Principal Components Analysis

Prepared by Prof. Jason Hsu

For use for UCLA Anderson MFE Program

Introduction

Principal components analysis (PCA) is a statistical technique used to reduce the dimensionality of data. It is an exploratory technique which specifies a linear factor structure between variables, and is especially useful when the data under consideration are correlated. If underlying data are uncorrelated then PCA has little utility.

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

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

Implementing Principal Component Analysis

In general, we are interested in understanding what factors lead to movements in an asset's return. One way of identifying these factors, which presumes no knowledge of any factors, and hence is entirely statistical in nature, is via principal component analysis (PCA). Originally developed as a data reduction process whereby the major sources of variance in a data set can be more parsimoniously represented by a smaller set of statistical factors, PCA can be used to identify the underlying statistical factors which explain comovement in stock returns. Each of these factors will be linear combinations of the observed variables we have in our data, and will be orthogonal to each other. That is, each factor is independent of each other, and variation in one is unrelated to variance in another. Mathematically, we want to transform the covariance matrix of our original data in such a way so as to maximize the variance of each of these orthogonal factors. Considering the covariance matrix, we will want to find a transformation such that we maximize the diagonal elements (largest variances) and reduce the off-diagonal covariances to zero (make factors independent). We can accomplish this by using an eigenvalue decomposition of the covariance matrix. Mathematically, we have:

X = matrix of returns for each asset;
$\underline{X}$ = matrix of demeaned returns for each asset;
Cx = covariance matrix of X;
P = transformation matrix;
Y = transformed data matrix;
Cy = covariance matrix of transformed data.

We want to transform our original data such that *Cy* is diagonal and *P* is orthonormal (that is, the vectors are mutually orthogonal or are perpendicular to each other). We start with

$$Y = \underline{X}P$$

So,

$$Cy = \left(\frac{1}{n}\right)Y'Y = \left(\frac{1}{n}\right)(\underline{X}P)'(\underline{X}P)$$

As a quick note, the results above also follow if you use the raw data matrix *X* instead of the de-trended matrix *X̲*, but the interpretation of *Cx* as a covariance matrix is facilitated by using the de-trended values.

Since a symmetric matrix, of which the covariance matrix is one, can be represented as the product of a matrix of eigenvectors and another with the eigenvalues on the diagonal, we let *E* be a matrix of eigenvectors, and *D* be a matrix with the eigenvalues on the diagonal and zeros on the off-diagonal, we can hence represent our matrix *Cx* as EDE^{-1}*.* Now, we make a choice here, that since we wanted *P* to be an orthonormal transformation, and matrices of eigenvectors are orthonormal, we choose *P = E*.

Thus, we have

$$Cy = P'PDP^{-1}P = (P'P)D(P^{-1}P) = D$$

where we used the fact that $P^{-1} = P'$ for orthogonal matrices. Note that this propitious choice for *P* allows us to diagonalize *Cy*. So, now we know how to transform our original data *X*, into data that diagonalizes the covariance matrix of the transformed data. Recall, this means that all of the variance is due to independent or orthogonal factors, which is what we want, and the basis vectors, or the new dimensions along which we maximize the variance or the orthogonal eigenvectors, are the columns of the *P* matrix. Now, the diagonal matrix *D*, has eigenvalues on the diagonal, and the relative size of the eigenvalues is an expression of the percent of total variance that that particular principal component explains. So, if we are looking to identify the 5, or 10 most important components, then we can simply choose the eigenvectors (which act as a mapping between our original data and the transformed data) associated with the 5 or 10 largest eigenvalues (numbers on the diagonals of the *D* matrix).

Now, in the case of the expected returns of an asset, we generally assume that the returns have the following structure:

$$R = bf + \varepsilon,$$

where *f* is the realization of the factor, *b* is the asset's sensitivity to the factor, and ε is idiosyncratic noise. Following Clarke, de Silva, and Thorley (2006), we define $\Omega = R'R$ as the sample covariance matrix. Note that we did not de-trend the returns or scale by *T*. However, if the number of assets is large relative to the number of observations, then we need to make use of the results of Connor and Korajczyk (1986, 1988) so as to ensure that the matrix is invertible. This is a point worth remembering. Since a 1000 by 1000 covariance matrix requires 1000*(1000-1)/2 unique covariances from a generally much shorter time series. This often leads to covariance matrices that can't be inverted, which prevents us from solving all sorts of important questions. To overcome this, we examine the *T* by *T* returns cross-product matrix $\Omega = RR'$ (note that we reversed the order of the two matrices). We then run principal components on the cross-product matrix Ω and take the top *K* components. The residuals are obtained by rearranging the previous equation to get:

$$e = R - bf\,.$$

We need the residuals here since the covariance matrix under PCA uses the components to give structure to the off-diagonal elements, but leaves the diagonal variances untouched. Thus:

$$\Omega = Var[R] = Var[bf + \varepsilon] = b'(ff')b + diag(ee') = bb' + diag(ee')$$

where *ff' = I* (be aware that the transposes may need to be reversed based on how your return series is set up). See MATLAB or SAS code for implementation details.

Determining Number of Principle Components

Given a number of assets, how do we choose the "right" number of principal components to include? There are no real answers, though there are some very good guidelines. First, we mentioned that the size of the eigenvalues tell us the relative importance of the factor that is associated with it, as well as the percentage of variance explained by that component. To calculate the amount that it explains, just sum up the value of all of the eigenvalues, and then divide each eigenvalue by the sum. Now, if you would like your factors to explain, say, 75% of the variance in your data set, then you choose as many factors as you need to so that the cumulate percentage explained is greater than 75% (For details of this method, see Example 1).

Now, one thing to be careful of here is the scale, or relative magnitude of the variances of each of the assets you are looking at. If one asset, or asset class, is far more volatile than another, then PCA will focus in on the high volatility assets, and the factors will seek to explain those returns, potentially to the relegation of your less volatile assets to minor importance (think about treasury bonds, commodities futures, and blue chip stocks). One way to address this is to standardize the returns by dividing by the standard deviation of the returns, and then use PCA with the standardized returns. This is essentially the same as doing PCA on the correlation matrix rather than the covariance matrix. With PCA on the correlation matrix, the eigenvalues on the diagonal will now sum up to the number of variables. So, if there are 10 variables, then the sum of the eigenvalues will be 10, so we would expect that any major factor would at least be able to generate its share of variance. So, in this case, it is common to choose all components with eigenvalues greater than one.

Example 1 (Determine the number of principal components by cumulate percentage explained)

PCA will create as many factors as variables under consideration. For example, if one performed PCA on the historical returns of 100 stocks, then 100 principal components would be created by the statistical software package. Using all 100 factors is overkill if the variables under consideration are correlated. Ideally, only a few of the principal components are retained and explain a majority of the variation in the underlying data. The goal is to reduce the dimensionality of returns into a simpler factor model.

Statistical software packages will typically compute the sample covariance or correlation matrix for the data provided. Most packages use the correlation matrix as a default. Be sure to check what the software is assuming. The resulting PCA will be affected by the choice of using the sample covariance matrix or the sample correlation matrix- they are not equivalent even though the same data are used to compute them. Using the correlation matrix considers each variable on an "equal-footing," or as if each variable has unit variance. If the sample covariance matrix is used, then variables are considered with unequal variances. This can be hazardous when there is large dispersion in the estimates of variance in the underlying data - the variables with the highest variance will dominate the first principal component regardless of the correlation to the other data. This is typically an issue when data measured are in different units.

The eignevalues of the sample covariance matrix are calculated by the software. The resultant eigenvalues are then sorted by descending value. The largest eigenvalue is equal to the variance of the first principal component. The second largest eigenvalue is the variance of the second principal component, and so on for all 100 diagonal entries in the sample covariance matrix. In this example, the first 13 eigenvalues are shown in the Table 1 below:

Table 1 Principal Components

Principal Component	Eigenvalue	Proportion of Variance	Cumulative Variance
1	75	26.95%	26.95%
2	43	15.45%	42.40%
3	30	10.78%	53.18%
4	21	7.55%	60.73%
5	19	6.83%	67.55%
6	18	6.47%	74.02%
7	17	6.11%	80.13%
8	11	3.95%	84.08%
9	10	3.59%	87.67%
10	10	3.41%	91.09%
11	9	3.16%	94.25%
12	5	1.80%	96.05%
13	4	1.58%	97.63%

The first 10 principal components account for more than 90% of the cumulative variation in the returns of the 100 stocks under consideration. This is sufficient for the factor model being built, and these first 10 principal components will be selected as the factor model. The rest of the 90 principal components account for the remaining 8.91% of variation and can be discarded. There is no set rule for how many components should be used, however; typically the additional variation explained by each component is diminishing.

In this case, the PCA reduced the dimensionality of the 100 stocks under consideration to 10 while accounting for more than 90% of the returns variation. Assuming the PCA was well specified and there are no major violations of assumptions, the resultant factor model can be used as to forecast the covariance matrix of stock returns.

Example 2 (Determine the number of principal components by eigenvalues: scree plot)

A useful procedure to determine the number of principal components needed in an application is to examine the *scree plot,* which is the time plot of the eigenvalues ordered from the largest to the smallest. The scree plot of the eigenvalues to visually demonstrate the proportion of total variance each principal component is accounting for, and that you can throw away the lower principal components without losing much explanatory power. Figure 1a shows the scree plot for five stock returns. By looking at for an elbow in the scree plot, indicating that the remaining eigenvalues are relatively small and all about the same size, one can determine the appropriate number of components. For both plots in Figure 1, two components appear to be appropriate.

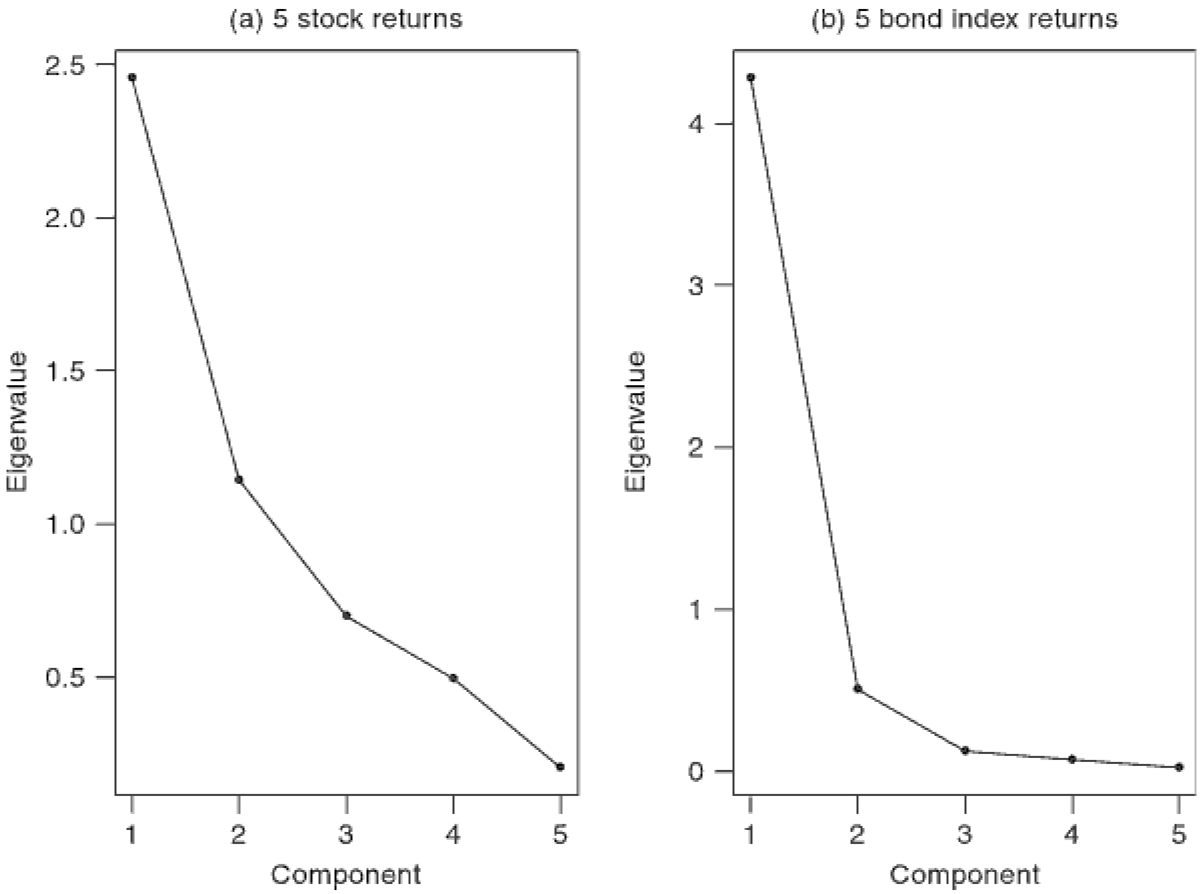

Figure 1 Scree plots for two 5-dimensional asset returns: (a) series of 5 stock returns and (b) bond index returns.

Expressing Covariance Matrix Using Identified Principal Components

Recall the original transformation we used

$$Y = XP \quad or \quad \hat{X} = YP^{-1} \text{ (matrix with means in each column)}$$

So, we are looking at approximating the original data matrix *X* with the transformed values and the eigenvectors, however, we now are only going to use *K* of the *N*, *K* < *N*, of the components. The matrix multiplication works out as follows:

$$\hat{X}_{T\times N} = Y_{T\times K} P^{-1}_{K\times N}$$

$\hat{X}$ has *T* time periods with *N* variables, we are only going to use the largest *K* principal components, so we use the *K* columns of those principal component scores from the *Y* matrix, and multiply that by the inverse of *P*. Recall, *P* is ($N \times N$), but since we only want the *K* largest eigenvectors, we only want the top *K* rows of P^{-1} (which is equal to *P'*). Now we have an approximation based on the *K* principal components, and we can find the covariance of the $\hat{X}$ matrix.

Now, in the specific case for our asset returns based on the factor model under the APCA approach, we found that

$$\Omega = b'(ff')b + diag(e'e)$$

If you have few enough assets relative to the number of time periods than you can use the eigenvectors directly

$$\Omega = (YP^{-1})'(YP^{-1}) + diag(e'e)$$

This is the sample covariance matrix which is based on the top principal components, to which we add the diagonal matrix of residual variances. If you check the sample covariance matrix (which is not demeaned as discussed above), you'll find that the diagonal elements are the same as those from the principal components based covariance matrix Ω above.

Creating Factor Mimicking Portfolios

Here we want to create a portfolio whose returns are highly correlated with those of any given principal component. One of the standard ways of doing this is to create a long-short portfolio that is long the assets with the highest weighting or loading on a given component (or more generally, any characteristic) and short those with the lowest loadings. If we consider the first principal component, then the weightings for each stock are found in the first column of our *P* matrix of eigenvectors. There is no clear cut answer for what percent to go long versus short, with answers varying between the top half minus the bottom half, to top tenth minus bottom tenth.

In the series of papers (Connor and Korajczyk (1986, 1988, 1993)) on using PCA to create statistical factors for APT models, Connor and Korajczyk generally use the realizations of the principal component as the realizations for that factor. Hence, the eigenvectors give you the weight for each stock in the factor mimicking portfolio. In this approach, the columns of the matrix *Y* will be the returns for each factor. This will be the preferred method for creating factor returns with PCA. In the MATLAB code, you will see that the correlation between the long-short portfolio approach and the first principal component is between .58 and .85 depending on the construction of the LS portfolio and the time period. It is also useful to note that in most asset pricing situations, the factors will be estimated over 5 year rolling horizons rather than the 18 year period used in this exercise.

Remarks

The following observations/suggestions which relate to statistical robustness checks are summarized from real financial industrial practice.

- PCA is sensitive to whether you are using the sample covariance matrix or the sample correlation matrix. If one uses the sample covariance matrix, if some variables have high variance relative to the other variables then they can dominate the first principal component *regardless* of the covariance structure. On the other hand, the sample correlation matrix considers each variable on an "equal-footing" (or all variables have equal variance) and is not subject to this potential bias, so there are benefits to both ways. Nevertheless, you should figure out what your statistical software package is using, because it will be assuming either the sample covariance or correlation matrix when building the principal components and this will meaningfully affect the outcome. There are advantages to both ways. The consideration for portfolio construction is that if some assets returns time series have extremely high volatility relative to the other assets then they can dominate the first component regardless of how correlated they are to the other assets.

- PCA can reduce the dimensionality of the data; however, if the original variables are already uncorrelated then you don't stand to gain much. This is probably not a concern for building a portfolio of many stocks as they likely commove to some extent, but should be considered when building portfolios with only a few different assets. One can quickly do a hypothesis test that the covariance matrix is diagonal to see if the data are indeed independent.

- Also, be careful of outliers in the input data. One of the assumptions of PCA is that there are no outliers in the underlying data - if there are outliers they can "break" the resultant principal component factors by making them non-normal (extreme skewness in the factors, for example). Each time you do a PCA, check that the resulting principal component scores are normally distributed - otherwise your factor structure will not be useful. This is also useful for detecting outliers in your data - it can be iteratively applied.

- Make sure all the units of the variables you are doing principal components analysis on are the same.

- Lastly, one of the benefits of PCA is that matrix singularity is not an issue because matrices are never inverted.

References

Campbell, J., Lo, A. and MacKinlay, C. (1997). The Econometrics of Financial Markets. Princeton University Press, Princeton, New Jersey.

Chan, L., Karceski, J. and Lakonishok, J. (1998). The Risk and Return from Factors. Journal *of Financial and Quantitative Analysis*. 33, 159-188.

Chincarini. L. and Kim, D. (2006). Quantitative Equity Portfolio Management. McGraw-Hill, New York.

Clarke, R., de Silva, H. and Thorley, S. (2006). Minimum-Variance Portfolios in the U.S. Equity Market. *The Journal of Portfolio Management*. 33(1), 10-24.

Connor, G. and Korajczyk, R. (1986). Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis. *Journal of Financial Economics*. 15(3), 373-394.

Connor, G. and Korajczyk, R. (1988). Risk and Return in an Equilibrium APT: Application of a New Test Methodology. *Journal of Financial Economics*. 21(2), 255-289.

Connor, G. and Korajczyk, R. (1993). A Test for the Number of Factors in an Approximate Factor Model. *Journal of Finance*. 48(4), 1263-91.

Shlens, J. (2009). A Tutorial on Principal Component Analysis. *http://www.snl.salk.edu/~shlens/pca.pdf*

Tsay, R. (2005). Analysis of Financial Time Series. 2nd edn. Wiley, Hoboken, New Jersey.