Statistical Methods & Applications (2002) 11: 71–82

Cross-validation methods in principal component analysis: a comparison

Giancarlo Diana, Chiara Tommasi

Dipartimento di Scienze Statistiche, Università di Padova, Via Cesare Battisti, 241, 35121 Padova, Italy
(e-mail: diana@stat.unipd.it; microbo@stat.unipd.it)

Abstract. In principal component analysis (PCA), it is crucial to know how many principal components (PCs) should be retained in order to account for most of the data variability. A class of "objective" rules for finding this quantity is the class of cross-validation (CV) methods. In this work we compare three CV techniques showing how the performance of these methods depends on the covariance matrix structure. Finally we propose a rule for the choice of the "best" CV method and give an application to real data.

Key words: Principal component analysis, cross-validation methods

1 Introduction

PCA is a very used method in all branches of science and technology. PCA decomposes high-dimensional data into a low-dimensional subspace component and a noise component. A central issue in PCA is to choose the number of PCs to be retained, that is the dimension of the subspace. Various rules have been proposed for determining the real dimensionality of the data. Jolliffe [8] and Jackson [6] provide short reviews of these methods. Some of them, as the percentage of total variation, the Kaiser's rule (Kaiser [9]) and the scree test (Cattell [2]) depend on subjective choices. On the other hand, other criteria, as the Bartlett's rule (Bartlett [1]), are statistically based methods but they require distributional assumptions which are often unrealistic. Some other procedures are *ad hoc* rules, whose justification is mainly that they are intuitively plausible. Many of them, as the indicator function and the embedded error (Malinowski [13]) come from the chemometricians. This work focuses on another class of rules developed in chemometrics, the CV methods. These methods are quite objective, in fact they do not require distributional assumptions. Despite of traditional criteria, CV methods are not based on the eigenvalues of the sample covariance matrix but on the predictive ability of different PC models.

The basic idea of CV methods is the use of different data sets for estimation and validation of each PC model. The training set is the data fraction used to estimate each PC model. While predictive ability of each PC model is computed on the other data, firstly cancelled to form the evaluation set. The number of cancellation groups is fixed and computations are repeated as many times as the cancellation groups are. Each time a group is used as evaluation set so that all the data are predicted once. Actually, these CV methods involve two subjective choices, i.e. how many cancellation groups must be done? And how? There is not a general answer. We try to reply to these questions through a simulation study and provide some guidelines in Sect. 4.

Minka [14] uses Bayesian model selection to determine the effective data dimensionality and compares this method with CV and other algorithms. This method seems more accurate but it depends on some distributional assumptions, hence it is not quite objective.

A CV procedure, which uses the singular decomposition of the data matrix, was proposed by Eastment and Krzanowski [3]. See also Krzanowski [10] and [11]. An improvement is given by Scarponi *et al.* [15]. This method is based on successively predicting each element in the data matrix. In other words, at each step the evaluation set is formed only by one item of the data matrix. These type of CV methods, known as leave-one-out methods, use the maximum amount of information at the estimation stage but they can be time expensive.

In this work we treat the CV procedures based on the nonlinear iterative partial least squares (NIPALS) algorithm (Wold and Lyttkens [18]). Sect. 2 provides a short description of simple cross-validation (SCV) (Wold [16]), double cross-validation (DCV) (Wold [17]) and full cross-validation (FCV) (Forina *et al.* [4]) methods. These techniques are different in dividing estimation and validation stages and in the used stopping rule. The performance of any CV method is conjectured to depend on the structure of the covariance matrix characterizing the experimental data. This conjecture seems to be confirmed by a simulation study. In Sect. 3 data set generation details are provided. In Sect. 4 we give some remarkable simulation results and some general conclusions. Finally, in Sect. 5 we apply the SCV, DCV and FCV methods to a real data set yet considered by other authors and compare our results with theirs.

2 CV methods

Let $\boldsymbol{\mu}$ be the mean of the p-dimensional random vector $\mathbf{x}$ with covariance matrix Σ. The PCs of $\mathbf{x}$ are $\boldsymbol{\psi} = \Gamma'(\mathbf{x} - \boldsymbol{\mu})$ where $\Gamma = [\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p]$ is a $p \times p$ orthogonal matrix of Σ eigenvectors. Through a PCA a $n \times p$ data matrix X may be approximated by the k-th PC model

$$X = \mathbf{1}_n \boldsymbol{\mu}' + \boldsymbol{\psi}_1 \boldsymbol{\gamma}_1' + \boldsymbol{\psi}_2 \boldsymbol{\gamma}_2' + \cdots + \boldsymbol{\psi}_k \boldsymbol{\gamma}_k' + E^{(k)} \tag{1}$$

where $\mathbf{1}_n$ denotes the n-dimensional unit vector. Parameter vectors $\boldsymbol{\mu}$, $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$, describe the systematic part of the data and the residual matrix $E^{(k)} = X -$

$\mathbf{1}_n \boldsymbol{\mu}' - \sum_{h=1}^{k} \boldsymbol{\psi}_h \boldsymbol{\gamma}_h'$ resumes the random part, consisting of both "model" and "measurement" errors.

For k varying from 1 to p we may have p different PC models. The "optimum" number k^* of PCs is that k value corresponding to the PC model with the "best" prediction properties.

All the considered CV procedures estimate vectors $\boldsymbol{\psi}_k$ and $\boldsymbol{\gamma}_k$ as first PC of the residual matrix $E^{(k-1)} = \boldsymbol{\psi}_k \boldsymbol{\gamma}_k' + E^{(k)}$ by the NIPALS method.

As noted before, CV methods divide data matrix X in T cancellation groups. In turn they are deleted to form the training sets. Each training set is used to estimate PC models and then the corresponding cancellation group serves to evaluate the just estimated model. For SCV method the cancellation groups are rows of X. Thus the number of evaluation sets, T, is fixed so that the number of rows in each cancellation group is an integer. On the other hand DCV and FCV methods are based on cancellation groups which are sets of items and not rows of X. For all the three methods evaluation sets may be formed through a cancellation matrix M. M is a $n \times p$ matrix whose items are natural numbers into $\{1, \ldots, T\}$. The (i, j) element of M assigns the (i, j) element of X to a specific cancellation group: items of X, corresponding to M elements with the same value, belong to the same cancellation group. A cancellation matrix for DCV and FCV can be obtained following a *diagonal scheme* or a *random rule*. In the first case the cancellation matrix is built setting the values from 1 to T along its diagonals sequentially, while in the latter case by a random selection in $\{1, \ldots, T\}$. For instance, if we have a 6×4 data matrix and we want to form $T = 3$ cancellation groups, then we may have the three following cancellation matrices,

$$M_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \end{bmatrix} \quad M_2 = \begin{bmatrix} 1 & 2 & 1 & 1 \\ 3 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 1 \\ 3 & 3 & 2 & 3 \end{bmatrix} \quad M_3 = \begin{bmatrix} 3 & 3 & 1 & 2 \\ 3 & 1 & 3 & 2 \\ 3 & 2 & 3 & 1 \\ 1 & 2 & 2 & 3 \\ 2 & 2 & 1 & 1 \\ 1 & 3 & 2 & 1 \end{bmatrix}$$

M_1 is a cancellation matrix for SCV, where entire rows of X are the cancellation groups. While, for DCV and FCV, cancellation matrix M_2 or M_3 may be used. Specifically, M_2 is formed following a diagonal scheme and M_3 using a random rule.

The aim of this work is to identify when the best choice is to use a specific CV method with a particular number of cancellation groups and a specific deleting scheme. The idea is that different "optimal" combinations of these three "parameters" correspond to different structures of covariance matrix Σ. Thus in our simulation study, given a covariance matrix Σ, we generate data matrices X lying in a subspace of known dimensionality k^*. Then we check whether the CV methods identify such k^* value. In Sects. 2.1, 2.2 and 2.3 we describe in more detail SCV, DCV and FCV techniques. DCV and FCV were proposed to improve SCV and DCV, respectively. The object was to get a complete independence between prediction and validation stages. Specifically, in SCV the validation data are directly used to predict themselves, as shown at step e.2 in Sect. 2.1. In DCV the two phases are

more disjoint but validation data are partially used at the estimation stage again, e.g. steps A.5 and B.8 in Sect. 2.2. Only with the last version of these CV methods the complete independence is got.

2.1 SCV method

This procedure is based on the following steps.

1. Divide the X rows into T groups of the same size (index $t = 1, \ldots, T$). Any partition is good since the rows are independent.
2. For t varying from 1 to T, compute the partial *Prediction Residual Error Square Sum*, $PRESS_t(k)$, as follows.
   a. Delete the t-th group of X rows, that is the evaluation set, E_t. Let $\tilde{n}_t$ be the number of remaining rows of X, which form the t-th training set. Thus, the training set is a $\tilde{n}_t \times p$ matrix, denoted by $\tilde{X}_t = [\tilde{x}_{i,j}]$.
   b. Set $k = 0$ and estimate the first PC model, $x_{i,j} = \mu_j + \varepsilon_{i,j}^{(0)}$, computing
   $$\tilde{\mu}_j = \textstyle\sum_{\{i,j:x_{i,j} \in \tilde{X}_t\}} \tilde{x}_{i,j}/\tilde{n}_t, \quad j = 1, \ldots, p.$$
   c. Form $(n-\tilde{n}_t)\times p$ prediction error matrix, $E_t^{(0)} = [x_{i,j} - \tilde{\mu}_j] = \left[e_{i,j}^{(0)}\right]$, where (i, j) are such that $x_{i,j} \in E_t$, and compute
   $$PRESS_t(0) = \sum_{\{i,j:x_{i,j} \in E_t\}} \left|e_{i,j}^{(0)}\right|^2.$$
   d. Form $\tilde{n}_t \times p$ residual matrix,
   $$\tilde{E}_t^{(0)} = [x_{i,j} - \tilde{\mu}_j] \quad \text{where} \quad x_{i,j} \in \tilde{X}_t.$$
   e. For k varying from 1 to $\min\{p-2, \tilde{n}_t - 2\}$,
      e.1. estimate the k-th PC model from $\tilde{X}_t$, computing ψ_k and γ_k as first PC and corresponding loading factors of $\tilde{E}_t^{(k-1)}$. Let $\tilde{\psi}_k^{(t)}$ and $\tilde{\gamma}_k^{(t)}$ be such estimates.
      e.2. In order to compute the predicted values for the evaluation set $E_t^{(k-1)}$, we cannot use vector $\tilde{\psi}_k^{(t)}$ since it is a $\tilde{n}_t \times 1$ vector. Thus, estimate the first PC of $E_t^{(k-1)}$ as $\psi_k^{(t)} = E_t^{(k-1)}\, \tilde{\gamma}_k^{(t)}$ and compute the predicted values as $\psi_k^{(t)}\, \tilde{\gamma}_k^{\prime(t)}$.
      e.3. Form $(n-\tilde{n}_t)\times p$ prediction error matrix $E_t^{(k)} = E_t^{(k-1)} - \psi_k^{(t)}\, \tilde{\gamma}_k^{\prime(t)} = \left[e_{i,j}^{(k)}\right]$, where (i, j) are such that $x_{i,j} \in E_t$, and compute
      $$PRESS_t(k) = \sum_{\{i,j:x_{i,j} \in E_t\}} \left|e_{i,j}^{(k)}\right|^2.$$
      e.4. Form $\tilde{n}_t \times p$ residual matrix,
      $$\tilde{E}_t^{(k)} = \tilde{E}_t^{(k-1)} - \tilde{\psi}_k^{(t)}\, \tilde{\gamma}_k^{\prime(t)}.$$
   f. Restore data matrix X.

3. Compute the total $PRESS(k)$ for each k value:

$$PRESS(k) = \sum_{t=1}^{T} PRESS_t(k).$$

4. Making F-tests on the quantity

$$\frac{[PRESS(k-1) - PRESS(k)]/n}{PRESS(k)/n(p-k-1)}$$

determine whether the last product term in model (1) is significant or not.

2.2 DCV method

A. Cross-validate μ as follows

A.1. Compute the residual square errors, $RSE(0)$, that at the first step is the sum of squared deviations from the origin

$$RSE(0) = \sum_{i=1}^{n}\sum_{j=1}^{p} (x_{ij} - \overline{x})^2 ,$$

where $\overline{x} = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}/np$.

A.2. Divide the X rows into T groups of the same size.

A.3. For t varying from 1 to T

A.3.1. delete the t-th group of rows, E_t, forming reduced matrix $\tilde{X}_t$. Compute $\tilde{\mu}_j$, $j = 1, \ldots, p$, as the average of $\tilde{x}_{ij}$ over all the $\tilde{n}_t$ objects of $\tilde{X}_t$.

A.3.2. Form the prediction errors as the differences between the deleted elements and $\tilde{\mu}_j$ and compute the corresponding partial sum of squares,

$$PRESS_t(0) = \sum_{\{i,j:\, x_{ij} \in E_t\}} (x_{ij} - \tilde{\mu}_j)^2 .$$

A.4. Form the total $PRESS(0)$ by summing the partial $PRESS_t(0)$ from step A.3,

$$PRESS(0) = \sum_{t=1}^{T} PRESS_t(0)$$

and compute the ratio $R(0) = PRESS(0)/RSE(0)$. A ratio $R(0)$ smaller than one shows that the predictions are improved by including μ in equation (1).

A.5. If $R(0) \leq 1$, compute $\hat{\mu}_j$, for $j = 1, \ldots, p$, as the average of x_{ij} over all the n objects of X. Form the $n \times p$ residual matrix $\tilde{E}^{(0)} = \left[\tilde{e}_{ij}^{(0)}\right]$, where $\tilde{e}_{ij}^{(0)} = x_{ij} - \hat{\mu}_j$. Go forward to step B.

A.6. If $R(0) > 1$, set $\tilde{E}^{(0)} = X$.

B. Cross-validate the k-th component.

B.1. Set $k = 0$.
B.2. Set $k = k + 1$.
B.3. Form the residual sum of squares, $RSE(k) = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(\tilde{e}_{ij}^{(k-1)}\right)^2$.
B.4. Divide the residual matrix $\tilde{E}^{(k-1)} = \left[\tilde{e}_{ij}^{(k-1)}\right]$ into T groups through a cancellation matrix, following a diagonal scheme or a random rule.
B.5. For t varying from 1 to T
  B.5.1. delete the t-th group in matrix $\tilde{E}^{(k-1)}$. Let E_t be the cancellation group.
  B.5.2. Give some convenient values to the just deleted elements of $\tilde{E}^{(k-1)}$, forming $\tilde{E}_t^{(k-1)}$. For instance, the missing values may be replaced by the average of the remaining elements in the same column.
  B.5.3. Estimate model (1) by NIPALS method. This method computes the estimates of ψ_{ik} and γ_{kj}, $i = 1, \ldots, n$, $j = 1, \ldots, P_j$ in model (1) as first PC and corresponding loading factors of the matrix $\tilde{E}_t^{(k-1)}$. Let $\tilde{\psi}_{ik}^{(t)}$ and $\tilde{\gamma}_{kj}^{(t)}$ be such estimates.
  B.5.4. Form the prediction errors as the differences between the deleted elements in $\tilde{E}^{(k-1)}$ and $\tilde{\psi}_{ik}^{(t)} \tilde{\gamma}_{kj}^{(t)}$. Sum the squares of these differences to give the partial $PRESS$,

$$PRESS_t(k) = \sum_{i \in E_t} \sum_{\{i,j:\tilde{e}_{ij}^{(k-1)} \in E_t\}} \left(\tilde{e}_{ij}^{(k-1)} - \tilde{\psi}_{ik}^{(t)}\, \tilde{\gamma}_{kj}^{(t)}\right)^2.$$

B.6. Form the total $PRESS$ from the partial sums,

$$PRESS(k) = \sum_{t=1}^{T} PRESS_t(k)$$

and compute the ratio $R(k) = PRESS(k)/RSE(k)$.
B.7. If $R(k) > 1$, then the procedure stops because the inclusion of the last term (the k-th) in model (1) do not improve the prediction errors, thus $k^* = k-1$. Otherwise, if $R(k) \leq 1$ proceed to step B.8.
B.8. Make a PCA on the complete matrix $\tilde{E}^{(k-1)}$ giving the final estimates $\hat{\psi}_{ik}^{(t)}$ and $\hat{\gamma}_{kj}^{(t)}$ of $\psi_{ik}^{(t)}$ and $\gamma_{kj}^{(t)}$, respectively ($i = 1, \ldots, n,\ j = 1, \ldots, p$). Form a new residual matrix $\tilde{E}^{(k)}$ subtracting $\hat{\psi}_{ik}^{(t)} \hat{\gamma}_{kj}^{(t)}$ from $\tilde{e}_{ij}^{(k-1)}$ for all i and j. Go back to step B.2.

2.3 FCV method

The following steps are followed.

1. Divide data matrix X into T groups.
2. For t varing from 1 to T,
   a. delete the t-th group, E_t, in data matrix X forming a $n \times p$ matrix $\tilde{X}_t = [\tilde{x}_{i,j}]$ with some missing values.

b. Set $k = 0$ and compute $\tilde{\mu}_j$, $j = 1, \ldots, p$, as the sample mean of $\tilde{x}_{i,j}$.
c. Use $\tilde{\mu}_j$ to estimate the missing data in $\tilde{X}_t$.
d. Compute the prediction errors, $e_{ij}^{(0)} = x_{i,j} - \tilde{\mu}_j$, where $x_{ij} \in E_t$, and

$$PRESS_t(0) = \sum_{\{i,j:x_{i,j} \in E_t\}} \left| e_{i,j}^{(0)} \right|^2 .$$

e. Form the $n \times p$ residual matrix, $\tilde{E}_t^{(0)} = \left[\tilde{e}_{ij}^{(0)} \right]$, where

$$\tilde{e}_{i,j}^{(0)} = \begin{cases} x_{ij} - \tilde{\mu}_j & \text{if } x_{ij} \in \tilde{X}_t \\ 0 & \text{if } x_{ij} \in E_t \end{cases}$$

f. For k varing from 1 to $p - 1$,
   f.1. estimate the k-th PC model from $\tilde{X}_t$, computing ψ_k and γ_k as first PC and corresponding loading factors of $\tilde{E}_t^{(k-1)}$. Let $\tilde{\psi}_k^{(t)}$ and $\tilde{\gamma}_k^{(t)}$ be such estimates.
   f.2. Form the prediction errors,

$$e_{ij}^{(k)} = e_{ij}^{(k-1)} - \tilde{\psi}_{ik}^{(t)} \, \tilde{\gamma}_{jk}^{(t)}$$

   and compute

$$PRESS_t(k) = \sum_{\{i,j:x_{i,j} \in E_t\}} \left| e_{i,j}^{(k)} \right|^2 .$$

   f.3. Compute the residual matrix

$$\tilde{E}_t^{(k)} = \tilde{E}_t^{(k-1)} - \tilde{\psi}_k^{(t)} \, \tilde{\gamma}_k^{\prime(t)}.$$

g. Restore data matrix X.

3. Form the total $PRESS(k)$ from the partial sums:

$$PRESS(k) = \textstyle\sum_{t=1}^{T} PRESS_t(k)$$

and compute the number of significant components as the k value which minimizes $PRESS(k)$.

3 Simulation: data generation

In order to generate data which approximately lie on a specific subspace, we consider the observable vector variable $\mathbf{x}$ decomposed as

$$\mathbf{x} = \mathbf{z} + \mathbf{u}, \tag{2}$$

where the unobservable random vector $\mathbf{z}$ can be considered as the "systematic" or "true" part while the unobservable random vector $\mathbf{u}$ is the " error of measurements". What distinguishes the systematic part is that it varies in a lower-dimensional linear

space of dimension k^*, where $k^* < p$. The random error $\mathbf{u}$ is assumed with null mean, equal variances and incorrelated components,

$$E(\mathbf{u}) = \mathbf{0}, \quad \mathrm{Var}(\mathbf{u}) = \sigma_u^2 I_p,$$

where σ_u^2 is unknown. This form of the covariance matrix may be appropriate when independent measurements are made with the same instrument. Furthermore, $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are mutually independent and also independent from $\mathbf{z}_1, \ldots, \mathbf{z}_n$. With this data model, appropriate in many situations related to chemistry and biological fields, the number of PCs to be retained is known to be k^*.

Let Σ_z denote the covariance matrix of the "true" part varing in a k^*-dimensional space. Of course Σ_z has rank k^*. Let $a_1, \ldots, a_{k^*}$ be the positive Σ_z eigenvalues. From model (2) we have that $\Sigma = \Sigma_z + \sigma_u^2 I_p$. As a consequence of the spectral decomposition theorem, the Σ eigenvalues are $\lambda_j = a_j + \sigma_u^2$ for $j = 1, \ldots, k^*$ and $\lambda_j = \sigma_u^2$ for $j = k^* + 1, \ldots, p$. We will call the first k^* eigenvalues as significant eigenvalues.

In order to generate random vector from a fixed covariance matrix, we choose k^* positive values a_j and the error variance σ_u^2 and so we obtain the diagonal matrix Λ of Σ eigenvalues. Then by the Heiberger's algorithm (Heiberger, 1978) we generate an orthonormal matrix Γ that we consider as Σ eigenvector matrix. Finally we obtain a covariance matrix with the desired structure using the spectral decomposition theorem, $\Sigma = \Gamma \Lambda \Gamma'$. At this point we can generate the data belonging to a subspace except for a random error. The X rows are n independent observations from a random vector with the covariance matrix obtained above. For example, $\mathbf{x}_i \sim N\left(0, \Gamma \Lambda^{1/2} \Gamma'\right)$, for $i = 1, \ldots, n$.

Moreover, through this procedure, we can generate covariance matrices with different structures. In the following, we consider *diagonal*, *block diagonal* and *general* covariance matrices and choose different number k^* of significant eigenvalues.

For each type of structure for Σ we generate 1000 $n \times p$ data matrices. For each replication we estimate k^* through the three CV methods, for three different numbers of evaluation sets and the two deleting schemes. We repeat this process for n equals to 20 and 50, for p equals to 8 and 10 and for two different values of σ_u^2. For each combination of the previous values of n, p, σ_u^2 and for each number of validation sets and deleting scheme, we have the number of times that each CV method chooses the real dimension k^*.

4 Results and discussion

In the simulation study, the artificial data sets come from covariance matrices which are different in their structures and eigenvalues. A first result is that whenever Σ is diagonal the best CV technique is the SCV method with 2 cancellation groups (see Fig. 1).

The reason is that with this method entire rows of X are deleted to make the evaluation sets and then NIPALS method, to estimate k^*, is immediately applied to the training set. While in the other two CV techniques, items of X are deleted

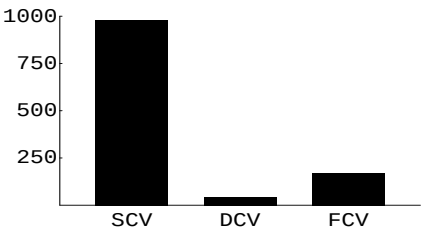

Fig. 1. Number of times that SCV, DCV and FCV choose $k^* = 3$ in 1000 replications. Specifically, Σ is a diagonal matrix, $n = 50$, $p = 8$, $\sigma_u^2 = 0.1$; $\lambda_1 = 5$, $\lambda_2 = 4$, $\lambda_3 = 3$

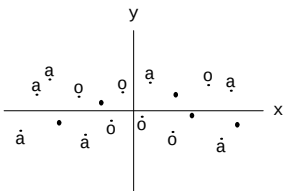

Fig. 2. A data set where the cancellation group is formed by the "a" point abscissae and "o" point ordinates

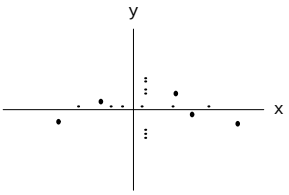

Fig. 3. The data set after the deleted data replacement

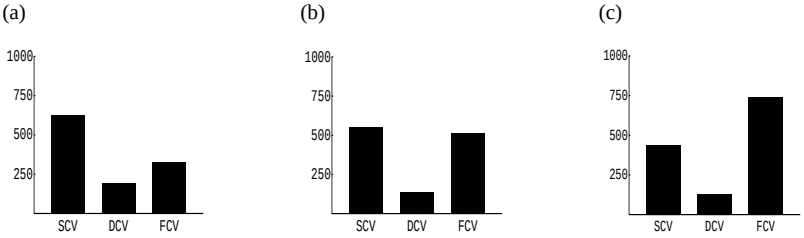

Fig. 4. Number of times that SCV, DCV and FCV choose $k^* = 5$ in 1000 replications. Specifically, Σ is a block diagonal matrix, $n = 20$, $p = 10$, $\sigma_u^2 = 0.1$; in (a) $\lambda_1 = 8$, $\lambda_2 = 4$, $\lambda_3 = 2.2$, $\lambda_4 = 1.9$, $\lambda_5 = 1.5$; in (b) $\lambda_1 = 5$, $\lambda_2 = 4.5$, $\lambda_3 = 4.5$, $\lambda_4 = 2$, $\lambda_5 = 1.6$; in (c) $\lambda_1 = 5$, $\lambda_2 = 4$, $\lambda_3 = 3$, $\lambda_4 = 3$, $\lambda_5 = 2.6$

to form a cancellation group. In order to estimate k^*, data matrix X must not have missing data, thus the previously deleted data are replaced by the average of the remaining data in the same column. Of course, some deleted data are coordinates associated to principal axes. Replacing these coordinates by a common value causes a reduction on the variability along principal axes. For example, Fig. 2 shows a data set where the x axis is the principal one. The cancellation group is formed by the "a" point abscissae and "o" point ordinates. After the data replacement, the variability along both axes has decreased but the reduction along the x axis is more evident, as shown in Fig. 3.

When Σ is not diagonal, there is a trade-off between SCV and FCV techniques. Whenever Σ significant eigenvalues are quite different, that is $\max_{j=1,\ldots,k^*} \lambda_j / \lambda_{j+1} > 1.5$, FCV method provides bad results while SCV performance is good. As $\max_{j=1,\ldots,k^*} \lambda_j / \lambda_{j+1}$ becomes smaller, FCV behaviour improves and SCV

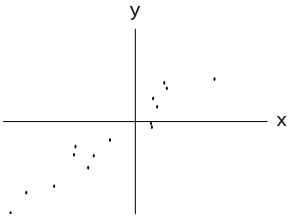

Figure 5. The original data set

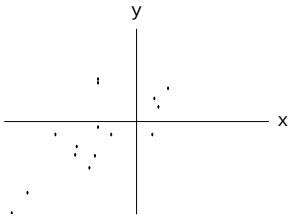

Figure 6. The data set after the deleted data replacement

performance becomes worse. This is well shown in Fig. 4, where the significant eigenvalue sum is the same but the eigenvalues become closer passing from (a) to (c). This behaviour also is due to the different way of forming the deleting sets. When the variability along the principal axes is large substituting the cancelled data with the column means of the remainder data changes so much original data that they have anymore the same "true" dimensionality. This is well illustrated by Figs. 5 and 6. On the other hand, SCV technique simply drops same points to form the cancellation group, thus it gives good results in this condition.

The DCV method seems to give the worst results. Other simulations shows that this method provides quite good results only when Σ eigenvalues are "very" different. Anyway, in this last case, DCV seems to perform better than FCV but worse than SCV.

Therefore our conjecture that the appropriate CV technique depends on the covariance matrix structure seems confirmed. We suggest to use SCV method with 2 cancellation groups whenever Σ is "quite" diagonal or when its eigenvalues satisfy $\max_{j=1,\ldots,k^*} \lambda_j/\lambda_{j+1} > 1.5$. Otherwise it seems better to use FCV method with the random deleting scheme and not less than p cancellation groups. The larger the number of cancellation groups more accurate the estimates for the deleting data, since the training set is larger. For this reason FCV technique gives best results when more than p cancellation groups are used. On the contrary, SCV method performs very well with only two deleting groups. As told in Sect. 2 with this method the deleting data estimates are based on the training set but also on the evaluation data. Thus, we have good estimates if we form training and evaluation sets dividing the data into two.

Finally, we observe that all CV techniques seem improve when the number of rows n increases and/or σ_u^2 decreases. The justification is quite obvious. If n increases, at each step of CV methods, we have more data to estimate PC models and if error variance σ_u^2 decreases the real dimension k^* becomes more evident.

When the variables are expressed in different scale it is convenient to standardize the data set and hence to consider correlation (rather than covariance) matrix. In the next section we provide a real data application where this standardization is necessary.

5 A real data analysis

Krzanowski [12] provides a data set which comprise 19 variables measured on each of 40 winged aphids (*alate adelges*) that had been caught in a light trap. Actually, that data set was previously studied by Jeffers [7]. Because of the disparate nature of the variables, both Jeffers [7] and Krzanowski [12] elected to standardize the data and then to make a PCA on the correlation matrix. Using standard methods, Jeffrey concludes that the essential dimensionality of the data is two, while Krzanowski [12], basing his analysis on the Eastment and Krzanowski's [3] method, suggests to retain four PCs.

Looking at the sample correlation matrix, it does not seem "quite" diagonal but the largest ratio of the pairs of subsequent eigenvalues is about 5.8. The whole list of eigenvalues can be found in Krzanowski [11]. Therefore following the guidelines given in Sect. 4 we have that, among the considered CV methods, SCV should be the most suitable for this data set. Table 1 gives the number of PCs that should be retained according to SCV, DCV and FCV, for four numbers of cancellation groups and two deleting schemes. As expected, SCV method gives the most reliable results.

Table 1. Number of principal components obtained from SCV, DCV and FCV, for different numbers of cancellation groups T and two deleting schemes.

		DCV		FCV	
T	SCV	random rule	diag. scheme	random rule	diag. scheme
2	5	2	2	1	1
5	4	2	2	2	2
10	4	2	2	2	2
20	4	2	2	2	2

Using two cancellation groups we get that five PCs should be retained, otherwise we obtain the same result as Krzanowski [12]. However the difference is slight.

*Acknowledgements.*The authors are thankful to the referee for his useful suggestions which led to improvement of the paper.

References

1. Bartlett MS (1950) Test of significance in factor analysis. *Br. J. Psych. Stat.* 3: 77–85
2. Cattel RB (1966) The Scree test for the number of factors. *Mult. Behav. Res.* 1: 245–276
3. Eastment HT, Krzanowski WJ (1982) Cross-validatory choice of the number of component analysis. *Technometrics* 24: 73–77
4. Forina M, Lanteri S, Boggia R, Bertran E (1993) Double cross full validation. *Química Analítica* 12: 128–135
5. Heiberger RM (1978) AS 127. Generation of random orthogonal matrices. *Applied Statistics* 27: 199–206
6. Jackson JE (1991) A user's guide to principal components. Wiley, New York
7. Jeffers JNR (1967) Two case studies in the application of principal components analysis. *Applied Statistics* 16: 225–236
8. Jolliffe IT (1986) Principal component analysis. Springer, Berlin Heidelberg New York

9. Kaiser HF (1960) The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* 20: 141–151
10. Krzanowski WJ (1983) Cross-validatory choice of the number inPprincipal component analysis; some sampling results. *J. Statist. Comput. Simul.* 18: 299–314
11. Krzanowski WJ (1987) Cross-validation in principal component analysis. *Biometrics* 43: 575–584
12. Krzanowski WJ (1987) Selection of variables to preserve multivariate data structure, using principal components. *Applied Statistics* 36: 22–33
13. Malinowski ER (1977) Theory of error in factor analysis. *Analytical Chemistry* 49: 606–612
14. Minka TP Automatic choice of dimensionality for PCA. Technical Report n. 514 (2000), MIT Media Laboratory, Vision and Modelling Group. http://citeseer.nj.nec.com/minkaooautomaic.html
15. Scarponi G, Moret I, Capodaglio G, Romanazzi M (1990) Cross-validation, influential observations and selection of variables in chemometric studies of wines by principal component analysis. *Journal of Chemometrics* 4: 217–240
16. Wold S (1976) Pattern recognition by means of disjoint principal components models. *Pattern Recognition* 8: 127–139
17. Wold S (1978) Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 20: 397–405
18. Wold H, Lyttkens E (1969) Nonlinear iterative partial least squares (NIPALS) estimation procedures. Bull. Intern. Statist. Inst.: Proc. 37th Session, pp. 1–15. London