For the given portfolio `Port`, we compute the mean `MuP` and the standard deviation `SigP` of the portfolio return, and we compute the one percentile of the corresponding normal distribution. We learn that only one percent of the time will the return be less than 8% while the above computation was telling us that it should be expected to be less than 9.2% with the same frequency. One cent on the dollar is not much, but for a large portfolio, things add up!

2.5 PRINCIPAL COMPONENT ANALYSIS

Dimension reduction without significant loss of information is one of the main challenges of data analysis. The internet age has seen an exponential growth in the research efforts devoted to the design of efficient codes and compression algorithms. Whether the data are comprised of video streams, images, and/or speech signals, or financial data, finding a basis in which these data can be expressed with a small (or at least a smaller) number of coefficients is of crucial importance. Other important domains of applications are cursed by the high dimensionality of the data. Artificial intelligence applications, especially those involving machine learning and data mining, have the same dimension reduction problems. Pattern recognition problems are closer to the hearts of traditional statisticians. Indeed, regression and statistical classification problems have forced statisticians to face the curse of dimensionality, and to design systematic procedures to encapsulate the important features of high dimensional observations in a small number of variables. Principal component analysis as presented in this chapter, offers an optimal solution to these dimension reduction issues.

2.5.1 Identification of the Principal Components of a Data Set

Principal component analysis (PCA, for short) is a data analysis technique designed for numerical data (as opposed to categorical data). The typical situation that we consider is where the data come in the form of a matrix $[x_{i,j}]_{i=1,\ldots,N,j=1,\ldots,M}$ of real numbers, the entry $x_{i,j}$ representing the value of the i-th observation of the j-th variable. As usual, we follow the convention of labeling the columns of the data matrix by the variables measured, and the rows by the individuals of the population under investigation. Examples are plentiful in most data analysis applications. We give below detailed analyses of several examples from the financial arena.

As we mentioned above, the N members of the population can be identified with the N rows of the data matrix, each one corresponding to an M-dimensional (row) vector of numbers giving the values of the variables measured on this individual. It is often desirable (especially when M is large) to reduce the complexity of the descriptions of the individuals and to replace the M descriptive variables by a smaller number of variables, while at the same time, losing as little information as possible. Let us consider a simple (and presumably naive) illustration of this idea. Imagine

momentarily that all the variables measured are scores of the same nature (for examples they are all lengths expressed in the same unit, or they are all prices expressed in the same currency, ...) so that it would be conceivable to try to characterize each individual by the mean, and a few other numerical statistics computed on all the individual scores. The mean:

$$\overline{x_{i\cdot}} = \frac{x_{i1} + x_{i2} + \cdots + x_{iM}}{M}$$

can be viewed as a linear combination of the individual scores with coefficients $1/M$, $1/M, \ldots, 1/M$. Principal component analysis, is an attempt to describe the individual features in the population in terms of M linear combinations of the original features, as captured by the variables originally measured on the N individuals. The coefficients used in the example of the mean are all non-negative and sum up to one. Even though this convention is very attractive because of the probabilistic interpretation which can be given to the coefficients, we shall use another convention for the linear combinations. We shall allow the coefficients to be of any sign (positive as well as negative) and we normalize them so that the sum of their squares is equal to 1. So if we were to use the mean, we would use the normalized linear combination (NLC, for short) given by:

$$\widetilde{x_{i\cdot}} = \frac{1}{\sqrt{M}} x_{i1} + \frac{1}{\sqrt{M}} x_{i2} + \cdots + \frac{1}{\sqrt{M}} x_{iM}.$$

The goal of principal component analysis is to search for the main sources of variation in the M-dimensional row vectors by identifying M linearly independent and orthogonal NLC's in such a way that a small number of them capture most of the variation in the data. This is accomplished by identifying the eigenvectors and eigenvalues of the covariance matrix C_x of the M column variables. This covariance matrix is defined by:

$$C_x[j, j'] = \frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \overline{x_{\cdot j}})(x_{ij'} - \overline{x_{\cdot j'}}), \qquad j, j' = 1, \ldots, M,$$

where we used the standard notation:

$$\overline{x_{\cdot j}} = \frac{x_{1j} + x_{2j} + \cdots + x_{Nj}}{N}$$

for the mean of the j-th variable over the population of N individuals. It is easy to check that the matrix C_x is symmetric (hence diagonalizable) and non-negative definite (which implies that all the eigenvalues are non-negative). One usually orders the eigenvalues in decreasing order, say:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M \geq 0.$$

The corresponding eigenvectors are called the loadings. In practice we choose c_1 to be a normalized eigenvector corresponding to the eigenvalue λ_1, c_2 to be a normalized eigenvector corresponding to the eigenvalue $\lambda_2, \ldots$, and finally c_M to be a

normalized eigenvector corresponding to the eigenvalue λ_M, and we make sure that all the vectors c_j are orthogonal to each other. This is automatically true when the eigenvalues λ_j are simple. See the discussion below for the general case. Recall that we say a vector is normalized if the sum of the squares of its components is equal to 1. If we denote by C the $M \times M$ matrix formed by the M column vectors containing the components of the vectors $c_1, c_2, \ldots, c_M$ in this order, this matrix is orthogonal (since it it a matrix transforming one orthonormal basis into another) and it satisfies:

$$C_x = C^t D C$$

where we use the notation ^{t} to denote the transpose of a matrix or a vector, and where D is the $M \times M$ diagonal matrix with $\lambda_1, \lambda_2, \ldots, \lambda_M$ on the diagonal. Notice the obvious lack of uniqueness of the above decomposition. In particular, if c_j is a normalized eigenvector associated to the eigenvalue λ_j, so is $-c_j$! This is something one should keep in mind when plotting the eigenvectors c_j, and when trying to find an intuitive interpretation for the features of the plots. However, this sign flip is easy to handle, and fortunately, it is the only form of non uniqueness when the eigenvalues are simple (i.e. nondegenerate). The ratio:

$$\frac{\lambda_j}{\sum_{j'=1}^{M} \lambda_{j'}}$$

of a given eigenvalue to the trace of C_x (i.e. the sum of its eigenvalues) has the interpretation of the proportion of the variation explained by the corresponding eigenvector, i.e. the loading c_j. In order to justify this statement, we appeal to the Raleigh-Ritz variational principle from linear algebra. Indeed, according to this principle, the eigenvalues and their corresponding eigenvectors can be characterized recursively in the following way. The largest eigenvalue λ_1 appears as the maximum:

$$\lambda_1 = \max_{x \in \mathbb{R}^M, \|x\|=1} x^t C_x x$$

while the corresponding eigenvector c_1 appears as the argument of this maximization problem:

$$c_1 = \arg \max_{x \in \mathbb{R}^M, \|x\|=1} x^t C_x x.$$

If we recall the fact that $x^t C_x x$ represents the quadratic variation (empirical variance) of the NLC's $x^t x_1.$, $x^t x_2.$,..., $x^t x_N.$, λ_1 can be interpreted as the maximal quadratic variation when we consider all the possible (normalized) linear combinations of the M original measured variables. Similarly, the corresponding (normalized) eigenvector has precisely the interpretation of this NLC which realizes the maximum variation.

As we have already pointed out, the first loading is uniquely determined up to a sign change if the eigenvalue λ_1 is simple. If this is not the case, and if we denote by m_1 the multiplicity of the eigenvalue λ_1, we can choose any orthonormal set $\{c_1, \cdots, c_{m_1}\}$ in the eigenspace of λ_1 and repeat the eigenvalue λ_1, m_1 times in the

list of eigenvalues (and on the diagonal of D as well). This lack of uniqueness is not a mathematical difficulty, it is merely annoying. Fortunately, it seldom happens in practice ! We shall assume that all the eigenvalues are simple (i.e. non-degenerate) for the remainder of our discussion. If they were not, we would have to repeat them according to their multiplicities.

Next, still according to the Raleigh-Ritz variation principle, the second eigenvalue λ_2 appears as the maximum:

$$\lambda_2 = \max_{x \in \mathbb{R}^M, \|x\|=1, x \perp c_1} x^t C_x x$$

while the corresponding eigenvector c_2 appears as the argument of this maximization problem:

$$c_2 = \arg \max_{x \in \mathbb{R}^M, \|x\|=1, x \perp c_1} x^t C_x x.$$

The interpretation of this statement is the following: if we avoid any overlap with the loading already identified (i.e. if we restrict ourselves to NLC's x which are orthogonal to c_1), then the maximum quadratic variation will be λ_2 and any NLC realizing this maximum variation can be used for c_2. We can go on and identify in this way all the eigenvalues λ_j (having to possibly repeat them according to their multiplicities) and the loadings c_j's.

Armed with a new basis of $\mathbb{R}^M$, the next step is to rewrite the data observations (i.e. the N rows of the data matrix) in this new basis. This is done by multiplying the data matrix by the *change of basis* matrix (i.e. the matrix whose columns are the eigenvectors identified earlier). The result is a new $N \times M$ matrix whose columns are called *principal components.* Their relative importance is given by the proportion of the variance explained by the loadings, and for that reason, one typically considers only the first few principal components, the remaining ones being ignored and/or treated as noise.

2.5.2 PCA with `S-Plus`

The principal component analysis of a data set is performed in `S-Plus` with the function `princomp`, which returns an object of class `princomp` that can be printed and plotted with generic methods. Illustrations of the calls to this function and of the interpretation of the results are given in the next subsections in which we discuss several financial applications of the PCA.

2.5.3 Effective Dimension of the Space of Yield Curves

Our first application concerns the markets of fixed income securities which we will introduce in Section 3.8. The term structure of interest rates is conveniently captured by the daily changes in the yield curve. The dimension of the space of all possible yield curves is presumably very large, potentially infinite if we work in the idealized world of continuous-time finance. However, it is quite sensible to try to approximate

these curves by functions from a class chosen in a parsimonious way. Without any a priori choice of the type of functions to be used to approximate the yield curve, PCA can be used to extract, one by one, the components responsible for the variations in the data.

PCA of the Yield Curve

For the purposes of illustration, we use data on the US yield curve as provided by the Bank of International Settlements (BIS, for short). These data are the result of a nonparametric processing (smoothing spline regression, to be specific) of the raw data. The details will be given in Section 4.4 of Chapter 4, but for the time being, we shall ignore the possible effects of this pre-processing of the raw data. The data are imported into an S-object named `us.bis.yield` which gives, for each of the 1352 successive trading days following January 3rd 1995, the yields on the US Treasury bonds for times to maturity

$$x = 0, 1, 2, 3, 4, 5, 5.5, 6.5, 7.5, 8.5, 9.5 \quad \text{months.}$$

We run a PCA on these data with the following S-Plus commands:

```
> dim(us.bis.yield)
[1] 1352   11
> us.bis.yield.pca <- princomp(us.bis.yield)
> plot(us.bis.yield.pca)
[1]  0.700000  1.900000  3.100000  4.300000  5.500000
[6] 6.700000  7.900000  9.099999  10.299999 11.499999
> title("Proportions of the Variance Explained
                                    by the Components")
```

The results are reproduced in Figure 2.20 which gives the proportions of the variation explained by the various components. The first three eigenvectors of the covariance matrix (the so-called loadings) explain 99.9% of the total variation in the data. This suggests that the effective dimension of the space of yield curves could be three. In other words, any of the yield curves from this period can be approximated by a linear combination of the first three loadings, the relative error being very small. In order to better understand the far reaching implications of this statement we plot the first four loadings.

```
> X <- c(0,1,2,3,4,5,5.5,6.5,7.5,8.5,9.5)
> par(mfrow=c(2,2))
> plot(X,us.bis.yield.pca$loadings[,1],ylim=c(-.7,.7))
> lines(X,us.bis.yield.pca$loadings[,1])
> plot(X,us.bis.yield.pca$loadings[,2],ylim=c(-.7,.7))
> lines(X,us.bis.yield.pca$loadings[,2])
> plot(X,us.bis.yield.pca$loadings[,3],ylim=c(-.7,.7))
> lines(X,us.bis.yield.pca$loadings[,3])
> plot(X,us.bis.yield.pca$loadings[,4],ylim=c(-.7,.7))
```

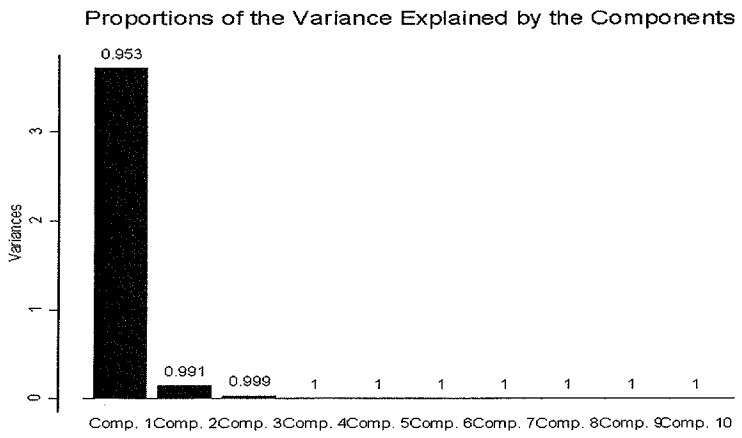

Fig. 2.20. Proportions of the variance explained by the components of the PCA of the daily changes in the US yield curve.

```
> lines(X,us.bis.yield.pca$loadings[,4])
> par(mfrow=c(1,1))
> title("First Four Loadings of the US Yield Curves")
```

The results are reproduced in Figure 2.21. The first loading is essentially flat, so a component on this loading will essentially represent the average yield over the maturities, and the effect of this most-important component on the actual yield curve is a parallel shift. Because of the monotone and increasing nature of the second loading, the second component measures the upward trend (if the component is positive, and the downward trend otherwise) in the yield. This second factor is interpreted as the tilt of the yield curve. The shape of the third loading suggests that the third component captures the curvature of the yield curve. Finally, the shape of the fourth loading does not seem to have an obvious interpretation. It is mostly noise (remember that most of the variations in the yield curve are explained by the first three components). These features are very typical, and they should be expected in most PCA's of the term structure of interest rates.

The fact that the first three components capture so much of the yield curve may seem strange when compared to the fact that some estimation methods, which we discuss later in the book, use parametric families with more than three parameters! There is no contradiction there. Indeed, for the sake of illustration, we limited the analysis of this section to the first part of the yield curve. Restricting ourselves to short maturities makes it easier to capture all the features of the yield curve in a small number of functions with a clear interpretation.

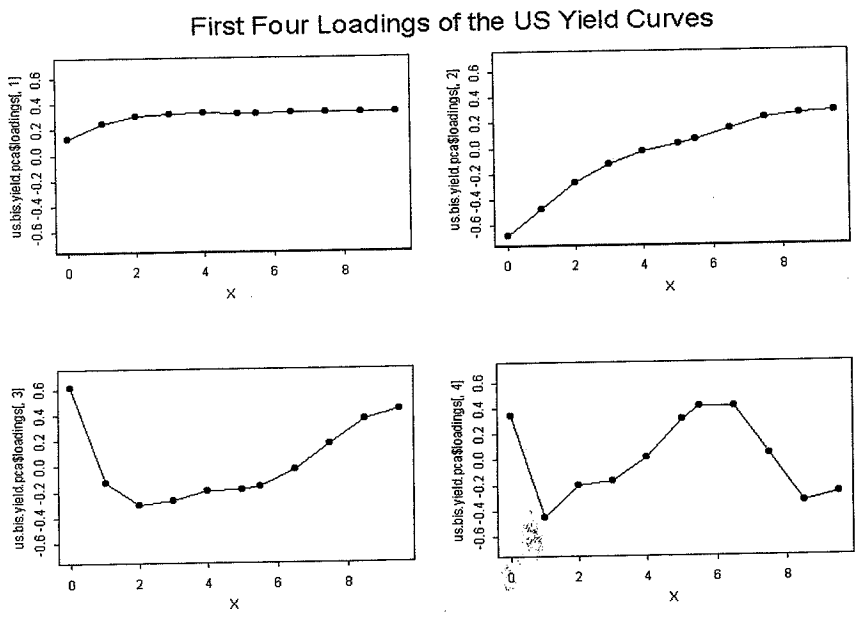

Fig. 2.21. From left to right and top to bottom, sequential plots of the first four US yield loadings.

2.5.4 Swap Rate Curves

Swap contracts have been traded publicly since 1981. As of today, they are the most popular fixed income derivatives. Because of this popularity, the swap markets are extremely liquid, and as a consequence, they can be used to hedge interest-rate risk of fixed income portfolios at a low cost. The estimation of the term-structure of swap rates is important in this respect and the PCA which we present below is the first step toward a better understanding of this term structure.

Swap Contracts and Swap Rates

As implied by its name, a swap contract obligates two parties to exchange (or swap) some specified cash flows at agreed upon times. The most common swap contracts are interest rate swaps. In such a contract, one party, say counter-party A, agrees to make interest payments determined by an instrument P_A (say, a 10 year US Treasury bond rate), while the other party, say counter-party B, agrees to make interest payments determined by another instrument P_B (say, the London Interbank Offer Rate – LIBOR for short) Even though there are many variants of swap contracts, in a typical contract, the principal on which counter-party A makes interest payments is equal to the principal on which counterparty B makes interest payments. Also, the

payment schedules are identical and periodic, the payment frequency being quarterly, semi-annually, ...

It is not difficult to infer from the above discussion that a swap contract is equivalent to a portfolio of forward contracts, but we shall not use this feature here. In this section, we shall restrict ourselves to the so-called plain vanilla contracts involving a fixed interest rate and the 3 or 6 months LIBOR rate.

We will not attempt to derive here a formula for the price of a swap contract, neither will we try to define rigorously the notion of swap rate. These derivations are beyond the scope of this book. See the Notes & Complements at the end of the chapter for references to appropriate sources. We shall use only the intuitive idea of the swap rate being a rate at which both parties will agree to enter into the swap contract.

PCA of the Swap Rates

Our second application of principal component analysis concerns the term structure of swap rates as given by the swap rate curves. As before, we denote by M the dimension of the vectors. We use data downloaded from `Data Stream`. It is quite likely that the raw data have been processed, but we are not quite sure what kind of manipulation is performed by `Data Stream`, so for the purposes of this illustration, we shall ignore the possible effects of the pre-processing of the data. In this example, the day t labels the rows of the data matrix. The latter has $M = 15$ columns, containing the swap rates with maturities T conveniently labeled by the times to maturity $x = T - t$, which have the values $1, 2, \ldots, 10, 12, 15, 20, 25, 30$ years in the present situation. We collected these data for each day t of the period from May 1998 to March 2000, and we rearranged the numerical values in a matrix $R = [r_{i,j}]_{i=1,\ldots,N,\ j=1,\ldots,M}$. Here, the index j stands for the time to maturity, while the index i codes the day the curve is observed.

The data is contained in the `S` object `swap`. The PCA is performed in `S-Plus` with the command:

```
> dim(swap)
[1] 496  15
> swap.pca <- princomp(swap)
> plot(swap.pca)
[1]  0.700000  1.900000  3.100000  4.300000  5.500000
[6] 6.700000  7.900000  9.099999  10.299999 11.499999
> title("Proportions of the Variance Explained by
                                        the Components")
> YEARS <- c(1,2,3,4,5,6,7,8,9,10,12,15,20,25,30)
> par(mfrow=c(2,2))
> plot(YEARS,swap.pca$loadings[,1],ylim=c(-.6,.6))
> lines(YEARS,swap.pca$loadings[,1])
> plot(YEARS,swap.pca$loadings[,2],ylim=c(-.6,.6))
> lines(YEARS,swap.pca$loadings[,2])
> plot(YEARS,swap.pca$loadings[,3],ylim=c(-.6,.6))
```

```
> lines(YEARS,swap.pca$loadings[,3])
> plot(YEARS,swap.pca$loadings[,4],ylim=c(-.6,.6))
> lines(YEARS,swap.pca$loadings[,4])
> par(mfrow=c(1,1))
> title("First Four Loadings of the Swap Rates")
```

Figure 2.22 gives the proportions of the variation explained by the various components, while Figure 2.23 gives the plots of the first four eigenvectors.

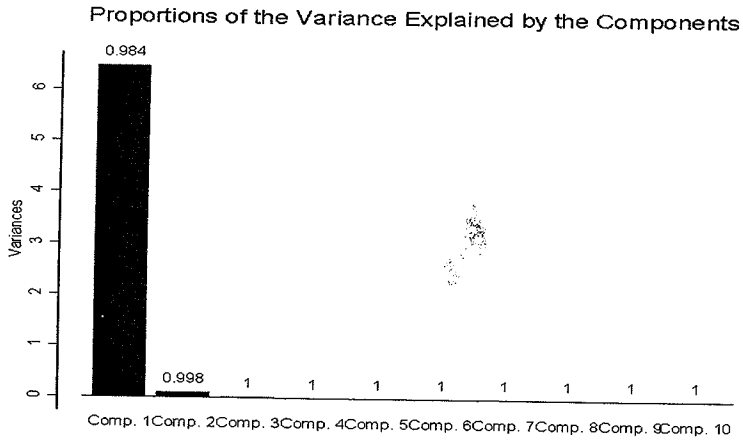

Fig. 2.22. Proportions of the variance explained by the components of the PCA of the daily changes in the swap rates for the period from May 1998 to March 2000.

Looking at Figure 2.23 one sees that the remarks made above, for the interpretation of the results in terms of a parallel shift, a tilt and a curvature component, do apply to the present situation as well.

Since such an overwhelming proportion of the variation is explained by one single component, it is often recommended to remove the effect of this component from the data, (here, that would amount to subtracting the overall mean rate level) and to perform the PCA on the transformed data (here, the fluctuations around the mean rate level).

APPENDIX 1: CALCULUS WITH RANDOM VECTORS AND MATRICES

The nature and technical constructs of this chapter justify our spending some time discussing the properties of random vectors (as opposed to random variables) and reviewing the fundamental results of the calculus of probability with random vectors. Their definition is very natural: a random vector is a vector whose entries are random