$$E(\text{RSS}) = \sum_{t=1}^{n} Ee_t^2 = \sum_{t=1}^{n} \left[E(y_t - \widehat{\boldsymbol{\beta}}^T \mathbf{x}_t)\right]^2 + \text{tr}\big(\text{Cov}(\mathbf{e})\big).$$

Combining this with (2.6) yields (1.31).

2.2 Principal component analysis (PCA)

2.2.1 Basic definitions

Definition 2.1. Let $\mathbf{V}$ be a $p \times p$ matrix. A complex number λ is called an *eigenvalue* of $\mathbf{V}$ if there exists a $p \times 1$ vector $\mathbf{a} \neq \mathbf{0}$ such that $\mathbf{Va} = \lambda \mathbf{a}$. Such a vector $\mathbf{a}$ is called an *eigenvector* of $\mathbf{V}$ corresponding to the eigenvalue λ.

We can rewrite $\mathbf{Va} = \lambda\mathbf{a}$ as $(\mathbf{V} - \lambda\mathbf{I})\mathbf{a} = 0$. Since $\mathbf{a} \neq \mathbf{0}$, this implies that λ is a solution of the equation $\det(\mathbf{V} - \lambda\mathbf{I}) = 0$. Since $\det(\mathbf{V} - \lambda\mathbf{I})$ is a polynomial of degree p in λ, there are p eigenvalues, not necessarily distinct. If $\mathbf{V}$ is symmetric, then all its eigenvalues are real and can be ordered as $\lambda_1 \geq \cdots \geq \lambda_p$. Moreover,

$$\text{tr}(\mathbf{V}) = \lambda_1 + \cdots + \lambda_p, \qquad \det(\mathbf{V}) = \lambda_1 \ldots \lambda_p. \tag{2.7}$$

If $\mathbf{a}$ is an eigenvector of $\mathbf{V}$ corresponding to the eigenvalue λ, then so is $c\mathbf{a}$ for $c \neq 0$. Moreover, premultiplying $\lambda\mathbf{a} = \mathbf{Va}$ by $\mathbf{a}^T$ yields

$$\lambda = \mathbf{a}^T\mathbf{Va}/||\mathbf{a}||^2. \tag{2.8}$$

In the rest of this section, we shall focus on the case where $\mathbf{V}$ is the covariance matrix of a random vector $\mathbf{x} = (X_1, \ldots, X_p)^T$. Not only are its eigenvalues real because $\mathbf{V}$ is symmetric, but they are also nonnegative since $\mathbf{V}$ is nonnegative definite; see Section 1.1.3. Consider the linear combination $\mathbf{a}^T\mathbf{x}$ with $||\mathbf{a}|| = 1$ that has the largest variance over all such linear combinations. To maximize $\mathbf{a}^T\mathbf{Va}(= \text{Var}(\mathbf{a}^T\mathbf{x}))$ over $\mathbf{a}$ with $||\mathbf{a}|| = 1$, introduce the Lagrange multiplier λ to obtain

$$\frac{\partial}{\partial a_i}\left\{\mathbf{a}^T\mathbf{Va} + \lambda(1 - \mathbf{a}^T\mathbf{a})\right\} = 0 \quad \text{for } i = 1, \ldots, p. \tag{2.9}$$

The p equations in (2.9) can be written as the linear system $\mathbf{Va} = \lambda\mathbf{a}$. Since $\mathbf{a} \neq \mathbf{0}$, this implies that λ is an eigenvalue of $\mathbf{V}$ and $\mathbf{a}$ is the corresponding eigenvector, and that $\lambda = \mathbf{a}^T\mathbf{Va}$ by (2.8).

Let $\lambda_1 = \max_{\mathbf{a}:||\mathbf{a}||=1} \mathbf{a}^T\mathbf{Va}$ and $\mathbf{a}_1$ be the corresponding eigenvector with $||\mathbf{a}_1|| = 1$. Next consider the linear combination $\mathbf{a}^T\mathbf{x}$ that maximizes $\text{Var}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}^T\mathbf{Va}$ subject to $\mathbf{a}_1^T\mathbf{a} = 0$ and $||\mathbf{a}|| = 1$. Introducing the Lagrange multipliers λ and η, we obtain

$$\frac{\partial}{\partial a_i}\{\mathbf{a}^T\mathbf{V}\mathbf{a} + \lambda(1 - \mathbf{a}^T\mathbf{a}) + \eta\mathbf{a}_1^T\mathbf{a}\} = 0 \text{ for } i = 1, ..., p.$$

As in (2.9), this implies that the Lagrange multiplier λ is an eigenvalue of $\mathbf{V}$ with corresponding unit eigenvector $\mathbf{a}_2$ that is orthogonal to $\mathbf{a}_1$. Proceeding inductively in this way, we obtain the eigenvalue $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ of $\mathbf{V}$ with the optimization characterization

$$\lambda_{k+1} = \max_{\mathbf{a}:\|\mathbf{a}\|=1, \mathbf{a}^T\mathbf{a}_j=0 \text{ for } 1\leq j\leq k} \mathbf{a}^T\mathbf{V}\mathbf{a}. \tag{2.10}$$

The maximizer $\mathbf{a}_{k+1}$ of the right-hand side of (2.10) is an eigenvector corresponding to the eigenvalue λ_{k+1}.

Definition 2.2. $\mathbf{a}_i^T\mathbf{x}$ is called the ith *principal component* of $\mathbf{x}$.

2.2.2 Properties of principal components

(a) From the preceding derivation, $\lambda_i = \text{Var}(\mathbf{a}_i^T\mathbf{x})$.

(b) The elements of the eigenvectors $\mathbf{a}_i$ are called *factor loadings* in PCA (principal component analysis), which can be carried out by the software package `princomp` in `R` or `Splus`. Since $\mathbf{a}_i^T\mathbf{a}_j = 0$ for $i \neq j$ and $\|\mathbf{a}_i\| = 1$, $(\mathbf{a}_1, \ldots, \mathbf{a}_p)$ is an orthogonal matrix and therefore we can decompose the identity matrix $\mathbf{I}$ as

$$\mathbf{I} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)(\mathbf{a}_1, \ldots, \mathbf{a}_p)^T = \mathbf{a}_1\mathbf{a}_1^T + \cdots + \mathbf{a}_p\mathbf{a}_p^T. \tag{2.11}$$

More importantly, summing $\lambda_i\mathbf{a}_i\mathbf{a}_i^T = \mathbf{V}\mathbf{a}_i\mathbf{a}_i^T$ over i and applying (2.11), we obtain the following decomposition of $\mathbf{V}$ into p rank-one matrices:

$$\mathbf{V} = \lambda_1\mathbf{a}_1\mathbf{a}_1^T + \cdots + \lambda_p\mathbf{a}_p\mathbf{a}_p^T. \tag{2.12}$$

(c) Since $\mathbf{V} = \text{Cov}(\mathbf{x})$ and $\mathbf{x} = (X_1, \ldots, X_p)^T$, $\text{tr}(\mathbf{V}) = \sum_{i=1}^p \text{Var}(X_i)$. Hence it follows from (2.7) that

$$\lambda_1 + \cdots + \lambda_p = \sum_{i=1}^p \text{Var}(X_i). \tag{2.13}$$

An important goal of PCA is to determine if the first few principal components can account for most of the overall variance $\sum_{i=1}^p \text{Var}(X_i)$. In view of (2.13), this amounts to determining whether

$$\Big(\sum_{i=1}^k \lambda_i\Big)\Big/\text{tr}(\mathbf{V}) \text{ is near 1 for some small } k. \tag{2.14}$$

The proportion in (2.14) can be evaluated by using `screeplot`, which plots $\lambda_i/\text{tr}(\mathbf{V})$ for each i, in `R` or `Splus`.

Singular-value decomposition

We can write the representation of $\mathbf{V}$ in (2.12) as

$$\mathbf{V} = \mathbf{Q}\text{diag}(\lambda_1, \ldots, \lambda_p)\mathbf{Q}^T, \tag{2.15}$$

where $\mathbf{Q} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$. The matrix $\mathbf{Q}$ is orthogonal, and (2.15) is called the *singular-value decomposition* of $\mathbf{V}$. We can use (2.15) to define the square root of $\mathbf{V}$ by $\mathbf{V}^{1/2} = \mathbf{Q}\text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})\mathbf{Q}^T$, noting that $\mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{Q}\text{diag}(\lambda_1, \ldots, \lambda_p)\mathbf{Q}^T = \mathbf{V}$.

PCA of sample covariance matrix and sampling theory of $\widehat{\lambda}_j$, $\widehat{\mathbf{a}}_j$

Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n independent observations from a multivariate population with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. The mean vector $\boldsymbol{\mu}$ can be estimated by $\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$, and the covariance matrix can be estimated by

$$\widehat{\mathbf{V}} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T/(n-1),$$

which is the sample covariance matrix. Let $\widehat{\mathbf{a}}_j = (\widehat{a}_{1j}, \ldots, \widehat{a}_{pj})^T$ be the eigenvector corresponding to the jth largest eigenvalue $\widehat{\lambda}_j$ of the sample covariance matrix $\widehat{\mathbf{V}}$. For fixed p, the following asymptotic results have been established under the assumption $\lambda_1 > \cdots > \lambda_p > 0$ as $n \to \infty$:

(i) $\sqrt{n}(\widehat{\lambda}_i - \lambda_i)$ has a limiting $N(0, 2\lambda_i^2)$ distribution. Moreover, for $i \neq j$, the limiting distribution of $\sqrt{n}(\widehat{\lambda}_i - \lambda_i, \widehat{\lambda}_j - \lambda_j)$ is that of two independent normals.

(ii) $\sqrt{n}(\widehat{\mathbf{a}}_i - \mathbf{a}_i)$ has a limiting normal distribution with mean 0 and covariance matrix

$$\sum_{h\neq i} \frac{\lambda_i\lambda_h}{(\lambda_i - \lambda_h)^2}\mathbf{a}_h\mathbf{a}_h^T.$$

Since $\mathbf{a}_i$ is not uniquely defined because multiplication by -1 still preserves the eigenvector property $\mathbf{V}\mathbf{a} = \lambda\mathbf{a}$ with $\|\mathbf{a}\| = 1$, the preceding result means that $\widehat{\mathbf{a}}_i$ and $\mathbf{a}_i$ are chosen so that they have the same sign.

In view of (ii) above, even when p is large so that $\widehat{\mathbf{V}}$ estimates a larger number, $p(p+1)/2$, of parameters, if many eigenvalues λ_h are small compared with the largest λ_i's, then they have small contributions to the sum in (ii), in which $\lambda_i\lambda_h/(\lambda_i - \lambda_h)^2 \doteq \lambda_h/\lambda_i$ when λ_h is much smaller than λ_i, noting that

$\text{tr}(\mathbf{a}_h \mathbf{a}_h^T) = 1$. Hence the covariance structure of the data can be approximated by a few principal components with large eigenvalues. In this approximation, only a few, say k, principal components are involved in the covariance matrix of the estimate $\widehat{\mathbf{a}}_j$ for $1 \le j \le k$.

Representation of data in terms of principal components

Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$, $i = 1, \dots, n$, be the multivariate sample. Let $\mathbf{X}_k = (x_{1k}, \dots, x_{nk})^T$, $1 \le k \le p$, and define

$$\mathbf{Y}_j = \widehat{a}_{1j}\mathbf{X}_1 + \cdots + \widehat{a}_{pj}\mathbf{X}_p, \qquad 1 \le j \le p,$$

where $\widehat{\mathbf{a}}_j = (\widehat{a}_{1j}, \dots, \widehat{a}_{pj})^T$ is the eigenvector corresponding to the jth largest eigenvalue $\widehat{\lambda}_j$ of the sample covariance matrix $\widehat{\mathbf{V}}$ with $\|\widehat{\mathbf{a}}_j\| = 1$. Since the matrix $\widehat{\mathbf{A}} := (\widehat{a}_{ij})_{1 \le i,j \le p}$ is orthogonal (i.e., $\widehat{\mathbf{A}}\widehat{\mathbf{A}}^T = \mathbf{I}$), it follows that the observed data $\mathbf{X}_k$ can be expressed in terms of the "principal components" $\mathbf{Y}_j$ as

$$\mathbf{X}_k = \widehat{a}_{k1}\mathbf{Y}_1 + \cdots + \widehat{a}_{kp}\mathbf{Y}_p. \tag{2.16}$$

Moreover, the sample correlation matrix of the transformed data y_{tj} is the identity matrix.

PCA of correlation matrices

As shown above, the key ingredient of PCA is the decomposition $\mathbf{V} = \lambda_1 \mathbf{a}_1 \mathbf{a}_1^T + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}_p^T$ of a nonnegative definite matrix $\mathbf{V}$. An alternative to $\text{Cov}(\mathbf{x})$ is the correlation matrix $\text{Corr}(\mathbf{x})$, which is also nonnegative definite, consisting of the correlation coefficients $\text{Corr}(X_i, X_j)$, $1 \le i, j \le p$. In fact, $\text{Corr}(\mathbf{x}) = \text{Cov}(X_1/\sigma_1, \dots, X_p/\sigma_p)$ is itself a covariance matrix, where σ_i is the standard deviation of X_i. Since a primary goal of PCA is to uncover low-dimensional subspaces around which $\mathbf{x}$ tends to concentrate, scaling the components of $\mathbf{x}$ by their standard deviations may work better for this purpose, and applying PCA to sample correlation matrices may give more stable results.

2.2.3 An example: PCA of U.S. Treasury-LIBOR swap rates

PCA can be implemented by the following software packages:

```
R/Splus:    princomp, screeplot, biplot.princomp;
MATLAB:     princomp, pcacov, biplot.
```

We now apply PCA to account for the variance of daily changes in swap rates with a few principal components (or factors). Details of interest rate swap contracts, which involve exchanging the U.S. Treasury bond rate with the London Interbank Offered Rate (LIBOR), and the associated swap rates for different maturities are given in Chapter 10. The top panel of Figure 2.1 plots the daily swap rates r_{kt} for four of the eight maturities $T_k = 1, 2, 3, 4, 5, 7, 10,$ and 30 years from July 3, 2000 to July 15, 2005. The data are obtained from www.Economagic.com. Let $d_{kt} = r_{kt} - r_{k,t-1}$ denote the daily changes in the k-year swap rates during the period. The middle and bottom panels plot the differenced time series d_{kt} for 1-year and 30-year swap rates, respectively. Further discussion of these plots will be given in Chapter 5 (Section 5.2.3). The sample mean vector and the sample covariance and correlation matrices of the difference data $\{(d_{1t}, \dots, d_{8t}), 1 \le t \le 1256\}$ are

$$\widehat{\boldsymbol{\mu}} = -\left(2.412\ 2.349\ 2.293\ 2.245\ 2.196\ 2.158\ 2.094\ 1.879\right)^T \times 10^{-3}.$$

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.233 \\ 0.300\ 0.438 \\ 0.303\ 0.454\ 0.488 \\ 0.297\ 0.453\ 0.492\ 0.508 \\ 0.292\ 0.451\ 0.494\ 0.514\ 0.543 \\ 0.268\ 0.421\ 0.466\ 0.490\ 0.513\ 0.498 \\ 0.240\ 0.384\ 0.431\ 0.458\ 0.477\ 0.472\ 0.467 \\ 0.169\ 0.278\ 0.318\ 0.344\ 0.362\ 0.365\ 0.369\ 0.322 \end{pmatrix} \times 10^{-2},$$

$$\widehat{\mathbf{R}} = \begin{pmatrix} 1.000 \\ 0.941\ 1.000 \\ 0.899\ 0.983\ 1.000 \\ 0.862\ 0.961\ 0.988\ 1.000 \\ 0.821\ 0.925\ 0.960\ 0.979\ 1.000 \\ 0.787\ 0.901\ 0.945\ 0.973\ 0.986\ 1.000 \\ 0.729\ 0.850\ 0.902\ 0.941\ 0.948\ 0.978\ 1.000 \\ 0.618\ 0.742\ 0.803\ 0.852\ 0.866\ 0.912\ 0.951\ 1.000 \end{pmatrix}.$$

Table 2.1 gives the results of PCA using both the covariance and correlation matrices. Also given there are eigenvalues, factor loadings (eigenvectors), and the proportions of overall variance explained by the principal components.

Swap rate movements indicated by PCA

Let $x_{tk} = (d_{tk} - \widehat{\mu}_k)/\widehat{\sigma}_k$, where $\widehat{\sigma}_k$ is the sample standard deviation of d_{tk} and is given by the square root of the kth diagonal element of $\widehat{\boldsymbol{\Sigma}}$. We can use (2.16) to represent $\mathbf{X}_k = (x_{1k}, \dots, x_{nk})^T$ in terms of the principal components $\mathbf{Y}_1, \dots, \mathbf{Y}_8$. Table 2.1 shows the PCA results using the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ and the sample correlation matrix $\widehat{\mathbf{R}}$. From part (b) of Table 2.1,

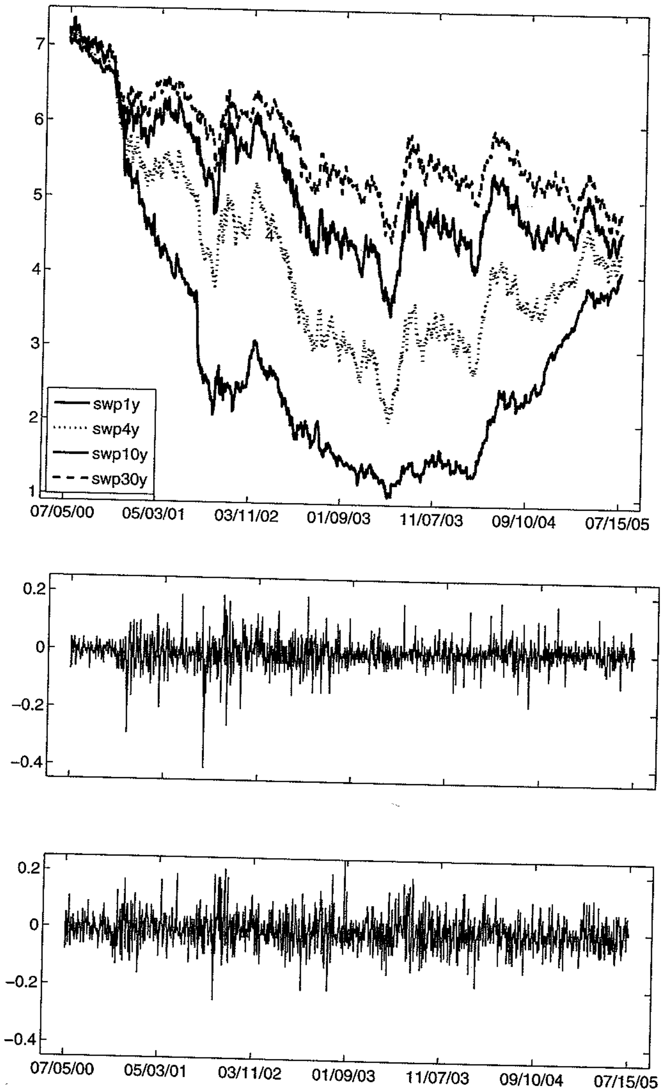

Fig. 2.1. Swap rates from July 3, 2000 to July 15, 2005. Top panel: the original time series. Middle and bottom panels: the differenced series for 1-year (middle) and 30-year (bottom) swap rates.

Table 2.1. PCA of covariance and correlation matrices of swap rate data.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
(a) Using sample covariance matrix								
Eigenvalue ($\times 10^4$)	324.1	18.44	3.486	1.652	0.984	0.473	0.292	0.253
Proportion	0.926	0.053	0.010	0.005	0.003	0.001	0.001	0.001
Factor loadings	0.231	0.491	−0.535	−0.580	−0.081	−0.275	0.006	−0.030
	0.351	0.431	−0.150	0.279	0.049	0.725	−0.050	0.246
	0.381	0.263	0.078	0.456	0.057	−0.265	−0.001	−0.706
	0.393	0.087	0.175	0.299	−0.069	−0.534	0.062	0.652
	0.404	−0.054	0.486	−0.447	0.414	0.131	0.455	−0.054
	0.387	−0.211	0.238	−0.262	−0.089	0.052	−0.818	−0.038
	0.365	−0.395	−0.127	−0.034	−0.735	0.154	0.344	−0.103
	0.279	−0.541	−0.588	0.138	0.513	−0.032	−0.003	0.027
(b) Using sample correlation matrix								
Eigenvalue	7.265	0.548	0.103	0.041	0.022	0.011	0.006	0.005
Proportion	0.908	0.067	0.013	0.005	0.003	0.001	0.001	0.001
Factor loadings	0.324	0.599	−0.586	−0.402	−0.025	−0.175	0.002	−0.013
	0.356	0.352	0.037	0.436	0.004	0.717	−0.044	0.205
	0.364	0.187	0.215	0.432	0.007	−0.349	0.007	−0.691
	0.368	0.043	0.253	0.206	−0.076	−0.533	0.045	0.681
	0.365	−0.064	0.385	−0.455	0.499	0.151	0.482	-0.056
	0.365	−0.192	0.216	−0.342	−0.014	0.075	−0.811	−0.052
	0.356	−0.348	−0.059	−0.154	−0.768	0.143	0.324	−0.098
	0.328	−0.565	−0.589	0.267	0.393	−0.025	0.002	0.020

the proportion of the overall variance explained by the first principal component is $7.265/8 = 90.8\%$. Moreover, the second principal component explains $0.548/8 = 6.7\%$ and the third principal component explains $0.103/8 = 1.3\%$ of the overall variance. Hence 98.9% of the overall variance is explained by the first three principal components, yielding the approximation

$$(\mathbf{d}_k - \widehat{\mu}_k)/\widehat{\sigma}_k \doteq \widehat{a}_{k1}\mathbf{Y}_1 + \widehat{a}_{k2}\mathbf{Y}_2 + \widehat{a}_{k3}\mathbf{Y}_3 \tag{2.17}$$

for the differenced swap rates.

In Figure 2.2, the bottom panel plots the factor loadings of the first three principal components (or the entries of the eigenvectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$) versus the maturities of the swap rates. The top panel shows the variances of all principal components. The graphs of the factor loadings show the following constituents of interest rate or yield curve movements documented in the literature:

(a) *Parallel shift component.* The factor loadings of the first principal component are roughly constant over different maturities. This means that a change in the swap rate for one maturity is accompanied by roughly the same change for other maturities. Indeed, if $\widehat{a}_{11}, \dots, \widehat{a}_{p1}$ (the components of $\widehat{\mathbf{a}}_1$) are equal, then the first summand on the right-hand side of (2.17) is the same for all maturities T_k.

(b) *Tilt component.* The factor loadings of the second principal component have a monotonic change with maturities. Changes in short-maturity and long-maturity swap rates in this component have opposite signs.

(c) *Curvature component.* The factor loadings of the third principal component are different for the midterm rates and the average of short- and long-term rates, revealing a curvature of the graph that resembles the convex shape of the relationship between the rates and their maturities.

2.3 Multivariate normal distribution

2.3.1 Definition and density function

(i) An $m \times 1$ random vector $\mathbf{Z} = (Z_1, \dots, Z_m)^T$ is said to have the m-variate standard normal distribution if $Z_1, \dots, Z_m$ are independent standard normal random variables. The joint density function of $\mathbf{Z}$ is therefore

$$f(\mathbf{z}) = \prod_{i=1}^{m} \left\{ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2 \right) \right\} = (2\pi)^{-m/2} \exp(-\mathbf{z}^T \mathbf{z}/2).$$

(ii) An $m \times 1$ random vector $\mathbf{Y}$ is said to have a multivariate normal distribution if it is of the form $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}$, where $\mathbf{Z}$ is standard m-variate normal and $\boldsymbol{\mu}$ and $\mathbf{A}$ are $m \times 1$ and $m \times m$ nonrandom matrices.

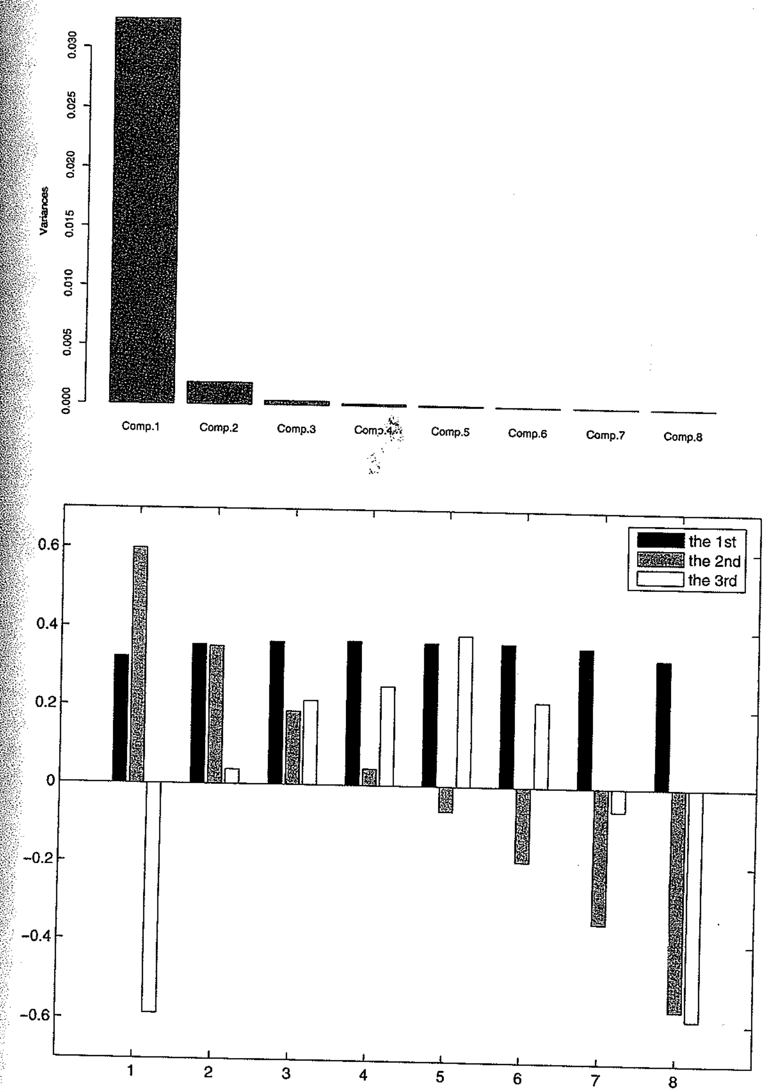

Fig. 2.2. PCA of correlation matrix. Top panel: variances of principal components. Bottom panel: eigenvectors of the first three principal components, which represent the parallel shift, tilt, and convexity components of swap rate movements.