

Determining an Adequate Number of Principal Components

Stanley L. Sclove

Abstract

The problem of choosing the number of PCs to retain is analyzed in the context of *model selection*, using so-called *model selection criteria* (MSCs). For a prespecified set of models, indexed by $k = 1, 2, \dots, K$, these model selection criteria (MSCs) take the form $MSC_k = nLL_k + a_n m_k$, where, for model k , LL_k is the maximum log likelihood, m_k is the number of independent parameters, and the constant a_n is $a_n = \ln n$ for BIC and $a_n = 2$ for AIC. The maximum log likelihood LL_k is achieved by using the maximum likelihood estimates (MLEs) of the parameters. In Gaussian models, LL_k involves the logarithm of the mean squared error (MSE). The main contribution of this chapter is to show how to best use BIC to choose the number of PCs, and to compare these results to *ad hoc* procedures that have been used. Findings include the following. These are stated as they apply to the eigenvalues of the correlation matrix, which are between 0 and p and have an average of 1. For considering an additional PC_{k+1} , with AIC, inclusion of the additional PC_{k+1} is justified if the corresponding eigenvalue λ_{k+1} is greater than $\exp(-2/n)$. For BIC, the inclusion of an additional PC_{k+1} is justified if $\lambda_{k+1} > n^{1/n}$, which tends to 1 for large n . Therefore, this is in approximate agreement with the average eigenvalue rule for correlation matrices, stating that one should retain dimensions with eigenvalues larger than 1.

Keywords: reduction of dimensionality, principal components, model selection criteria, information criteria, AIC, BIC

1. Introduction and background

1.1 Introduction

Sometimes, researchers know how many principal components (PCs) they need. For example, to construct an optimal scatterplot, the scores of the sample on the first two principal components will be used to obtain an optimal plot. For an optimal three-dimensional scatterplot, the scores on the first three principal components will be used. In many applications, however, the researchers will question how many principal components they need. This chapter discusses the application of various methods to the problem of reduction of dimensionality, in the sense of choosing an adequate

number of principal components to retain to represent a dataset. The methods discussed include *ad hoc* methods, likelihood-based methods, and model selection criteria (MSCs), especially Akaike's information criterion (AIC) and Bayesian information criterion (BIC). This chapter applies the concepts of [1, 2] to this particular problem.

1.2 Background

To begin the discussion here, we first give a short review of some general background on the relevant portions of multivariate statistical analysis, which may be obtained from textbooks such as [3] or [4].

1.3 Sample quantities

Let $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ denote a sample of n p -dimensional random vectors

$$\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{pi})', \quad i = 1, 2, \dots, n. \quad (1)$$

Here, the transpose ($'$) means that the vectors are being considered as column vectors. The sample *mean vector* is

$$\bar{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i / n. \quad (2)$$

The $p \times p$ sample covariance matrix is denoted by

$$\mathbf{S} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' / (n - 1). \quad (3)$$

1.4 Population quantities

The sample covariance matrix \mathbf{S} estimates the true covariance matrix Σ of the random variables

$$X_1, X_2, \dots, X_p.$$

The true covariance matrix is

$$\Sigma = [\sigma_{u,v}]_{u,v=1,2,\dots,p}, \quad (4)$$

where

$$\sigma_{uv} = C[X_u, X_v], \quad (5)$$

the covariance of X_u and X_v , for $u \neq v$, $u, v = 1, 2, \dots, p$. For $u = v$, we have $C[X_v, X_v] = \mathcal{V}[X_v]$, the variance of X_v .

1.5 Principal components

The principal components of Σ are defined as *uncorrelated linear combinations of maximal variance*. Let us elaborate on this brief definition. First, a linear combination, say LC , of the p variables can be expressed as the vector product $\mathbf{a}'\mathbf{x}$ of two vectors \mathbf{a} and \mathbf{x} , that is,

$$LC = \mathbf{a}'\mathbf{x} = a_1x_1 + a_2x_2 + \dots + a_px_p. \quad (6)$$

Here, the vector \mathbf{a} is a vector of scalars a_1, a_2, \dots, a_p :

$$\mathbf{a}' = (a_1 \ a_2 \ \dots \ a_p). \quad (7)$$

These a_j are the coefficients in the linear combination. Such linear combinations are called *variates*. Principal components are also called *latent variables*.

The variance \mathcal{V} of a linear combination LC is

$$\mathcal{V}[LC] = \mathcal{V}[\mathbf{a}'\mathbf{X}] = \mathbf{a}'\Sigma\mathbf{a}. \quad (8)$$

This is estimated as $\mathbf{a}'\mathbf{S}\mathbf{a}$. This is to be maximized over \mathbf{a} . The derivative with respect to the vector \mathbf{a} is

$$\partial\mathbf{a}'\mathbf{S}\mathbf{a}/\partial\mathbf{a} = 2\mathbf{S}\mathbf{a}. \quad (9)$$

The solution is not unique: If \mathbf{a} is a solution to this set of equations, so is $c\mathbf{a}$, where c is any scalar constant. Therefore, a constraint is required to obtain a meaningful solution. A reasonable such constraint is the condition $\mathbf{a}'\mathbf{a} = 1$, that is, the squared length of the vector \mathbf{a} equals 1. This is of course equivalent to the length of \mathbf{a} , the quantity $\sqrt{\mathbf{a}'\mathbf{a}}$, being equal to 1.

A function incorporating the constraint, the *Lagrangian function*, is

$$L(\mathbf{S}, \mathbf{a}; \lambda) = \mathbf{a}'\mathbf{S}\mathbf{a} + \lambda(1 - \mathbf{a}'\mathbf{a}). \quad (10)$$

The partial derivatives of the function L with respect to \mathbf{a} and λ are

$$\partial L/\partial\mathbf{a} = 2\mathbf{S}\mathbf{a} - 2\lambda\mathbf{a} \quad (11)$$

and

$$\partial L/\partial\lambda = \partial\lambda(1 - \mathbf{a}'\mathbf{a})/\partial\lambda = 1 - \mathbf{a}'\mathbf{a}. \quad (12)$$

Setting these partial derivatives equal to zero gives the simultaneous linear equations

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a}, \quad (13)$$

and the equation

$$\mathbf{a}'\mathbf{a} = 1. \quad (14)$$

The simultaneous linear equations can be written as

$$\mathbf{S}\mathbf{a} - \lambda\mathbf{a} = 0, \quad (15)$$

where $\mathbf{0}$ is the zero vector, the vector whose elements are all zeroes. Factoring out \mathbf{a} on the right, we obtain

$$(\mathbf{S} - \lambda\mathbf{I})\mathbf{a} = 0. \quad (16)$$

For nontrivial solutions, the determinant of the coefficient matrix $\mathbf{S} - \lambda\mathbf{I}$ must be zero, that is, we must have $\det(\mathbf{S} - \lambda\mathbf{I}) = 0$. This condition is a polynomial equation of degree p in λ . Denote the p roots by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. These roots are the *eigenvalues* (also called *latent values*). Their sum is the trace of \mathbf{S} ; their product is the determinant of \mathbf{S} .

The corresponding Eigen equations are

$$\mathbf{S}\mathbf{a}_j = \lambda_j\mathbf{a}_j, \quad j = 1, 2, \dots, p. \quad (17)$$

1.5.1 Values of PCs in terms of Xs

The j th principal component (PC), C_j , is the linear combination of the form

$$C_j = \mathbf{a}'_j\mathbf{x} = a_{1j}x_1 + a_{2j}x_2 + \dots + a_{pj}x_p, \quad (18)$$

where $\mathbf{a}'_j = (a_{1j}, a_{2j}, \dots, a_{pj})$. That is to say, for $j = 1, 2, \dots, p$, the value of the j th PC for Individual i is $c_{ji} = \mathbf{a}'_j\mathbf{x}_i$, $i = 1, 2, \dots, n..$

The equation for the j th PC in terms of the vector $\mathbf{x} = (x_1x_2 \dots x_p)'$ is $c_j = \mathbf{a}'_j\mathbf{x}$, $j = 1, 2, \dots, p$. Let \mathbf{c} be the p -vector of values of the p PCs. Then, $\mathbf{c} = \mathbf{A}'\mathbf{x}$, where $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_p]$ is the $p \times p$ matrix whose columns are the eigenvectors.

1.5.2 Values of Xs in terms of PCs

The inverse relation is

$$\mathbf{x} = \mathbf{A}'^{-1}\mathbf{c} = \mathbf{B}\mathbf{c}, \quad (19)$$

where

$$\mathbf{B} = \mathbf{A}'^{-1}, \quad (20)$$

where \mathbf{B} is the matrix of *loadings* of the X_v on the PCs C_j . Actually, \mathbf{A} is an orthonormal matrix (meaning that its columns are of length one and are pairwise orthogonal), so $\mathbf{A}^{-1} = \mathbf{A}'$. Thus, $\mathbf{B} = \mathbf{A}$. Therefore,

$$\mathbf{x} = \mathbf{A}'^{-1}\mathbf{c} = \mathbf{A}\mathbf{c}. \quad (21)$$

Letting $\mathbf{a}^{(v) \prime}$ be the v th row of the matrix \mathbf{A} , that is,

$$\mathbf{a}^{(v) \prime} = (a_{v1}, a_{v2}, \dots, a_{vp}), \quad (22)$$

we have, for $v = 1, 2, \dots, p$, the representation of each variable X_v in terms of the variables C_1, C_2, \dots, C_p that are the principal components,

$$X_v = a_{v1}C_1 + a_{v2}C_2 + \dots + a_{vp}C_p. \quad (23)$$

In terms of the first k PCs, this is

$$X_v = a_{v1}C_1 + a_{v2}C_2 + \dots + a_{vk}C_k + \varepsilon_v, \quad (*) \quad (24)$$

where the error ε_v is

$$\varepsilon_v = a_{vk+1}C_{k+1} + a_{vk+2}C_{k+2} + \dots + a_{vp}C_p. \quad (25)$$

The covariance matrix can be represented in terms of its *principal idempotents* $\mathbf{a}_j\mathbf{a}'_j$ as

$$\mathbf{S} = \sum_{j=1}^p \lambda_j \mathbf{a}_j \mathbf{a}'_j. \quad (26)$$

It follows as a result of this representation that the best approximation of rank k to \mathbf{S} is the eigenvalue weighted sum of the first k principal idempotents,

$$\mathbf{S}^{(k)} = \sum_{j=1}^k \lambda_j \mathbf{a}_j \mathbf{a}'_j. \quad (27)$$

The weights are all non-negative, recalling that, for a symmetric matrix, such as a covariance matrix, the eigenvalues are non-negative.

2. Some *ad hoc* arithmetic procedures for determining an appropriate number of PCs

2.1 Procedure based on the average of the eigenvalues

The mean $\bar{\lambda}$ of the eigenvalues is the sum over the number

$$\bar{\lambda} = \sum_{j=1}^p \lambda_j / p. \quad (28)$$

The sum of the eigenvalues turns out to be equal to the *trace* of the covariance matrix; therefore, the mean eigenvalue is equal to the trace divided by p .

One procedure for deciding on the number of PCs to retain is to retain those for which the eigenvalues are greater than average, that is, greater than $\bar{\lambda}$. When working in terms of the correlation matrix, this average value is 1. To see this, recall that the correlation matrix is a special case of the covariance matrix, namely, the correlation matrix is the covariance matrix of the standardized variables. It is often preferable to work in terms of the correlation matrix rather than the covariance matrix, to control the effects of different units of measurement and different variances. If a variable has high variance relative to the other variables, the PC will be pulled in the direction of the variable with large variance.

When \mathbf{S} is taken to be the sample *correlation* matrix, the trace of the matrix is simply p , and therefore, the mean $\bar{\lambda}$ of the eigenvalues is 1.

2.2 An *ad hoc* arithmetic procedure based on retaining a prescribed proportion of the total variance

Another *ad hoc* procedure is to retain a number of PCs sufficient to account for a prescribed proportion, say, 90% of the total variance, that total variance being trace $\mathbf{S} = \sum_{j=1}^p \lambda_j$. The Figure 90% is of course somewhat arbitrary, so it might be good to have some somewhat more objective criteria based on the pattern of the eigenvalues.

2.3 Procedure based on the decrease of the eigenvalues

Another procedure—a graphical procedure—is to plot $\lambda_1, \lambda_2, \dots, \lambda_p$ against $1, 2, \dots, p$. The λ s are in decreasing order, so one then looks for a dropoff—an elbow—in the curve and retains a number of PCs corresponding to the point before the leveling off of the curve, if it does indeed take an elbow shape. Such a plot, of the eigenvalues versus $1, 2, \dots, p$, is called a *scree* plot, “scree” being the debris at the foot of a glacier (or, more generally, a collection of broken rock fragments at the base of crags, mountain cliffs, volcanoes, or valley shoulders).

3. Model selection criteria AIC and BIC for the number of PCs

Let us now delve a bit further into mathematical statistics and consider some more objective, numerical criteria, in particular, the information criteria AIC and BIC. Let us see what a Gaussian model would imply about AIC and BIC. The maximum log likelihood for the model (*) approximating the p variables in terms of k PCs is

$(2\pi)^{-np/2} |\hat{\Sigma}_k|^{-n/2} C(n, p, k)$, where $C(n, p, k)$ is a constant depending upon the sample size, n , the number of variables, p , and k , the Model k being considered, $k = 1, 2, \dots, K$, and $|\Sigma_k|$ denotes the determinant of the residual covariance matrix Σ_k .

The determinant of the covariance matrix is the product of the eigenvalues,

$$|\Sigma| = \prod_{j=1}^p \lambda_j. \quad (29)$$

For a model based on the first k PCs, the determinant of the residual covariance matrix is the product of the remaining, smaller eigenvalues, $\prod_{j=k+1}^p \lambda_j$.

The model selection criterion AIC—Akaike’s information criterion [5–7]—is based on an estimate of the logarithm of the cross-entropy of the K proposed models with a null model. That is, for alternative models indexed by $k = 1, 2, \dots, K$, AIC_k is an estimate of the log cross-entropy of the proposed Model k with the null model. The cross-entropy of the distribution with the probability density function $q(\mathbf{x})$ relative to a distribution with the probability density function $p(\mathbf{x})$ is defined as $H(p, q) = -\mathcal{E}_p[\ln q(\mathbf{X})] = -\int \ln q(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$.

The Bayesian information criterion (BIC) [8] is based on a large-sample estimate of the posterior probability pp_k of Model k , $k = 1, 2, \dots, K$. More precisely, BIC_k is an approximation to $-2 \ln pp_k$.

Formulated in this way, these model selection criteria (MSCs) are, thus, smaller-is-better criteria and take the form

$$MSC_k = -2 \ln \max L_k + a_n m_k, \quad k = 1, 2, \dots, K, \quad (30)$$

where L_k is the likelihood for Model k , $a_n = \ln n$ for BIC_k , $a_n = 2$ (not depending upon n) for AIC_k , and m_k is the number of independent parameters in Model k . The first term is a lack-of-fit (LOF) term, and the second term is a penalty term based on the number of parameters used. With AIC, the penalty is two units per parameter; with BIC, the penalty is $\ln n$ units per parameter. For $n \geq 8$, $\ln n$ exceeds 2: for sample sizes greater than 7, the penalty per parameter with BIC exceeds that for AIC. Therefore, relative to AIC, BIC tends to favor more parsimonious models—models with a smaller number of parameters.

Note that

$$pp_k \approx C \exp(-BIC_k/2), \quad (31)$$

where C is a constant. Thus, BIC values can be converted to values on a scale of 0–1. This is done by exponentiating $-BIC_k/2$, summing the values, and dividing by the sum. That is,

$$pp_k \approx \exp(-BIC_k/2) / \sum_{j=1}^K \exp(-BIC_j/2). \quad (32)$$

To relate the maximum likelihood to the eigenvalues, note that for the PC model,

$$-2 \ln \max L_k = n \ln \prod_{j=k+1}^p \lambda_j = n \sum_{j=k+1}^p \ln \lambda_j. \quad (33)$$

The model selection criteria can be written as

$$MSC_k = \text{Deviance}_k + \text{Penalty}_k, \quad (34)$$

where $\text{Deviance}_k = n \ln \max L_k$ is a measure of lack of fit and $\text{Penalty}_k = a_n m_k$. Inclusion of an additional PC is justified if the criterion value decreases, that is, if $MSC_{k+1} < MSC_k$. For PCs, this is

$$n \sum_{j=k+2}^p \ln \lambda_j + (k+1)a_n < n \sum_{j=k+1}^p \ln \lambda_j + k a_n. \quad (35)$$

This is

$$a_n < n \ln \lambda_{k+1} = \ln(\lambda_{k+1}^n), \quad (36)$$

or

$$\exp[a_n] < \lambda_{k+1}^n, \quad (37)$$

or

$$\lambda_{k+1} > \exp [a_n/n] \tag{38}$$

or

$$\lambda_{k+1} > \exp [-a_n/n]. \tag{39}$$

Thus, for AIC, the inclusion of the additional PC_{k+1} is justified if λ_{k+1} is greater than $\exp (-2/n)$.

For BIC, the inclusion of an additional PC_{k+1} is justified if

$$\lambda_{k+1} > \exp (\ln n/n) = [\exp (\ln n)]^{1/n} = n^{1/n}. \tag{40}$$

The quantity $n^{1/n}$ tends to 1 for large n . Therefore, this procedure is in approximate agreement with the average eigenvalue rule for correlation matrices, stating that one should retain dimensions with eigenvalues larger than 1.

4. Examples

4.1 An artificial example

The synthesis/analysis paradigm can be useful for understanding a problem. This means synthesizing (simulating) a dataset, so that you know the model and parameter values, and then applying your analysis method to see how well it performs. In the present context, it is interesting to simulate a dataset of measurements of rectangles, with variables length (L) and width (W) and also some functions of those such as perimeter = 2 L + 2 W and difference = L–W. In one synthesis, we took L to be Normal with a mean of 10 and a variance of 1, W was Normal with a mean of 10 and a variance of 1, PERI = 2 L + 2 W plus N(0,1) error, and DIFF = L–W plus N(0,1) error. The eigensystem was computed, and as expected, it is noted that there are two large eigenvalues, with subsequent ones dropping off a lot in value and being close to zero. The eigenvalues of the correlation matrix were 1.91, 1.83, 0.21, and 0.05.

4.2 A real example

Next, we consider the principal component analysis of a sample from the Los Angeles (LA) Heart Study. This was a long-term study, 1947–1972. It was a study among Civil Servants of Los Angeles county. LA civil servants, 2252, randomly selected, ages 21–70, received a battery of examinations for “routine” cardiovascular disease (CVD) risk factors.

The variables include age, systolic blood pressure (SYS), diastolic blood pressure (DIAS), weight (WT), height (HT), and coronary incident, a binary variable indicating whether the individual had a coronary incident during the course of the study. Blood pressure is reported as a bivariate variable, (SYS, DIAS). SYS is the pressure when the heart pumps, and DIAS is the pressure when the heart relaxes.

In the textbook [9], data for a sample of $n = 100$ men were studied. (Data on the same variables for another sample of 100 men are also given in [9]. Results can be compared and contrasted between the two samples.) Although, of course, the emphasis in the Heart Study was on explaining and predicting the coronary incident variable, here, we focus on the first five variables, their representation in terms of a smaller

number of PCs, and the interpretations of the PCs. we did the PC analysis; it was not in the LA Heart Study or the textbook.

We used Minitab statistical software for the analysis. Aspects of the analysis are shown as follows.

The lower-triangular portion of the correlation matrix for the five variables is shown in **Table 1**. The highest correlation is 0.835, between SYS and DIAS. The next highest correlation, 0.426, is between HT and WT.

4.3 Principal component analysis in the example

Note that an eigenvector can be multiplied by -1 , changing the signs of all its elements. In the following, this is done with PC1 so that SYS and DIAS have positive loadings. Our interpretations, related to the scientific/medical context of the study, are BPtotal, SIZE, AGE, OVERWT, and BPdiff and are written below the eigenvectors. The interpretations are based on which loadings are large and which are small, that is, on the relative sizes of the loadings. Taking 0.6 as a cutoff point, in PC1, SYS and DIAS have loadings above this, while the other variables have loadings less than this (in fact, less than 0.4), so PC1 can be interpreted as an index of total BP. In PC2, the variables WT and HT have large loadings with the same sign, so PC2 can be interpreted as SIZE (**Tables 2 and 3**).

	AGE	SYS	DIAS	WT	
SYS	0.342				
DIAS	0.354	0.835			<= NOTE highest r of 0.835 is btw SYS and DIAS
WT	-0.009	0.261	0.308		
HT	-0.332	-0.088	-0.099	0.426	<= NOTE next highest r of 0.426 is btw HT and WT

Correlations: AGE, SYS, DIAS, WT, HT.
Cell Contents: Pearson correlation.

Table 1.
 Correlation matrix of five variables—LA heart data.

Eigenanalysis of the correlation matrix					
Eigenvalue	2.1894	1.5382	0.6617	0.4485	0.1621
Proportion	0.438	0.308	0.132	0.090	0.032
Cumulative	0.438	0.746	0.878	0.968	1.000
Variable	PC1	PC2	PC3	PC4	PC5
AGE	-0.394	-0.365	0.800	-0.269	0.005
SYS	-0.615	0.050	-0.342	-0.174	0.687
DIAS	-0.624	0.063	-0.291	-0.049	-0.721
WT	-0.252	0.616	0.373	0.642	0.078
HT	0.117	0.694	0.141	-0.695	-0.051

Principal component analysis: AGE, SYS, DIAS, WT, HT.

Table 2.
 PCs of heart data.

Variable	PC1	PC2	PC3	PC4	PC5
AGE	0.394	-0.365	0.800	-0.269	0.005
SYS	0.615	0.050	-0.342	-0.174	0.687
DIAS	0.624	0.063	-0.291	-0.049	-0.721
WT	0.252	0.616	0.373	0.642	0.078
HT	-0.117	0.694	0.141	-0.695	-0.051
Interpretations (edited in to the computer output):					
	BPtotal	SIZE	AGEindex	OVERWT	BPdiff

Table 3.
PC1 is multiplied by -1.

As above, denote the eigensystem in terms of the eigenpairs

$$(\lambda_v, \mathbf{a}_v), \quad v = 1, 2, \dots, p. \tag{41}$$

Then, the eigensystem equations are

$$\mathbf{S} \mathbf{a}_v = \lambda_v \mathbf{a}_v, \quad v = 1, 2, \dots, p. \tag{42}$$

Here, \mathbf{S} is taken to be the correlation matrix. Let $\mathbf{1}'_v = (0 \ 0 \dots \ 1 \dots \ 0 \dots)$, the vector with 1 in the v th position and zeroes elsewhere. The covariance between a variable X_v and a PC C_u is $C[X_v, C_u] = C[\mathbf{1}'_v \mathbf{X}, \mathbf{a}'_u \mathbf{X}] = \mathbf{1}'_v \Sigma \mathbf{a}_u = \mathbf{1}'_v \lambda_u \mathbf{a}_u = \lambda_u a_{uv}$, where a_{uv} is the v th element of the vector \mathbf{a}_u . The coefficient of correlation is $\text{Corr}[X_v, C_u] = C[X_v, C_u] / \text{SD}[X_v] \text{SD}[C_u] = \lambda_u a_{uv} / \sigma_v \sqrt{\lambda_u} = \sqrt{\lambda_u} a_{uv} / \sigma_v$. When the covariance matrix used is the correlation matrix, each standard deviation $\sigma_v = 1$, and therefore, this correlation is $\sqrt{\lambda_u} a_{uv}$. A correlation of size greater than 0.6 corresponds to more than $0.6^2 \times 100\% = 36\%$ of variance explained. The variable X_v has a correlation higher than 0.6 with the component C_u if its loading in C_u , the value a_{uv} , is greater than $0.6 / \sqrt{\lambda_u}$. These values are appended to **Table 4**. Loadings larger than

Variable	PC1	PC2	PC3	PC4	PC5
AGE	0.394	-0.365	0.800	-0.269	0.005
SYS	0.615	0.050	-0.342	-0.174	0.687
DIAS	0.624	0.063	-0.291	-0.049	-0.721
WT	0.252	0.616	0.373	0.642	0.078
HT	-0.117	0.694	0.141	-0.695	-0.051
Eigenvalue, λ	2.1894	1.5382	0.6617	0.4485	0.1621
Square root, $\sqrt{\lambda}$	1.48	1.24	0.81	0.67	0.40
$0.6/\sqrt{\lambda}$	0.40	0.48	0.74	0.90	1.50
Interpretations	BPtotal	SIZE	AGE	OVERWT	BPdiff

Table 4.
Loadings corresponding to correlations > 0.6 are boldface.

No. of PCs, k	λ_k	$\lambda_k > 1?$	$\ln \lambda_k$	$N \ln \lambda_k$	for BIC: $N \ln \lambda_k > -4.61?$	for AIC: $N \ln \lambda_k > -2?$
1	2.19	Yes	0.78	78.36	Yes	Yes
2	1.54	Yes	0.43	43.06	Yes	Yes
3	0.66	No	-0.41	-41.29	No	No
4	0.45	No	-0.80	-80.18	No	No
5	0.16	No	-1.82	-181.95	No	No

Table 5.
Estimating the number of PCs by various methods.

this cutoff value are in boldface. (The cutoff point of 0.6 is somewhat arbitrary; one might use, for example, a cutoff of 0.5.)

One can also focus on the pattern of loadings within the different PCs for the interpretation of the PCs. To reiterate this process and the interpretations, we have the following:

PC1: SYS and DIAS have large loadings with the same sign; we interpret PC1 as BPindex, or BPtotal.

PC2: WT and HT have large loadings with the same sign; we interpret PC2 as the man's SIZE.

PC3: Only AGE has a large loading, so we interpret PC3 simply as AGE.

PC4: WT and HT have large loadings with opposite signs; we interpret PC4 as OVERWEIGHT.

PC5: SYS and DIAS have large loadings with opposite signs; we interpret PC5 as BPdrop.

We continue to marvel at how readily interpretable the PCs are. This simplicity is attained even without using a factor analysis model and using rotation to simplify the pattern of the loadings.

4.4 Employing the criteria in the example

To compare and contrast the methods, **Table 5** shows the eigenvalues and the results according to the various criteria for deciding on the adequate number of PCs. According to the rule based on the average eigenvalue, the dimension is retained if its eigenvalue is greater than 1 (when working in terms of the correlation matrix). For BIC, the k th PC is retained if

$$n \ln \lambda_k > -a_n, \tag{43}$$

where $a_n = \ln n$. Here, $n = 100$ and $\ln n = \ln 100$, approximately 4.61. For AIC, the k th PC is retained if $n \ln \lambda_k > -2$. In this example, the methods agree on retaining $k = 2$ PCs.

We feel that we should remark that, though it is the case that two PCs are suggested, the fourth and fifth PCs do have simple and interesting interpretations. It is just that they do not improve the fit very much. The third PC is essentially a single variable, age.

5. Discussion

The focus here has been on determining the number of dimensions needed to represent a complex of variables adequately. The algebraic solution devolves upon the

analysis of properties of the covariance matrix of the variables, especially through its eigensystem.

5.1 Regression on principal components

Next, we consider applying principal component analysis in the context of *multiple regression*. In this context, there is, of course, a response variable Y and explanatory variables X_1, X_2, \dots, X_p . One may transform the X s to their principal components, as this may aid in the interpretation of the results of the regression. In addition, the number of significant regression coefficients may be decreased. In such *regression on principal components* (see, e.g., [10]), however, one should not necessarily eliminate the principal components with small eigenvalues, as they may still be strongly related to the response variable.

The value of the Bayesian information criterion for Model k is

$$\text{BIC}_k = -2LL_k + m_k \ln n, \quad (44)$$

for alternative models indexed by $k = 1, 2, \dots, K$, where LL_k is the maximum log likelihood for Model k , that is, $LL_k = \max \ln L_k$ and m_k is the number of independent parameters in Model k . For linear regression models with Gaussian-distributed errors, $-2LL_k = \text{Const.} + n \ln \text{MSE}_k$ and so BIC takes the form

$$\text{BIC}_k = n \ln \text{MSE}_k + m_k \ln n, \quad (45)$$

where here MSE_k is the maximum likelihood estimate (MLE) of the mean squared error (MSE) of Model k , with divisor n , of the error variance.

The total number of subsets of p things is 2^p . Therefore, with p explanatory variables, there are 2^p alternative models—“subset regressions”—(including the model where no explanatory variables are used and the fitted value of Y is simply \bar{y}). For example, if there are three X s, the eight subsets are X_1 alone, X_2 alone, X_3 alone, (X_1, X_2) , (X_1, X_3) , (X_2, X_3) , (X_1, X_2, X_3) , and the empty set. It would usually seem to be expedient to evaluate all 2^p regression models—regressions on all 2^p subsets of principal components, using adjusted R-square, AIC, and/or BIC rather than reducing the number of models considered by regressing on only a few principal components. That is, in the context of regression on principal components, it is probably wise *not* to reduce the number of principal components, for, as stated above, it is conceivable that some principal components with small eigenvalues may nevertheless be important in explaining and predicting the response variable.

5.2 Some related recent literature

Other researchers have considered the problem of the choice of the number of principal components. For example, Bai et al. [11] examined the asymptotic consistency of AIC and BIC for determining the number of significant principal components in high-dimensional problems. The focus in this chapter has not necessarily been on high-dimensional problems.

Some various applications from recent literature involving choosing the number of principal components include the following. The method presented here could possibly be applied in these applications.

For example, a good book on the topic of model selection and testing, covering many aspects, is [12]. In recent years, various econometricians have examined the problems of diagnostic testing, specification testing, semiparametric estimation, and model selection. In addition, various researchers have considered whether to use model testing and model selection procedures to decide upon the models that best fit a particular dataset. This book explores both the issues with application to various regression models, including models for arbitrage pricing theory. Along the lines of model selection criteria, the book references, e.g., [8], the foundational paper for BIC.

Next, we mention some recent papers, which show applications of model selection in various research areas.

One such paper is [13], an application of principal component analysis and other methods to water quality assessment in a lake basin in China.

Another is [14], on feature selection for *classification* using principal component analysis.

As mentioned, a particularly interesting application of principal component analysis is in regression and logistic regression. We have mentioned the paper [10] on using principal component analysis in regression, taking several principal components to replace the set of explanatory variables. Another interesting application is in [15], on using principal components in *logistic* regression.

6. Conclusions

The problem of choice of the number of principal components to use to represent a complex of variables—a multivariate sample—has been considered in this chapter.

In addition to some *ad hoc* arithmetic criteria, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) have been applied here to the choice of the number of principal components to represent a dataset. The results have been compared and contrasted with *ad hoc* criteria such as retaining those principal components that explain more than an average amount of the total variance. The use of BIC is seen to correspond rather closely to the rule of retaining PCs whose eigenvalues are larger than average.

Acknowledgements

There are no further acknowledgements.

Authors' contributions

Stanley L. Sclove is the sole author.

Funding

There was no funding other than the author's usual salary at the university.

Competing interests

There are no competing interests.

Availability of data and material

The source of data used is a book that is referenced and available.

Abbreviations


AIC	Akaike's information criterion
BIC	Bayesian information criterion
DIAS	diastolic blood pressure
HT	height
LC	linear combination
LL	maximum log likelihood
MLE	maximum likelihood estimate
MSE	mean squared error
PC	principal component
SYS	systolic blood pressure
WT	weight

Author details

Stanley L. Sclove
University of Illinois at Chicago, Chicago, Illinois, USA

*Address all correspondence to: sclslove@uic.edu

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Sclove SL. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*. 1987;52(1987):333-343. DOI: 10.1007/BF02294360
- [2] Sclove SL. Principal components. In: Darity WA editor. *International Encyclopedia of the Social Sciences*, 2nd edition. Detroit, USA: Macmillan Reference
- [3] Anderson TW. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. New York, NY: Wiley; 2002
- [4] Johnson RJ, Wichern DW. *Applied Multivariate Statistical Analysis*. 6th ed. Upper Saddle River, NJ: Pearson; 2008
- [5] Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F, editors. *2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, USSR, September 2-8, 1971. Budapest: Akadémiai Kiadó; 1973. pp. 267-281 Republished in Kotz S, Johnson NL editors. *Breakthroughs in Statistics*, I. Berlin, Germany: Springer-Verlag; 1992. pp. 610-624
- [6] Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974;19(6): 716-723. DOI: 10.1109/TAC.1974.1100705
- [7] Akaike H. Prediction and entropy. In: Atkinson AC, Fienberg SE, editors. *A Celebration of Statistics*, Springer. NY: New York; 1985. pp. 1-24
- [8] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978;6:461-464 Available from: <http://www.jstor.org/stable/2958889>
- [9] Dixon WJ, Massey FJ Jr. *Introduction to Statistical Analysis*. 3rd ed. New York: McGraw-Hill; 1969
- [10] Massy WF. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*. 1965;60(309): 234-256. DOI: 10.1080/01621459.1965.10480787
- [11] Bai Z, Choi KP, Fujikoshi Y. Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *The Annals of Statistics*. 2018;46(3): 1050-1076. DOI: 10.1214/17-AOS1577
- [12] Bhatti MI, Al-Shanfari H, Hossain MZ. *Econometric Analysis of Model Selection and Model Testing*. Oxfordshire, England, UK: Routledge; 2017
- [13] Xu S, Cui Y, Yang C, Wei S, Dong W, Huang L, et al. The fuzzy comprehensive evaluation (FCE) and the principal component analysis (PCA) model simulation and its applications in water quality assessment of Nansi Lake Basin, China. *Environmental Engineering Research*. 2021;26(2):222-232
- [14] Omuya EO, Okeyo GO, Kimwele MW. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*. 2021;174: 114765
- [15] Aguilera AM, Escabias M, Valderrama MJ. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics and Data Analysis*. 2006;50(8): 1905-1924