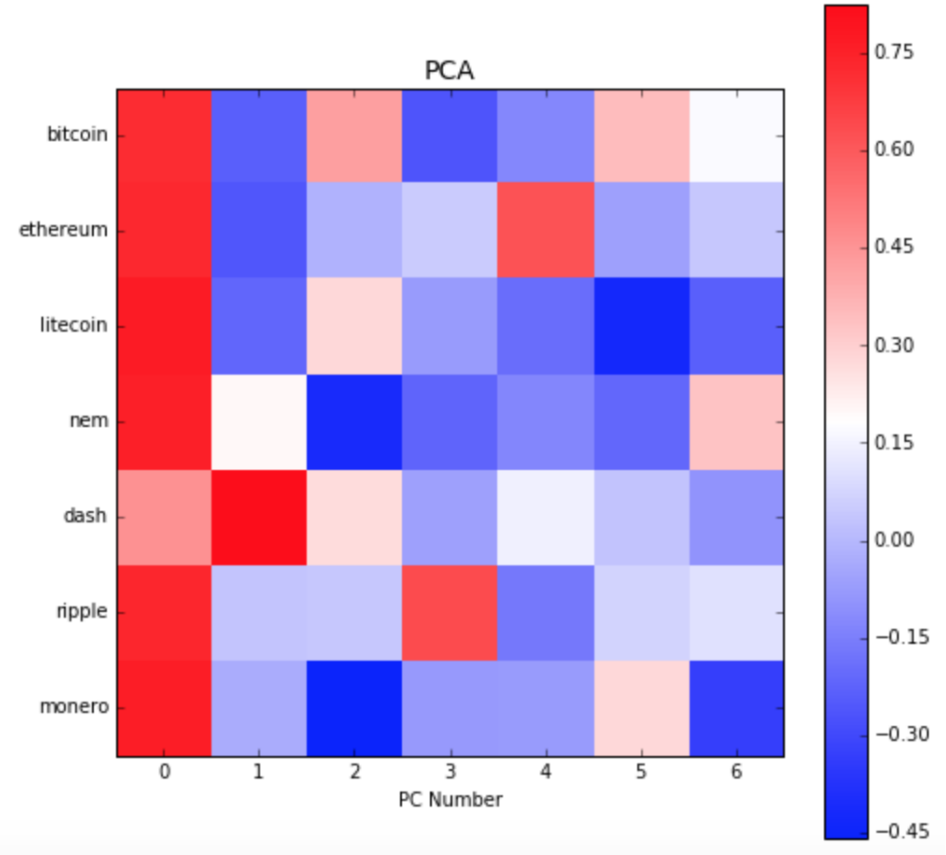
?????? ADD (General price trend in cryptocurrencies is not very strong-maybe we compare with S&P500and this is the initial indicator of the fact that the patterns of the cryptos do not tend to move in parallel)

Component Loadings

In multivariate space, the correlation between the principal component and the original variables (crytocurrencies) 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 $*\sqrt{AbsoluteEigenvalues}$

Principal component loading diagram is shown below:



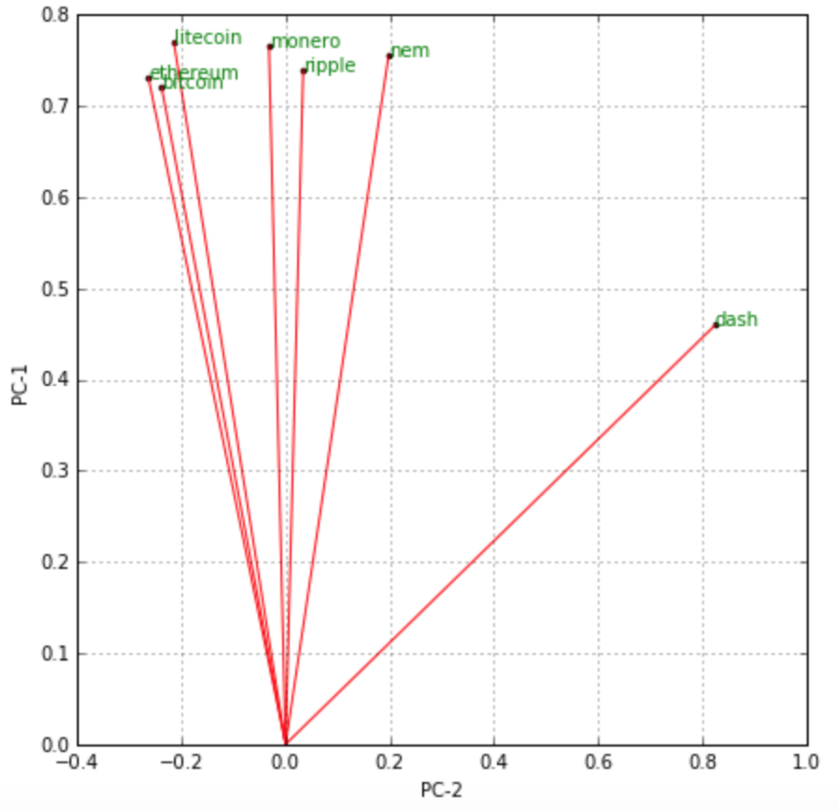
Loading table:

PrincipalComponent	Bitcoin	Ethereum	Litecoin	Nem	Dash	Ripple	Monero
1	0.720	0.731	0.769	0.755	0.459	0.739	0.766
2	0.238	-0.263	-0.215	0.198	0.825	0.034	-0.032
3	0.427	-0.018	0.282	-0.414	0.271	0.039	-0.460
4	-0.269	0.047	-0.075	-0.220	-0.059	0.639	-0.080
5	-0.128	0.623	-0.198	-0.132	0.146	-0.164	-0.074
6	0.352	-0.063	-0.425	-0.210	0.029	0.068	0.280
7	0.171	0.040	-0.236	0.330	-0.095	0.105	-0.332

The first principal component is strongly correlated with six out of seven currencies. The first principal component increases with increasing Bitcoin, Ethereum, Litecoin, Nem, Ripple and Monero scores. This suggests that these six currencies are likely to vary together (positvely correlated). If one increases, then the remaining ones tend to as well. Furthermore, we see that the correlation of first principal component with those six currencies are quite similar.

We also notice that the second principal component correlates most strongly with Dash. In fact, we could state that based on the correlations of 0.825, this principal component is primarily a measure of Dash.

We construct the bi-plot of relative weights of each cryptocurrency in the first two PC components (PC-1 and PC-1) arising from the previous analysis:



This diagram also demonstrate the distinct movement between Dash and the rest of cryptocurrencies. All seven variables have positive values in the PC1 axis, while Bitcoin, Ethereum, Litecoin, Nem, Ripple and Monero are negative in PC2's and Dash is positive. Since all the variables are positive in PC1, those which constrain the system the most are Litecoin and (then) Monero and Nem (in PC1 axis).The PC2 (t has much smaller variance) contrasts Dash from everything else.

In order to further confirm the weak correlation for each pair of currencies (cryptocurrency time-series), we make use of two distinct tools, namely, one-factor linear regression (hence its R^2 metric) and Kendalls rank correlation metric of τ.

3 - XXX

(a) Introduction