Application of PCA and PVCA to a portfolio of 20 stocks in South Africa

Wouter Bezuidenhout[a]

[a]*Stellenbosch Univeristy, South Africa*

Abstract

This paper investigates two dimension reduction topics: principal component analysis and principal variance component analysis. The techniques are applied to a portfolio of 20 blue chip South African stocks for the period from 2013/01/01 - 2021/10/29. The importance of the paper is methodological, specifically the application of PVCA to the South African Market.

Keywords: Principal Component Analysis, Principal Variance Component Analysis

JEL classification C38, C58

1. Introduction

Elementary analysis of financial returns usually assume time-invariant or constant covariances (volatility), known as homoskedasticity. It is well-known that financial returns posses conditional heteroskedastic, or time-varying conditional covariances, known as volatility [tsay2]. The pursuit of modeling conditional heteroskedasticity is important, and resulted in a Nobel Prize for Robert Engle. My paper extends this idea to argue that is more useful to known if volatility is common between different time-series. The challenge with multivariate volatility modeling is the growth in dimensions. For K variables, the covariance matrix has K(K+1)/2 processes to estimate. Therefore, I implement two dimension reduction techniques to engage a more manageable set of estimates. I use two methods to explore this: a principal component analysis (PCA) and a principal variance component analysis (PVCA).

PCA finds structure in the covariance matrix to locate low-dimensional sub-spaces that contain most of the variation of the data [ruppert]. PCA is not feature selection, as each PCA is just a linear combination of all original variables. PCA is especially useful for highly correlated variables like financial returns, as PCAs are uncorrelated from one another. An example of how this is useful, say an analyst would be interested in the behavior of an index with 50 stocks. After implementing PCA, the analyst could focus on prediction of a dozen PCAs responsible for more than 90 percent of the

Email address: 20088418@sun.ac.za (Wouter Bezuidenhout)

variation in the data, instead of focusing on all 50 stocks simultaneously (Ruppert & Matteson, 2011). On the other hand, PVCA is a generalization of PCA to detect common volatility factors in returns. The aim of PVCA is to find a small number of common volatility components in order to find a linear combination of the series with no conditional heteroscedasticity [hu].

The computation required for PVCA is strenuous, and to keep a positive semi-definite matrix is difficult as dimensions grow. In Hu & Tsay (2014), the author uses 7 currencies, whereas in Engle & Susmel (1993), the author uses 18 indexes. Therefore, for my application, I have decided to apply these techniques to a portfolio of 20 blue chip stocks listed on the JSE. My aim is the showcasing of the two methodologies applied to equity returns. The paper is structured as follows. Section 2 discusses the methodologies for PCA and PVCA in detail so that the paper is as self-contained as possible. Section 3 discusses the data's properties and transformations implemented. Section 4 discusses the results of the two methods. Section 5 concludes.

2. Methodologies

Data is required to look as follows: one needs a sample $\mathbf{Y}_i = (Y_{i,1}, ..., Y_{i,d}), i = 1, ..., n$ of d-dimensional random vectors with mean vector μ and covariance matrix $\mathbf{\Sigma}$. PCA is focused on extracting structure from the covariance matrix. PCA produces zero contemporaneous correlations, meaning PCA overlooks the dynamic dependence between the volatility processes (James, Witten, Hastie & Tibshirani, 2013). PVCA focuses on the dynamic dependence of volatility. The motivation with PVCA is with a small number of common volatility components, one can find a linear combinations of the return series that contains no conditional heteroscedasticity [hu]. In PCA, one performs a spectral decomposition of the covariance matrix. Following Hu & Tsay (2014), the authors extend this to propose a sample estimate of a cumulative generalized kurtosis matrix to summarize the dynamic volatility dependence of the multivariate time series. Spectral analysis of this generalized matrix is then used to define PVCs. In order to determine that no conditional heteroscedasticity is present in the PVC process, the authors conduct a generalized Ling–Li test statistic. It is worth noting that Engle & Susmel (1993) conducted a study with a similar aim, but used noticeably different methods. Engle & Susmel (1993) conducted a pairwise procedure to test for no conditional heteroscedasticity after modeling using GARCH and M-GARCH models.

3. Data

The return series that I have chosen for my application is a portfolio of 20 blue chip, large cap equities listed on the Johannesburg Stock Exchange (JSE). The sample period is from 2013/01/01 to 2021/10/29. The return series has been logged. The following 20 stocks are in the portfolio. I initially

tried to conduct this with the ALSI Top 40, but there are numerous nuances with such an application that require further thinking in order to achieve.

Short name (ticker)	Sector
BHP Group (BHP)	Resources
Anglo American (AGL)	Resources
Sasol (SOL)	Resources
Anglogold Ashanti (ANG)	Resources
Richemont (CFR)	Industrials
MTN Group (MTN)	Industrials
Shoprite (SHP)	Industrials
Mondi (MNP)	Industrials
Aspen Pharmaceuticals (APN)	Industrials
Naspers (NPN)	Industrials
Vodacom (VOD)	Industrials
Standard Bank (SBK)	Financials
Firstrand (FSR)	Financials
ABSA (ABG)	Financials
Growthpoint (GRT)	Financials
Nedbank (NED)	Financials
Investec Ltd. (INL)	Financials
Investec Plc. (INP)	Financials
Remgro (REM)	Financials
Sanlam (SLM)	Financials

Table 3.1: Portfolio of 20 stocks

In order to get an idea of the series, I plot them below. Figure 3.1 shows no noticeable missing periods of data, furthermore, one can already see periods of high variance in some stock returns. There is a period in 2020 where Sasol (SOL) shows especially high variance. In the next section, I run the PCA and PVCA, and discuss their results.

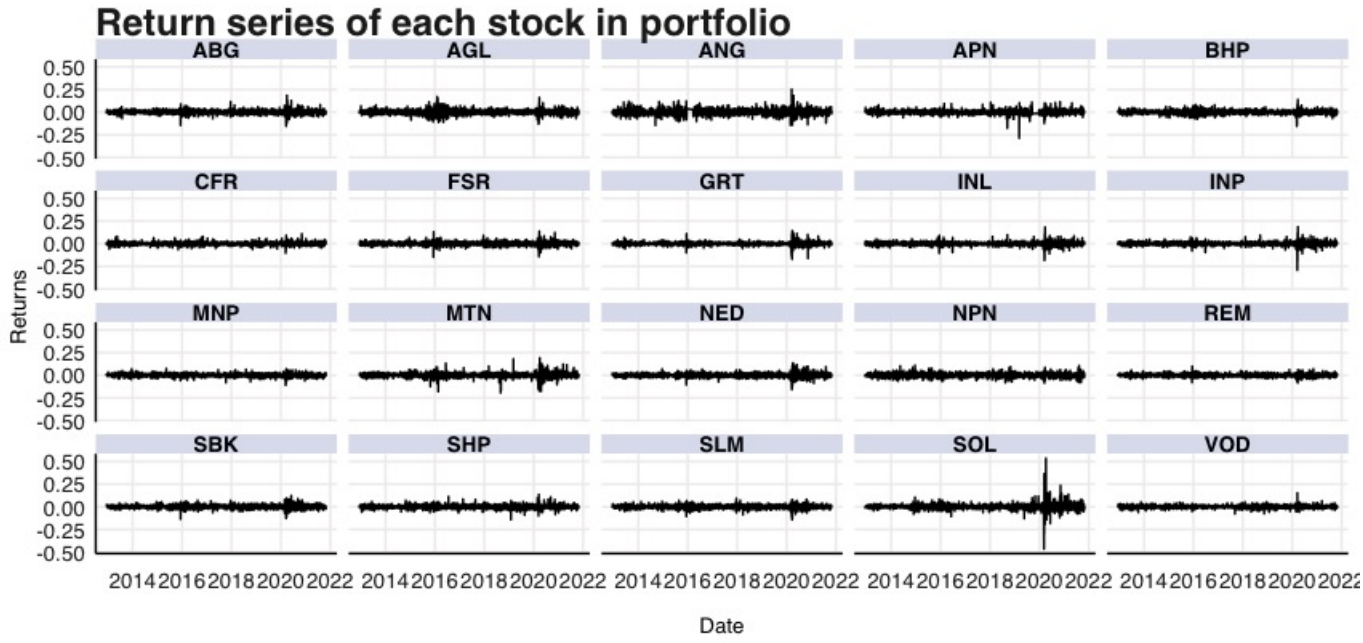

Figure 3.1: Stock returns of the portoflio

4. Results

4.1. Principal Component Analysis

It is recommended to standardize the variables when one conducts PCA. If variables are on different scales, then variables with greater variance in magnitude will dominate the analysis and therefore bias the results (James, Witten, Hastie & Tibshirani, 2013). In my case, with logged returns, it should not be necessary, but I do so for completeness. Furthermore, I check for missing returns and find a small amount in Anglogold Ashanti (ANG) and Aspen Pharmaceuticals (APN). I impute these by drawing from their distribution. I calculate the PCAs using the prcomp function in R from the stats package. This function decomposes the d-dimensional into d contemporaneously uncorrelated PCAs, and ranks them based on the amount of variability explained by each. Figure 4.1 displays the scree plot of the PCAs. A scree plot shows the amount of variability explained by each PCA. There are 20 PCAs, but I have chosen to display the first 12. The first PCA explains 36,5 percent of the variability of all 20 stock returns. To complement this, Table 4.1 shows the cumulative proportions of variance explained by each PCA. With 11 PCAs, one can explain more than 90 percent of the variance in the portfolio of 20 stock returns. Figure 4.2 takes a closer look at PCA 1 to understand which shares underlie its importance. The figure shows that Sasol (SOL) represents most of the contribution, followed by Nedbank (NED), Absa (ABG), Firstrand (FSR), MTN and Standard Bank (SBK).

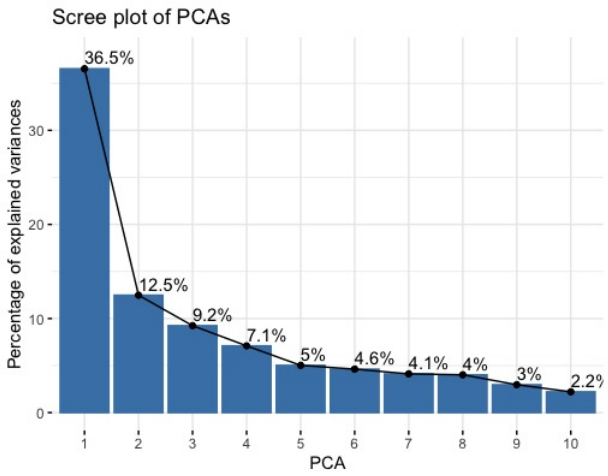

Figure 4.1: Scree plot fo PCAs

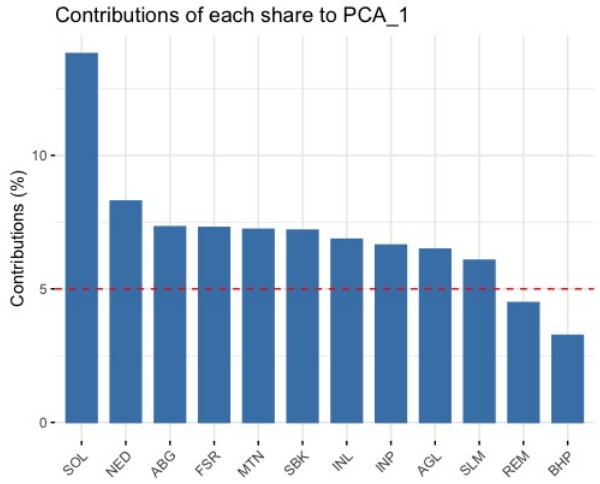

Figure 4.2: Individual contributions to PCA 1

Category	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Eigenvalues	43,3	30,3	21,6	13,4	10,3	8,4	7,2	6,4	3,5	2,4	1,8	1,6
Prop of Variance (%)	36,5	12,5	9,2	7,1	5	4,6	4,1	4	3	2,2	2,1	1,6
Cum Prop (%)	36,5	49	58,3	65,4	70,4	75	79,1	83,2	86,1	88,3	90,4	92,1

Table 4.1: Importance of PCAs

For a deeper look at the PCAs, it is worth investigating the eigenvectors. Appendix Table 5.1 shows the eigenvectors for each PCA. Following the approach of Ruppert & Matteson (2011), the first eigenvector has only negative values, meaning for an increase in PCA 1 all returns should decrease. Eigenvector 2 has 7 negative values and 13 positive values, where the negative values are all four resources stocks, as well as the following industrials: Naspers, Mondi and Richemont. Variation along this eigenvector has resource stocks and three industrials moving in the opposite direction to other returns. It is not uncommon to see resource move opposite to the market, and therefore this PCA accounts for 12,5% of variation. For eigenvector three, all values are negative except for Anglo American, BHP, Richemont, Investec Ltd, Investec Plc, Mondi and Sasol. This PCA accounts for 9,2% of the variation. Both PCA 2 and 3 should not be over-confidently interpreted, but rather modeled quantitatively. In conclusion to the PCAs, logically, it seems that sectors have common variation and that PCA 1 represents an overall market factor. The PCAs derived are now able to be used in prediction modeling by an analyst who will now have less dimensions to work with opposed to having to work with all 20 stocks. There is no golden rule about how many PCAs to use, but I argue that 90 percent of variation explained by 11 PCAs is suitable. For interest sake, if I modeled stocks only from the same sector (say Resources), one would observe the 90 percent threshold reached in a handful of PCAs opposed to a dozen. In the next subsection, I turn to the results of the PVCA.

4.2. Principal Variance Component Analysis

Hu & Tsay (2014) proposed the idea of PVCA, the generalization of PCA methods to focus more directly on modeling common multivariate volatility. The method requires the specification of a Vector-autoregression (VAR) model to account for serial correlation. According to the Akaike Information Criterion, my data requires only two lags where as in Hu & Tsay (2014), the authors uses 5 lags. After running Portmanteau tests to detect serial correlation, I am not completely satisfied that serial correlation is not present, however, no additional amount of lags remove this. Additionally, the data has been logged and differenced (working with returns), and therefore I proceed.

After running the model and producing 20 PVCAs, I conduct univariate ARCH tests on all 20 PVCAs. I find that the 19th PVCA has no conditional heteroscedasticity. In the Hu & Tsay (2014) paper, this result was found with their seventh of seven currencies. This implies that there certainly exists common volatility factors in my portfolio of 20 stocks (Hu & Tsay, 2014). Figure 4.3 presents the time series of selected PVCAs. Although, the F-Tests confirm no conditional heteroscedasticity, the variances do not look much different with the eye. Additional ACF plots are not informative either.

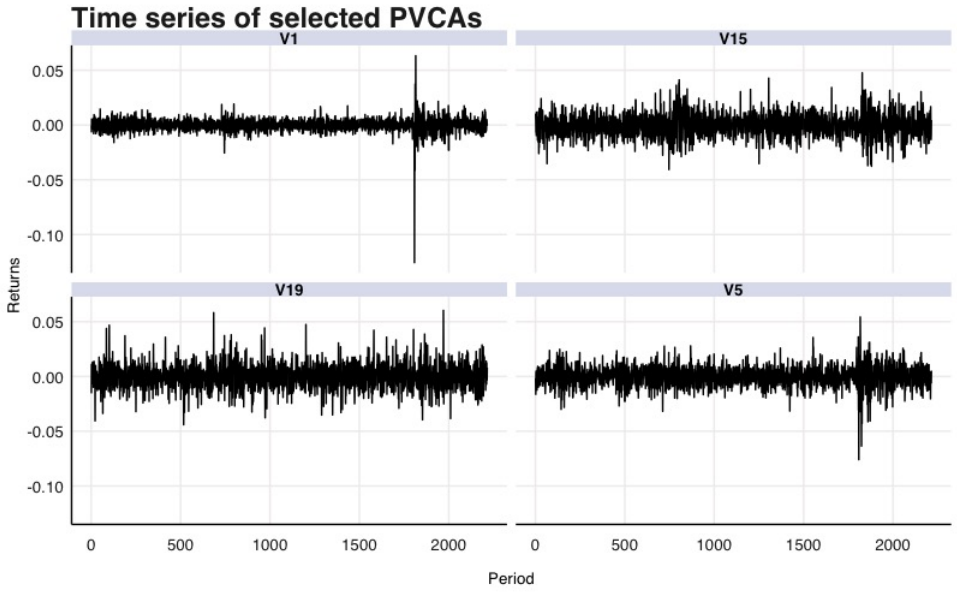

Figure 4.3: Time Series of selected PVCAs

Figure 4.4 displays the scree plot of the PVCAs. In this case, the scree plot shows the amount of volatility explained by each PVCA. The first PVCA explains 27,7 percent, the second 19,4 percent and the third 13,8 percent, of the volatility of all 20 stock returns. Cumulatively, the first three PVCAs account for 61 percent of volatility. Table 4.1 shows the cumulative proportions of volatility explained by each PVCA more specifically. With 12 PVCAs, one can explain 96,2 percent of the volatility in the portfolio of 20 stock returns. Figure 4.5 takes a closer look at PVCA 1 to understand which shares

underlie its importance. The figure shows that Anglo American (ANG) and Shoprite (SHP) are most important in the first PVCA. This is a surprising result. In order to understand the PVCAs better, I turn to looking at their eigenvectors.

Table 4.2 in the appendix contains the eigenvectors for the individual stocks for the first 12 PVCAs (explaining 96% of volatility). Note that I have rounded off for the values for presentability. If the F-test for conditional heteroscedasticity not present at PVCA 19, then my interpretation is that the first 18 PVCAs account for all the common and time-varying volatility, whereafter the 19th PVCA shows time-invariant covariance. Overall, the PVCA model is a good attempt at a unique method to modeling common volatility. Additionally, a more interesting application might highlight the usefulness of the method more. Hu & Tsay (2014) had 7 variables in their sample, whereas I have 20, and Engle & Susmel (1993) had 18 - although Engle & Susmel (1993) used vastly different methods. In my calculations, as one added more variables, the process immediately became more complex computationally and mathematically.

PVCA	1	2	3	4	5	6	7	8	9	10	11	12
Eigenvalues	43,3	30,3	21,6	13,4	10,3	8,4	7,2	6,4	3,5	2,4	1,8	1,6
Prop of Volatility (%)	27,7	19,4	13,8	8,5	6,6	5,4	4,6	4,1	2,3	1,6	1,2	1
Cum Prop (%)	27,7	47,1	60,9	69,4	76	81,4	86	90,1	92,4	94	95,2	96,2

Table 4.2: PVCAs

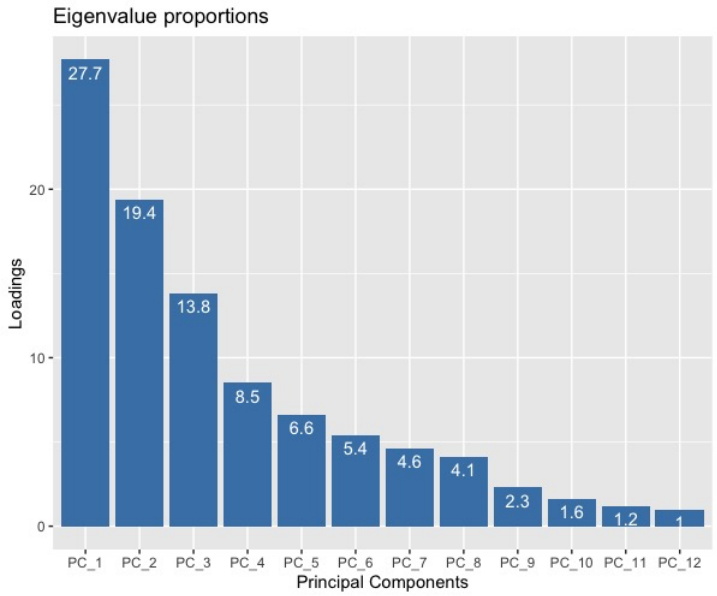

Figure 4.4: PVCAs proportions explained

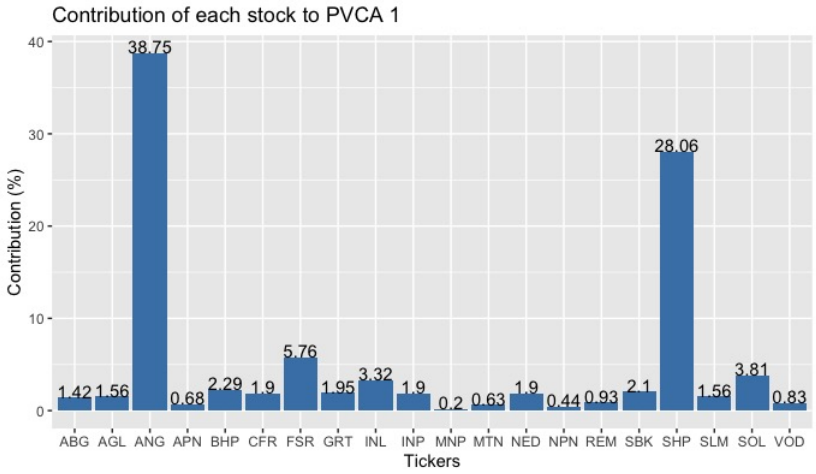

Figure 4.5: Individual contributions to PVCA 1

5. Conclusion

Modelling common volatility has been the pursuit of numerous authors in the past. Options traders are often concerned with modelling common implied volatility, too. However, the challenge with multivariate volatility modeling is the growth in dimensions of parameters required to be estimated. For K variables, the covariance matrix has K(K+1)/2 processes to estimate. Dimension reduction techniques are therefore crucial if one is to build a model that is accurate and efficient. In this paper, I have implemented two dimension reduction techniques, a principal component analysis and a principal variance component analysis. PCA is concerned with investigating the structure of covariance matrix in order to represent the high-dimensional series with a low-dimensional uncorrelated PCAs. PVCAs generalize the process of PCAs by estimating a cumulative generalized kurtosis matrix to summarize the dynamic volatility dependence of the multivariate time series. The motivation of PVCAs is to find a PVCA combination that is a linear combinations of the return series that contains no conditional heteroscedasticity.

My PCA model was successful and it logically it seemed that sectors were moving together within the PCA. With 12 PCAs, my model could account for more than 92 percent of the variation of the 20-stock portfolio. The noticeable contributions to explaining variation within PCA 1 were Sasol, Nedbank, Absa, Firstrand, MTN and Standard Bank. This PCA is ready to be used in further analysis. The PVCA model was a good attempt at a unique method to modelling common volatility. The first 12 PVCAs account for 96,2 percent of the volatility. The F-tests that I conducted to test for the presence of conditional heteroscedasticity showed that PVCA 19 had no conditional heteroscedasticity. This means that the common volatility was accounted for in the previous PVCAs. Surprisingly, in the first PVCA, Anglo American and Shoprite were the most important contributors. In conclusion, more work

needs to be done to elaborate on the R-package that models PVCA and a more suitable application needs to be conducted to really test the usefulness of the method.

References

10 Engle, R.F. & Susmel, R. 1993. Common volatility in international equity markets. *Journal of Business & Economic Statistics*. 11(2):167–176.

Hu, Y.-P. & Tsay, R.S. 2014. Principal volatility component analysis. *Journal of Business & Economic Statistics*. 32(2):153–164.

James, G., Witten, D., Hastie, T. & Tibshirani, R. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

Katzke, N.F. 2017. *Texevier: Package to create elsevier templates for rmarkdown*. Stellenbosch, South Africa: Bureau for Economic Research.

Ruppert, D. & Matteson, D.S. 2011. *Statistics and data analysis for financial engineering*. Vol. 13. Springer.

Tsay, R.S. 2013. *Multivariate time series analysis: With r and financial applications*. John Wiley & Sons.

Tsay, R.S. 2014. *An introduction to analysis of financial data with r*. John Wiley & Sons.

Appendix

Appendix A

Eigenvectors of PCA Process												
Ticker	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
BHP	-0.05	-0.39	0.04	-0.43	-0.04	0.39	0.14	-0.04	-0.48	0.12	0.41	-0.21
AGL	0.03	-0.12	0.03	-0.08	0.06	-0.05	0.00	0.11	-0.06	0.05	0.13	0.36
CFR	0.04	0.06	0.07	0.05	-0.16	-0.03	-0.02	-0.08	0.00	-0.10	0.14	-0.05
MTN	-0.01	0.06	-0.09	-0.08	-0.04	0.01	0.06	0.12	-0.10	-0.15	-0.27	0.10
SOL	-0.08	0.27	0.05	-0.03	-0.08	0.02	0.00	-0.24	-0.13	-0.27	-0.38	-0.49
NPN	-0.01	-0.02	-0.04	-0.18	-0.09	0.01	0.06	-0.10	-0.03	0.02	-0.16	-0.16
SBK	0.04	0.02	-0.06	-0.48	0.12	-0.41	-0.16	-0.39	-0.06	-0.05	-0.07	0.19
FSR	0.12	0.38	-0.53	-0.07	-0.02	0.24	0.00	0.16	-0.02	-0.15	0.18	-0.14
SHP	-0.57	0.44	-0.34	-0.34	-0.67	-0.06	-0.72	0.68	-0.19	0.44	0.17	0.24
ANG	0.79	-0.30	0.63	0.30	0.58	-0.08	0.59	-0.37	0.16	-0.29	-0.21	-0.28
SLM	-0.03	0.08	-0.14	0.10	0.01	-0.03	0.03	-0.01	-0.18	0.46	0.29	0.17
REM	0.02	-0.03	-0.05	0.24	0.20	0.19	-0.09	-0.15	-0.28	0.03	-0.06	-0.02
ABG	-0.03	-0.21	0.17	0.40	-0.23	0.58	-0.02	0.12	0.19	-0.24	-0.10	0.05
APN	0.01	0.06	-0.04	0.04	0.06	-0.03	0.00	-0.05	0.07	-0.17	0.31	0.02
VOD	-0.02	-0.20	0.25	0.06	0.03	0.09	-0.08	0.12	-0.08	0.01	-0.41	-0.11
NED	0.04	0.23	-0.13	0.12	0.07	-0.25	0.05	-0.07	0.48	-0.14	-0.09	0.47
GRT	-0.04	0.09	0.11	0.21	-0.01	-0.12	-0.01	-0.02	-0.21	-0.10	0.22	0.01
INP	-0.04	0.17	-0.08	0.11	-0.13	0.09	0.06	-0.23	0.45	0.49	-0.14	-0.21
MNP	0.00	-0.33	-0.18	0.06	-0.05	-0.27	0.00	0.03	0.05	0.01	0.01	-0.04
INL	-0.07	0.16	0.07	0.12	-0.17	-0.25	0.23	0.12	-0.20	0.03	-0.02	0.23

Figure 5.1: Eigenvectors of PCAs

Eigenvectors of PVCA Process												
Ticker	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
BHP	-0.05	-0.39	0.04	-0.43	-0.04	0.39	0.14	-0.04	-0.48	0.12	0.41	-0.21
AGL	0.03	-0.12	0.03	-0.08	0.06	-0.05	0.00	0.11	-0.06	0.05	0.13	0.36
CFR	0.04	0.06	0.07	0.05	-0.16	-0.03	-0.02	-0.08	0.00	-0.10	0.14	-0.05
MTN	-0.01	0.06	-0.09	-0.08	-0.04	0.01	0.06	0.12	-0.10	-0.15	-0.27	0.10
SOL	-0.08	0.27	0.05	-0.03	-0.08	0.02	0.00	-0.24	-0.13	-0.27	-0.38	-0.49
NPN	-0.01	-0.02	-0.04	-0.18	-0.09	0.01	0.06	-0.10	-0.03	0.02	-0.16	-0.16
SBK	0.04	0.02	-0.06	-0.48	0.12	-0.41	-0.16	-0.39	-0.06	-0.05	-0.07	0.19
FSR	0.12	0.38	-0.53	-0.07	-0.02	0.24	0.00	0.16	-0.02	-0.15	0.18	-0.14
SHP	-0.57	0.44	-0.34	-0.34	-0.67	-0.06	-0.72	0.68	-0.19	0.44	0.17	0.24
ANG	0.79	-0.30	0.63	0.30	0.58	-0.08	0.59	-0.37	0.16	-0.29	-0.21	-0.28
SLM	-0.03	0.08	-0.14	0.10	0.01	-0.03	0.03	-0.01	-0.18	0.46	0.29	0.17
REM	0.02	-0.03	-0.05	0.24	0.20	0.19	-0.09	-0.15	-0.28	0.03	-0.06	-0.02
ABG	-0.03	-0.21	0.17	0.40	-0.23	0.58	-0.02	0.12	0.19	-0.24	-0.10	0.05
APN	0.01	0.06	-0.04	0.04	0.06	-0.03	0.00	-0.05	0.07	-0.17	0.31	0.02
VOD	-0.02	-0.20	0.25	0.06	0.03	0.09	-0.08	0.12	-0.08	0.01	-0.41	-0.11
NED	0.04	0.23	-0.13	0.12	0.07	-0.25	0.05	-0.07	0.48	-0.14	-0.09	0.47
GRT	-0.04	0.09	0.11	0.21	-0.01	-0.12	-0.01	-0.02	-0.21	-0.10	0.22	0.01
INP	-0.04	0.17	-0.08	0.11	-0.13	0.09	0.06	-0.23	0.45	0.49	-0.14	-0.21
MNP	0.00	-0.33	-0.18	0.06	-0.05	-0.27	0.00	0.03	0.05	0.01	0.01	-0.04
INL	-0.07	0.16	0.07	0.12	-0.17	-0.25	0.23	0.12	-0.20	0.03	-0.02	0.23

Figure 5.2: Eigenvectors of PVCA