8.2 Prove eq. (8.10).

1.9 Principal Component Analysis$^{\mathrm{B}}$

The LS procedure was introduced in the previous sections as a *descriptive* device. Given a certain set of observations, how do we fit a linear relation as well as possible? The method answers this question by minimizing the sum of squares of the discrepancies. We shall meet LS very frequently in later chapters, but its role will be different because it will be based on a statistical model. Here, however, we shall continue with the descriptive interpretation and discuss the so-called principal component technique.

Our starting point consists of n observations on K variables, which will be arranged in an $n \times K$ matrix $\mathbf{X}$. An example is given in Table 1.1, which contains time series data on 17 components of total income and outlay in the

Inventory Revaluation Adjustment (x_9)	Net Rent Received by Individuals (x_{10})	Entrepreneurial Withdrawals (x_{11})	Dividends (x_{12})	Adjustment for Depreciation (x_{13})	Interest (x_{14})	Dividends and Interests from Abroad, etc. (x_{15})	Adjustment for Depreciation and Depletion (x_{16})	Foreign Balance (x_{17})
−88	490	1079	296	−751	398	−105	−41	51
−16	516	1134	374	−839	421	−95	−85	25
16	563	1194	368	−834	437	−96	−68	78
−35	546	1250	427	−848	458	−96	−60	36
170	514	1245	462	−918	470	−105	−55	8
78	508	1262	492	−917	494	−104	−55	42
6	494	1288	534	−941	527	−114	−46	63
70	492	1338	612	−1000	560	−127	−55	35
412	426	1277	577	−953	572	−109	−29	56
323	303	1121	434	−842	571	−61	−1	9
147	209	975	275	−726	552	−37	35	16
−227	211	902	225	−682	500	−45	36	22
−149	190	910	290	−718	485	−75	4	35
−72	214	952	373	−738	465	−110	−2	−27
−16	219	1012	486	−768	461	15	−18	−47
−64	258	1123	502	−850	469	−221	−57	−26
110	258	1106	353	−857	459	−175	−52	61

United States in the 17 years, 1922 to 1938.[13] Hence $n = K = 17$, but the equality of n and K should be regarded as accidental. The problem is: can we describe each of these K variables by a linear function of a small number of other variables with a high degree of accuracy? This would be trivially true if all variables moved proportionally; one single variable would then suffice to describe the behavior of all K variables. Let us start with one variable.

The First Principal Component

Our single variable takes n values, to be arranged in a column vector $\mathbf{p}$. At this stage $\mathbf{p}$ is not yet determined, but we proceed as if it were. If all variables behave proportionally, each column of $\mathbf{X}$ is equal to some scalar multiple of $\mathbf{p}$. This implies $\mathbf{X} = \mathbf{pa}'$, where $\mathbf{a}'$ is the K-element row vector consisting of these scalar multiples, one for each column of $\mathbf{X}$. Note that the product $\mathbf{pa}'$ remains unchanged when $\mathbf{p}$ is multiplied by some scalar $c \neq 0$ and $\mathbf{a}$ by $1/c$. By imposing

$$\mathbf{p}'\mathbf{p} = 1 \tag{9.1}$$

we shall be able to obtain uniqueness except for sign (i.e., $\mathbf{p}$ and $\mathbf{a}$ may still be replaced by $-\mathbf{p}$ and $-\mathbf{a}$, respectively).

Obviously, one should expect that $\mathbf{X} = \mathbf{pa}'$ will not hold exactly in general,

[13] The numerical example is based on R. Stone (1947). Tables 1.1 through 1.5 are derived from this article, but the figures reproduced here have fewer decimal places.

so that there will be a nonzero matrix of discrepancies, $\mathbf{X} - \mathbf{pa}'$, for whatever vectors $\mathbf{p}$ and $\mathbf{a}$. Our criterion is to select these vectors such that the sum of squares of all Kn discrepancies is minimized. It is easily verified that the sum of squares of all elements a_{ij} of an $m \times n$ matrix $\mathbf{A}$ can be written as the trace of $\mathbf{A}'\mathbf{A}$:

$$\operatorname{tr} \mathbf{A}'\mathbf{A} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2$$

Therefore our objective is to minimize

$$\begin{aligned}(9.2)\quad \operatorname{tr} (\mathbf{X} - \mathbf{pa}')'(\mathbf{X} - \mathbf{pa}') &= \operatorname{tr} \mathbf{X}'\mathbf{X} - \operatorname{tr} \mathbf{ap}'\mathbf{X} - \operatorname{tr} \mathbf{X}'\mathbf{pa}' + \operatorname{tr} \mathbf{ap}'\mathbf{pa}' \\ &= \operatorname{tr} \mathbf{X}'\mathbf{X} - 2\mathbf{p}'\mathbf{Xa} + \mathbf{a}'\mathbf{a}\end{aligned}$$

where use is made of $\operatorname{tr} \mathbf{ap}'\mathbf{X} = \operatorname{tr} \mathbf{p}'\mathbf{Xa} = \mathbf{p}'\mathbf{Xa}$ and similarly $\operatorname{tr} \mathbf{ap}'\mathbf{pa}' = \operatorname{tr} \mathbf{p}'\mathbf{pa}'\mathbf{a} = \mathbf{p}'\mathbf{pa}'\mathbf{a} = \mathbf{a}'\mathbf{a}$. We differentiate (9.2) with respect to $\mathbf{a}$ for given $\mathbf{p}$ and put the derivative equal to zero. This gives

$$(9.3)\qquad \mathbf{a} = \mathbf{X}'\mathbf{p}$$

which expresses the coefficient vector $\mathbf{a}$ in $\mathbf{p}$, whatever $\mathbf{p}$ may be. When substituting (9.3) into (9.2) we obtain $\operatorname{tr} \mathbf{X}'\mathbf{X} - \mathbf{p}'\mathbf{XX}'\mathbf{p}$, which shows that our next task is to maximize $\mathbf{p}'\mathbf{XX}'\mathbf{p}$ for variations in $\mathbf{p}$ subject to (9.1). So we form the Lagrangian expression $\mathbf{p}'\mathbf{XX}'\mathbf{p} - \lambda(\mathbf{p}'\mathbf{p} - 1)$ and differentiate it with respect to $\mathbf{p}$. This gives $2\mathbf{XX}'\mathbf{p} - 2\lambda\mathbf{p}$, so that the condition on $\mathbf{p}$ becomes

$$(9.4)\qquad (\mathbf{XX}' - \lambda\mathbf{I})\mathbf{p} = \mathbf{0}$$

Hence, $\mathbf{p}$ is a characteristic vector of the $n \times n$ positive semidefinite matrix $\mathbf{XX}'$ corresponding to root λ. To find out which root is to be taken we premultiply (9.4) by $\mathbf{p}'$, which gives $\mathbf{p}'\mathbf{XX}'\mathbf{p} = \lambda\mathbf{p}'\mathbf{p} = \lambda$. Since our objective is to maximize $\mathbf{p}'\mathbf{XX}'\mathbf{p}$, we should take the largest root of $\mathbf{XX}'$.[14] Furthermore, by premultiplying (9.4) by $\mathbf{X}'$ we obtain

$$(9.5)\qquad (\mathbf{X}'\mathbf{X} - \lambda\mathbf{I})\mathbf{X}'\mathbf{p} = (\mathbf{X}'\mathbf{X} - \lambda\mathbf{I})\mathbf{a} = \mathbf{0}$$

in view of (9.3). Hence the coefficient vector $\mathbf{a}$ is a characteristic vector of the matrix $\mathbf{X}'\mathbf{X}$ corresponding to the largest root except that it is *not* normalized such that it has unit length (see Problem 9.1). Note also that (9.4) and (9.3) imply $\lambda\mathbf{p} = \mathbf{X}(\mathbf{X}'\mathbf{p}) = \mathbf{Xa}$ and, hence,

$$(9.6)\qquad \mathbf{p} = \frac{1}{\lambda}\mathbf{Xa}$$

The vector $\mathbf{p}$ thus derived gives the best linear description of the $\mathbf{X}$ columns in the LS sense. It is known as the *first principal component* of the K variables represented in $\mathbf{X}$. The addition "first" will become clear immediately; it will induce us to add a subscript 1 to $\mathbf{p}$, $\mathbf{a}$, and λ of (9.4) to (9.6).

[14] See Problem 9.2 for an analysis of the conditional maximum along the lines of the last two paragraphs of Section 1.8.

Other Principal Components

The matrix $\mathbf{X}$ is now approximated by $\mathbf{p}_1\mathbf{a}_1'$ and, therefore, the discrepancy matrix is $\mathbf{X} - \mathbf{p}_1\mathbf{a}_1'$. We may then ask whether this matrix of residual elements can in turn be described by another matrix of unit rank, $\mathbf{p}_2\mathbf{a}_2'$, so that we obtain $\mathbf{p}_1\mathbf{a}_1' + \mathbf{p}_2\mathbf{a}_2'$ as a more accurate approximation to $\mathbf{X}$. This question will be answered under the conditions

$$\mathbf{p}_2'\mathbf{p}_2 = 1 \qquad \mathbf{p}_1'\mathbf{p}_2 = 0 \tag{9.7}$$

the first of which is analogous to (9.1), while the second requires that the two principal components be orthogonal.[15] The procedure is precisely the same as before except that we should replace $\mathbf{X}$ by $\mathbf{X} - \mathbf{p}_1\mathbf{a}_1'$. So we minimize

$$\begin{aligned}\operatorname{tr}(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1' - \mathbf{p}_2\mathbf{a}_2')'(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1' - \mathbf{p}_2\mathbf{a}_2')&\\ &= \operatorname{tr}(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1')'(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1') - 2\operatorname{tr}(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1')'\mathbf{p}_2\mathbf{a}_2' + \operatorname{tr}\mathbf{a}_2\mathbf{p}_2'\mathbf{p}_2\mathbf{a}_2'\\ &= \operatorname{tr}(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1')'(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1') - 2\mathbf{a}_2'\mathbf{X}'\mathbf{p}_2 + \mathbf{a}_2'\mathbf{a}_2\end{aligned}$$

where use has been made of (9.7) in the last step. Minimization with respect to $\mathbf{a}_2$ gives

$$\mathbf{a}_2 = \mathbf{X}'\mathbf{p}_2 \tag{9.8}$$

The function to be minimized for variations in $\mathbf{p}_2$ is then

$$\operatorname{tr}(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1')'(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1') - \mathbf{p}_2'\mathbf{X}\mathbf{X}'\mathbf{p}_2$$

so that $\mathbf{p}_2'\mathbf{X}\mathbf{X}'\mathbf{p}_2$ is to be maximized subject to (9.7). We construct the Lagrangian expression $\mathbf{p}_2'\mathbf{X}\mathbf{X}'\mathbf{p}_2 - \lambda_2(\mathbf{p}_2'\mathbf{p}_2 - 1) - \mu\mathbf{p}_1'\mathbf{p}_2$, differentiate it with respect to $\mathbf{p}_2$, and equate the result to zero:

$$2\mathbf{X}\mathbf{X}'\mathbf{p}_2 - 2\lambda_2\mathbf{p}_2 - \mu\mathbf{p}_1 = \mathbf{0} \tag{9.9}$$

We then premultiply by $\mathbf{p}_1'$, which gives $2\mathbf{p}_1'\mathbf{X}\mathbf{X}'\mathbf{p}_2 = \mu\mathbf{p}_1'\mathbf{p}_1 = \mu$. This implies $\mu = 0$ because $\mathbf{X}\mathbf{X}'\mathbf{p}_1 = \lambda_1\mathbf{p}_1$ [see (9.4)] and hence $\mathbf{p}_1'\mathbf{X}\mathbf{X}'\mathbf{p}_2 = \lambda_1\mathbf{p}_1'\mathbf{p}_2 = 0$. Therefore we can simplify (9.9) to

$$(\mathbf{X}\mathbf{X}' - \lambda_2\mathbf{I})\mathbf{p}_2 = \mathbf{0} \tag{9.10}$$

which shows that $\mathbf{p}_2$ is a characteristic vector of $\mathbf{X}\mathbf{X}'$ corresponding to root λ_2. This vector should be orthogonal to the characteristic vector $\mathbf{p}_1$ which

[15] It can be shown that precisely the same second principal component is obtained when this orthogonality condition is not imposed. We shall not give a detailed proof but shall confine ourselves to stating that the basic reason is that (1) the first principal component is orthogonal to the columns of the discrepancy matrix $\mathbf{X} - \mathbf{p}_1\mathbf{a}_1'$, because $(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1')'\mathbf{p}_1 = \mathbf{X}'\mathbf{p}_1 - \mathbf{a}_1 = \mathbf{0}$ follows from (9.3), and (2) if the second principal component is to give the best linear approximation of these columns in the LS sense, it must be orthogonal to any vector that is orthogonal to all of these columns.

Table 1.2

Sums of Squares and Products of Deviations from Means

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	Number of Variable
644.6	444.8	245.4	357.9	277.2	229.3	−109.8	105.6	65.7	94.2	129.4	101.8	−88.7	6.9	−25.5	−28.5	6.0	1
	328.5	195.8	239.4	174.2	139.3	−63.5	83.0	30.7	54.2	81.4	70.2	−57.2	0.4	−20.1	−19.9	1.2	2
		265.6	185.3	76.1	121.3	−56.4	72.2	−39.5	60.2	43.6	24.0	−23.5	−21.0	−11.2	−18.8	2.6	3
			232.4	159.2	163.9	−79.2	66.1	23.5	74.7	78.0	52.3	−48.1	−1.8	−10.7	−18.3	4.9	4
				138.6	116.8	−59.2	30.2	43.6	50.0	64.8	44.4	−42.5	8.2	−8.6	−11.1	5.8	5
					133.7	−69.8	31.9	21.0	64.5	59.1	29.4	−33.6	−1.4	−5.9	−12.8	7.4	6
						46.5	−15.6	−10.5	−35.8	−31.7	−12.0	17.7	1.2	6.7	7.1	−5.3	7
							45.5	− 2.8	14.2	15.6	17.2	−10.5	−4.3	−4.9	−6.4	−3.6	8
								41.3	7.3	18.0	14.0	−13.6	8.0	−0.4	−0.8	1.5	9
									33.7	26.8	10.5	−14.0	−2.3	−2.5	−6.4	4.5	10
										31.4	19.6	−19.7	2.5	−4.6	−5.9	3.4	11
											19.5	−14.3	3.8	−2.8	−3.2	−0.3	12
												13.7	−2.7	3.5	3.5	−1.7	13
													4.4	0.6	1.3	0.1	14
														4.4	1.7	−0.6	15
															2.1	−0.6	16
																1.9	17

Table 1.3

Correlation Coefficients Corresponding to Table 1.2

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	Number of Variable
1	.97	.59	.92	.93	.78	−.63	.62	.40	.64	.91	.91	−.94	.13	−.48	−.78	.17	1
	1	.66	.87	.82	.66	−.51	.68	.26	.52	.80	.88	−.85	.01	−.53	−.77	.05	2
		1	.75	.40	.64	−.51	.66	−.38	.64	.48	.33	−.39	−.61	−.33	−.80	.12	3
			1	.89	.93	−.76	.64	.24	.84	.91	.78	−.85	−.06	−.34	−.83	.24	4
				1	.86	−.74	.38	.58	.73	.98	.85	−.97	.33	−.35	−.66	.36	5
					1	−.88	.41	.28	.96	.91	.58	−.79	−.06	−.24	−.77	.47	6
						1	−.34	−.24	−.90	−.83	−.40	.70	.09	.47	.73	−.57	7
							1	−.07	.36	.41	.58	−.42	−.30	−.35	−.66	−.39	8
								1	.19	.50	.50	−.57	.60	−.03	−.09	.17	9
									1	.82	.41	−.65	−.19	−.21	−.77	.56	10
										1	.79	−.95	.21	−.40	−.73	.44	11
											1	−.87	.41	−.30	−.50	−.04	12
												1	−.35	.46	.66	−.34	13
													1	.14	.43	.03	14
														1	.55	−.22	15
															1	−.31	16
																1	17

corresponds to the largest root; at the same time, λ_2 should be as large as possible because the objective is to maximize $\mathbf{p}_2'\mathbf{X}\mathbf{X}'\mathbf{p}_2 = \lambda_2$. Hence the second principal component $\mathbf{p}_2$ is a characteristic vector corresponding to the second largest root λ_2. We assume that λ_1 and λ_2 are different; the case of multiple positive roots of $\mathbf{X}\mathbf{X}'$ is characterized by a lack of uniqueness of the principal components and will not be discussed here.

We can go on in this way by deriving r principal components, r being the rank of $\mathbf{X}\mathbf{X}'$ (and of $\mathbf{X}$). The ith such component minimizes the sum of squares of the discrepancies that are left after the earlier components $\mathbf{p}_1, \ldots, \mathbf{p}_{i-1}$ have done their work, and the minimization takes place subject to the unit-length constraint $\mathbf{p}_i'\mathbf{p}_i = 1$ and the orthogonality constraints $\mathbf{p}_1'\mathbf{p}_i = \cdots = \mathbf{p}_{i-1}'\mathbf{p}_i = 0$. The result is that $\mathbf{p}_i$ is a characteristic vector of $\mathbf{X}\mathbf{X}'$ corresponding to the ith largest root λ_i.

A Numerical Example

The principal components $\mathbf{p}_1, \mathbf{p}_2, \ldots$ can thus be determined by computing the matrix $\mathbf{X}\mathbf{X}'$ and then finding its characteristic roots and vectors. One may also take $\mathbf{X}'\mathbf{X}$ instead of $\mathbf{X}\mathbf{X}'$ and compute the characteristic vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots$, after which the $\mathbf{p}$'s are computed from (9.6). Following the latter procedure, we obtain for the data of Table 1.1 the matrix of sums of squares and products given in Table 1.2. (Note that all 17 variables have been measured as deviations from the means. The corresponding correlation coefficients are given in Table 1.3 for comparison purposes.) The left-hand part of Table 1.4 contains the elements of the first three principal components together with the roots $\lambda_1, \lambda_2, \lambda_3$. These roots may be used to measure the relative importance of the corresponding components. The argument is based on the criterion used: the sum of squares of all Kn discrepancies. These discrepancies are of the form $\mathbf{X} - \mathbf{p}_1\mathbf{a}_1' - \cdots - \mathbf{p}_i\mathbf{a}_i'$ after the use of the ith component. Consequently, *before* any component is used the discrepancies are the elements of $\mathbf{X}$, and their sum of squares is tr $\mathbf{X}'\mathbf{X}$. The first principal component reduces this sum of squares to [see (9.2) to (9.4)]:

$$(9.11) \qquad \begin{aligned} \operatorname{tr}(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1')'(\mathbf{X} - \mathbf{p}_1\mathbf{a}_1') &= \operatorname{tr}\mathbf{X}'\mathbf{X} - 2\mathbf{p}_1'\mathbf{X}\mathbf{a}_1 + \mathbf{a}_1'\mathbf{a}_1 \\ &= \operatorname{tr}\mathbf{X}'\mathbf{X} - \mathbf{p}_1'\mathbf{X}\mathbf{X}'\mathbf{p}_1 \\ &= \operatorname{tr}\mathbf{X}'\mathbf{X} - \lambda_1 \end{aligned}$$

It can be shown in the same way that the second principal component accounts for an additional reduction of the sum of squares of the discrepancies equal to λ_2, and so on. Thus, by dividing $\lambda_1, \lambda_2, \lambda_3$ by tr $\mathbf{X}'\mathbf{X}$ we obtain three ratios which can be regarded as measuring the degree to which the variation of the K variables is accounted for by the corresponding principal component. In this case the first component accounts for more than 80

Table 1.4

Three Principal Components and the Proportion of the Variance of Each Variable Accounted for by Each Component

				Number of	Proportion Accounted for by			
Year	$\mathbf{p}_1$	$\mathbf{p}_2$	$\mathbf{p}_3$	Variable	$\mathbf{p}_1$	$\mathbf{p}_2$	$\mathbf{p}_3$	Rest
1922	−.13	.33	.27	1	.97	.02	.01	.00
1923	.08	.28	.17	2	.91	.00	.09	.01
1924	.08	.14	.29	3	.50	.49	.00	.01
1925	.17	.21	.20	4	.95	.01	.03	.02
1926	.26	.07	.13	5	.83	.12	.04	.01
1927	.24	−.04	.16	6	.73	.01	.26	.01
1928	.27	−.03	.10	7	.50	.00	.34	.16
1929	.37	.09	−.08	8	.42	.10	.11	.37
1930	.14	−.52	−.00	9	.09	.72	.03	.16
1931	−.15	−.50	.13	10	.55	.03	.40	.02
1932	−.45	−.34	.21	11	.84	.06	.08	.02
1933	−.46	.16	11	12	.72	.14	.06	.08
1934	−.31	.16	− 18	13	.83	.13	.01	.03
1935	−.18	.14	− 30	14	.00	.75	.00	.25
1936	−.01	.12	−.47	15	.21	.00	.05	.74
1937	.12	.01	−.47	16	.71	.09	.01	.20
1938	−.02	−.09	−.27	17	.04	.01	.39	.56
λ_i	1605	211	121					
$\lambda_i/\text{tr}\,\mathbf{X}'\mathbf{X}$	.808	.106	.061					

percent of the total variation, the second for more than 10 percent, and the third for 6 percent. What remains for all other principal components is only 2½ percent.

Analysis of Individual Variables

It is also interesting to consider the contributions of the three principal components to the "explanation" of the behavior of each of the 17 variables separately. The argument is as follows. Let $\mathbf{x}_h$ be one of the K columns of $\mathbf{X}$ (the observations on the hth variable) and consider the linear relation

$$\mathbf{x}_h = b_{1h}\mathbf{p}_1 + b_{2h}\mathbf{p}_2 + b_{3h}\mathbf{p}_3 + \text{discrepancy vector} \tag{9.12}$$

where the b's are coefficients that are still to be specified. Let us specify them according to the LS principle. Applying (7.8) we find

$$\begin{bmatrix} b_{1h} \\ b_{2h} \\ b_{3h} \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1'\mathbf{p}_1 & \mathbf{p}_1'\mathbf{p}_2 & \mathbf{p}_1'\mathbf{p}_3 \\ \mathbf{p}_2'\mathbf{p}_1 & \mathbf{p}_2'\mathbf{p}_2 & \mathbf{p}_2'\mathbf{p}_3 \\ \mathbf{p}_3'\mathbf{p}_1 & \mathbf{p}_3'\mathbf{p}_2 & \mathbf{p}_3'\mathbf{p}_3 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{p}_1'\mathbf{x}_h \\ \mathbf{p}_2'\mathbf{x}_h \\ \mathbf{p}_3'\mathbf{x}_h \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1'\mathbf{x}_h \\ \mathbf{p}_2'\mathbf{x}_h \\ \mathbf{p}_3'\mathbf{x}_h \end{bmatrix}$$

where the second equality sign is based on the unit-length condition $\mathbf{p}_i'\mathbf{p}_i = 1$ and the orthogonality condition $\mathbf{p}_i'\mathbf{p}_j = 0$, $i \neq j$. Now $\mathbf{p}_1'\mathbf{x}_h$ (and hence b_{1h}) is the hth element of the vector $\mathbf{p}_1'\mathbf{X} = \mathbf{a}_1'$; see (9.3). Therefore we can write (9.12) in the following specific form:

$$\text{(9.13)} \qquad \mathbf{x}_h = a_{1h}\mathbf{p}_1 + a_{2h}\mathbf{p}_2 + a_{3h}\mathbf{p}_3 + \mathbf{v}_h$$

where a_{ih} is the hth element of the vector $\mathbf{a}_i$ and $\mathbf{v}_h$ is a discrepancy vector (whose length is minimized by the LS method).

If we premultiply each side of (9.13) by its own transpose we obtain

$$\begin{aligned}\mathbf{x}_h'\mathbf{x}_h &= a_{1h}^2\mathbf{p}_1'\mathbf{p}_1 + a_{2h}^2\mathbf{p}_2'\mathbf{p}_2 + a_{3h}^2\mathbf{p}_3'\mathbf{p}_3 + \mathbf{v}_h'\mathbf{v}_h \\ &\quad + 2(a_{1h}a_{2h}\mathbf{p}_1'\mathbf{p}_2 + \cdots + a_{1h}\mathbf{p}_1'\mathbf{v}_h + \cdots)\end{aligned}$$

The cross-product terms on the second line are all zero because of the orthogonality of the $\mathbf{p}$'s mutually and that of the $\mathbf{p}$'s and $\mathbf{v}_h$, $\mathbf{v}_h$ being an LS discrepancy vector which is orthogonal to the vectors of the regressors (the $\mathbf{p}$'s). Taking account of $\mathbf{p}_i'\mathbf{p}_i = 1$ we thus find

$$\text{(9.14)} \qquad \mathbf{x}_h'\mathbf{x}_h = a_{1h}^2 + a_{2h}^2 + a_{3h}^2 + \mathbf{v}_h'\mathbf{v}_h$$

The sum of squares of the values taken by the hth variable can thus be decomposed into parts attributable to the first principal component (a_{1h}^2), the second (a_{2h}^2), the third (a_{3h}^2), and the rest ($\mathbf{v}_h'\mathbf{v}_h$). The results for our numerical example (in the "relative" form $a_{ih}^2/\mathbf{x}_h'\mathbf{x}_h$, $\mathbf{v}_h'\mathbf{v}_h/\mathbf{x}_h'\mathbf{x}_h$) are given in the last four columns of Table 1.4. They indicate that some of the variables are much more closely related to the second principal component than to the first (take x_9 and x_{14}). By adding the contributions of any given component to all K variables we obtain the λ's which refer to all these variables jointly:

$$\text{(9.15)} \qquad \sum_{h=1}^{K} a_{ih}^2 = \mathbf{a}_i'\mathbf{a}_i = \mathbf{p}_i'\mathbf{X}\mathbf{X}'\mathbf{p}_i = \lambda_i\mathbf{p}_i'\mathbf{p}_i = \lambda_i$$

An Interpretation of the Three Principal Components

It is interesting to observe that the principal components of the present data allow a rather simple interpretation. We recall that the variables are all components of total income and total outlay of the United States. Consider then (1) total income, (2) the yearly change in total income, and (3) time. (The last variable is an ordinary linear trend which takes the value 1 in 1922, 2 in 1923, and so on.) The nine correlations of these three variables and the first three principal components are given in Table 1.5. It appears that the first component is very highly correlated with total income, that the second is highly correlated with the change in income, and that the third is rather highly correlated with time. The conclusion of the analysis is that the behavior

Table 1.5

Correlation Coefficients of Principal Components and Certain Economic Variables

Economic Variables	$\mathbf{p}_1$	$\mathbf{p}_2$	$\mathbf{p}_3$
Total income	.995	−.041	.057
Same, annual change	−.056	.948	−.124
Time	−.369	−.282	−.836

of the 17 variables during the period 1922–1938 can be described rather accurately by linear combinations of total income, of its change, and of a linear time trend.

However, the economic interpretation of principal components in general is no easy matter. In principle it is conceivable that there is a limited number of principal factors which dominate the behavior of economic variables, but there is no reason to assume that these factors satisfy the orthogonality condition $\mathbf{p}_i'\mathbf{p}_j = 0$. It just happened that in this example income, its change, and time have low correlations.

Principal Components Depend on the Origin and Scale of the Variables

The numerical results presented here are based on variables which are measured as deviations from the means. If the "natural" zeros are used instead of the means, one obtains different principal components. If the variables are standardized (measured as deviations from the means and subsequently divided by the standard deviations), the matrix $\mathbf{X}'\mathbf{X}$ becomes a matrix of correlation coefficients and the principal components are changed again. The latter procedure is often applied in psychology, where the variables frequently have no common unit of measurement; it was not applied here, since the variables are all in dollars per year. This dependence on the unit of measurement is obviously a weakness of the principal component technique. To show that there is such a dependence, it is sufficient to consider the case $K = 2$, so that $\mathbf{X}'\mathbf{X}$ is a 2×2 matrix, and to inspect the roots (5.3) of such a 2×2 matrix. It is also intuitively plausible. If a variable is measured in such small units that its numerical values dominate those of the other $K - 1$ variables, the first principal component will reflect the behavior of this particular variable rather closely; see Problem 9.3.

Problems

9.1 Let $\mathbf{X}$ be the matrix of observations on certain variables. Write $\mathbf{p}_i = (1/\lambda_i)\mathbf{X}\mathbf{a}_i$ for the ith principal component, where λ_i is a positive root of

$\mathbf{X}'\mathbf{X}$ (or of $\mathbf{XX}'$). Prove that the length of the weight vector $\mathbf{a}_i$ is $\sqrt{\lambda_i}$ if $\mathbf{p}_i$ has unit length.

9.2 To verify the second-order condition for the first principal component along the lines of the last two paragraphs of Section 1.8, recall that it is stated below eq. (9.3) that the problem is to maximize $\mathbf{p}_1'\mathbf{XX}'\mathbf{p}_1$ subject to $\mathbf{p}_1'\mathbf{p}_1 = 1$. (It is advantageous to use a subscript 1 here.) The Lagrangian function and the matrix of its second-order derivatives with respect to $\mathbf{p}_1$ and $\mathbf{p}_1'$ are then

$$F(\mathbf{p}_1, \lambda_1) = \mathbf{p}_1'\mathbf{XX}'\mathbf{p}_1 - \lambda_1(\mathbf{p}_1'\mathbf{p}_1 - 1) \tag{9.16}$$

$$\frac{\partial^2 F}{\partial \mathbf{p}_1\, \partial \mathbf{p}_1'} = 2(\mathbf{XX}' - \lambda_1\mathbf{I})$$

Prove that the second-order constrained maximum condition is

$$\mathbf{q}'(\mathbf{XX}' - \lambda_1\mathbf{I})\mathbf{q} < 0 \quad \text{for all} \quad \mathbf{q} \neq \mathbf{0} \quad \text{satisfying} \quad \mathbf{p}_1'\mathbf{q} = 0 \tag{9.17}$$

where $\mathbf{q} = [q_\alpha] = [dp_{\alpha 1}]$, $dp_{\alpha 1}$ being an infinitesimal change in the αth element of $\mathbf{p}_1$. Prove that we may specify $\mathbf{q}'\mathbf{q} = 1$ without real loss of generality. Also prove that

$$\mathbf{q}'\mathbf{XX}'\mathbf{q} = \mathbf{q}'\left(\sum_{i=1}^{r} \lambda_i \mathbf{p}_i \mathbf{p}_i'\right)\mathbf{q} = \sum_{i=2}^{r} \lambda_i (\mathbf{p}_i'\mathbf{q})^2$$

where λ_i and $\mathbf{p}_i$ are the ith root and a corresponding characteristic vector, respectively, of $\mathbf{XX}'$ and r is the rank of this matrix. Conclude, using $\mathbf{q}'\mathbf{q} = 1$,

$$\mathbf{q}'(\mathbf{XX}' - \lambda_1\mathbf{I})\mathbf{q} = \sum_{i=2}^{r} \lambda_i(\mathbf{p}_i'\mathbf{q})^2 - \lambda_1 \tag{9.18}$$

Finally, prove that $(\mathbf{p}_2'\mathbf{q})^2 + \cdots + (\mathbf{p}_r'\mathbf{q})^2 \leq 1$, and use this to prove that the second-order condition is satisfied when the largest root λ_1 of $\mathbf{XX}'$ is a simple root. [Hint for $\sum (\mathbf{p}_i'\mathbf{q})^2 \leq 1$: run an LS regression of $\mathbf{q}$ on $\mathbf{p}_2, \ldots, \mathbf{p}_r$ along the lines of eqs. (9.12) to (9.14).]

9.3 Consider the matrix $\mathbf{X}'\mathbf{X}$ of (9.5) and suppose that the first variable is measured in a different unit such that $x_{\alpha 1}$ becomes $cx_{\alpha 1}$. Prove that, if c is sufficiently large, the new matrix $\mathbf{X}'\mathbf{X}$ satisfies approximately $(1/c)^2\mathbf{X}'\mathbf{X} \approx (\sum x_{\alpha 1}^2)\mathbf{i}_1\mathbf{i}_1'$, where $\mathbf{i}_1$ is the first column of the $K \times K$ unit matrix. Next prove the following statements on the basis of this approximation: the new $\mathbf{X}'\mathbf{X}$ has unit rank, its (only) positive root is $c^2 \sum x_{\alpha 1}^2$, and $\mathbf{i}_1$ is a corresponding characteristic vector. Finally, conclude $\mathbf{a}_1 \approx \mathbf{i}_1$ and $\mathbf{p}_1' \approx [cx_{11} \cdots cx_{n1}]$, in both cases apart from a normalization factor.