New Cross-Validation Methods in Principal Component Analysis

J. Camacho, J. Picó
Department of Systems Engineering and Control.

A. Ferrer
Department of Applied Statistics, Operations Research and Quality.

Technical University of Valencia
Cno. de Vera s/n, 46022, Valencia, Spain

Two novel cross-validation algorithms to select the number of principal components (PCs) in Principal Components Analysis (PCA) are presented. These algorithms yield a 100% of effectiveness in determining the correct number of PCs in all the simulated data sets studied, for measurement noise levels up to a 30% and 40%.

Nomenclature $(*)$ stands for the data used to fit the PCA model and $(\#)$ stands for the data predicted in each cross-validation iteration.

1. Traditional methods

1.1 Wold (1978)

The proposal by Wold is a very fast cross-validation algorithm based on the NIPALS procedure. In each iteration, the computations are performed from the residuals. Wold suggested to include PCs to the PCA model whereas the following index R is below 1:

$$R_a = \frac{PRESS_a}{SSE_{a-1}}$$

where $PRESS_a$ is the sum-of-squares of prediction errors computed for a PCs, and SSE_{a-1} is the sum of squared residuals after $a-1$ PCs have been extracted.

Drawback: The index used to select the number of PCs is heuristic whereas the threshold -1- imposes a hard condition.

1.2 Eastment and Krzanowski (1982)

Leave-one-out procedure based on the singular value decomposition (SVD) algorithm. This approach includes in the PCA model all the PCs up to the last one for which the following index W exceeds 1:

$$W_a = \frac{(PRESS_{a-1} - PRESS_a)/DOF_a}{PRESS_a/DOF_{rem}}$$

where DOF_a is the number of degrees-of-freedom (DOFs) used to fit the a-th PC and DOF_{rem} is the remaining DOFs after the a-th PC has been added to the model.

Drawback: The index used to select the number of PCs is heuristic whereas the threshold -1- imposes a hard condition.

1.3 Leave-n-objects-out (LnOO)

For each PC ($a = 1...A$)
  For each group of objects ($g = 1...G$)
    Form $\mathbf{X}^*$ with data from all groups but g
    Form $\mathbf{X}^\#$ with data from g
    Calibrate a PCA model from $\mathbf{X}^*$,
      obtaining $\mathbf{P}^*_a$ and $\mathbf{T}^*_a$
    $\mathbf{T}^\#_a = \mathbf{X}^\# \cdot \mathbf{P}^*_a$
    $\hat{\mathbf{X}}^\# = \mathbf{T}^\#_a \cdot \mathbf{P}^{*t}_a$
    $\mathbf{E}_g = \mathbf{X}^\# - \hat{\mathbf{X}}^\#$
  end
  $PRESS_a = \sum_{n=1}^{N}\sum_{m=1}^{M} e^2_{n,m}$
end

Drawback: The $PRESS_a$ is monotonously decreasing with a and so its minimum cannot be used directly to select the number of PCs.

1.4 Leave-n-samples-out (LnSO)

For each PC ($a = 1...A$)
  For each group of objects ($g = 1...G$)
    Form $\mathbf{X}^*$ with data from all groups but g
    Form $\mathbf{X}^\#$ with data from g
    Calibrate a PCA model from $\mathbf{X}^*$,
      obtaining $\mathbf{P}^*_a$ and $\mathbf{T}^*_a$
    For each group of variables ($h = 1...H$)
      Set $\mathbf{X}^\#_h = 0$
      $\mathbf{T}^\#_a = \mathbf{X}^\# \cdot \mathbf{P}^*_a$
      $\hat{\mathbf{X}}^\# = \mathbf{T}^\#_a \cdot \mathbf{P}^{*t}_a$
      Restore its actual value to $\mathbf{X}^\#_h$
      $\mathbf{E}_{g,h} = \mathbf{X}^\#_h - \hat{\mathbf{X}}^\#_h$
    end
  end
  $PRESS_a = \sum_{n=1}^{N}\sum_{m=1}^{M} e^2_{n,m}$
end

Drawback: PCs modelling independent variables do not reduce the PRESS and so, they are not selected.

2. Proposed Algorithms

The approach of this poster is to correct the LnSO method by replicating the information in the data, so that independent variables are not independent any more. To reduce the effect of the measurement noise, the information is duplicated using the PCA subspace. Two choices:

$$\mathbf{X}_{aug} = [\mathbf{X}, \mathbf{T}_a] \qquad \mathbf{X}_{aug} = [\mathbf{X}, \mathbf{T}_a \cdot \mathbf{P}^t_a]$$

Computational efficiency: CLnSO needs the calibration of a PCA model for each of the $G \times H$ different groups of samples, whereas in fast-CLnSO, LnSO and LnOO a PCA model is fitted only for each of the G groups of objects. The algorithm by Eastment and Krzanowski (1982) needs of $G + H$ SVD runs and the one by Wold (1978) needs of G PCA runs.

2.1 Fast Corrected-leave-n-samples-out (fast-CLnSO)

For each PC ($a = 1...A$)
  Calibrate a PCA model from $\mathbf{X}$, obtaining $\mathbf{P}_a$ and $\mathbf{T}_a$
  For each group of objects ($g = 1...G$)
    Form $\mathbf{X}^*$ and $\mathbf{T}^*_a$ with data from all groups but g
    Form $\mathbf{X}^\#$ and $\mathbf{T}^\#_a$ with data from g
    $\mathbf{X}^*_{aug} = [\mathbf{X}^*, \mathbf{T}^*_a]$, remember not to scale $\mathbf{T}^*_a$
    Calibrate a PCA model from $\mathbf{X}^*_{aug}$, obtaining $\mathbf{P}^*_{aug,a}$ and $\mathbf{T}^*_{aug,a}$
    For each group of variables ($h = 1...H$)
      Set $\mathbf{X}^\#_h = 0$
      $\mathbf{X}^\#_{aug} = [\mathbf{X}^\#, \mathbf{T}^\#_a]$
      $\mathbf{T}^\#_{aug,a} = \mathbf{X}^\#_{aug} \cdot \mathbf{P}^*_{aug,a}$
      $\hat{\mathbf{X}}^\#_{aug} = \mathbf{T}^\#_{aug,a} \cdot \mathbf{P}^{*t}_{aug,a}$
      Restore its actual value to $\mathbf{X}^\#_h$
      $\mathbf{E}_{g,h} = \mathbf{X}^\#_h - \hat{\mathbf{X}}^\#_h$
    end
  end
  $PRESS_a = \sum_{n=1}^{N}\sum_{m=1}^{M} e^2_{n,m}$
end

2.2 Corrected-leave-n-samples-out (CLnSO)

For each PC ($a = 1...A$)
  Calibrate a PCA model from $\mathbf{X}$, obtaining $\mathbf{P}_a$ and $\mathbf{T}_a$
  For each group of objects ($g = 1...G$)
    Form $\mathbf{X}^*$ and $\mathbf{T}^*_a$ with data from all groups but g
    Form $\mathbf{X}^\#$ and $\mathbf{T}^\#_a$ with data from g
    For each group of variables ($h = 1...H$)
      $\mathbf{X}^*_{aug} = [\mathbf{X}^*, \mathbf{T}^*_a \cdot \mathbf{P}^t_{a,h}]$
      Calibrate a PCA model from $\mathbf{X}^*_{aug}$, obtaining $\mathbf{P}^*_{aug,a}$ and $\mathbf{T}^*_{aug,a}$
      Set $\mathbf{X}^\#_h = 0$
      $\mathbf{X}^\#_{aug} = [\mathbf{X}^\#, \mathbf{T}^\#_a \cdot \mathbf{P}^t_{a,h}]$
      $\mathbf{T}^\#_{aug,a} = \mathbf{X}^\#_{aug} \cdot \mathbf{P}^*_{aug,a}$
      $\hat{\mathbf{X}}^\#_{aug} = \mathbf{T}^\#_{aug,a} \cdot \mathbf{P}^{*t}_{aug,a}$
      Restore its actual value to $\mathbf{X}^\#_h$
      $\mathbf{E}_{g,h} = \mathbf{X}^\#_h - \hat{\mathbf{X}}^\#_h$
    end
  end
  $PRESS_a = \sum_{n=1}^{N}\sum_{m=1}^{M} e^2_{n,m}$
end

3. Experimental Results

3.1 First simulated data set

10 observable variables from 8 latent variables:
$x_i = lv_j + lv_k, i\epsilon\{1,..,6\}, j \neq k\epsilon\{1,..,4\}$
$x_i = lv_j + lv_k, i\epsilon\{7,8,9\}, j \neq k\epsilon\{5,6,7\}$
$x_{10} = lv_8$

% Noise	R	W	L1SO	fast-CL1SO	CL1SO
10%	2	2	6	**8**	**8**
20%	2	2	6	**8**	**8**
30%	2	2	6	**8**	**8**
40%	2	2	6	**8**	**8**
50%	2	0	6	9	9

3.2 Second simulated data set

27 observable variables from 12 latent variables:
$x_i = lv_j, i\epsilon\{1,..,12\}, j\epsilon\{1,..,12\}$
$x_i = lv_j + lv_k, i\epsilon\{13,..,27\}, j \neq k\epsilon\{1,..,6\}$

% Noise	R	W	L1SO	fast-CL1SO	CL1SO
10%	6	11	**12**	**12**	**12**
20%	6	11	**12**	**12**	**12**
30%	6	6	**12**	**12**	**12**
40%	6	11	**12**	13	**12**
50%	6	6	**12**	19	17
60%	6	6	10	20	20

3.3 Third simulated data set

50 observable variables from 15 latent variables:
$x_i = lv_j, i\epsilon\{1,..,5\}, j\epsilon\{1,..,5\}$
$x_i = lv_j + lv_k, i\epsilon\{6,..,50\}, j \neq k\epsilon\{6,..,15\}$

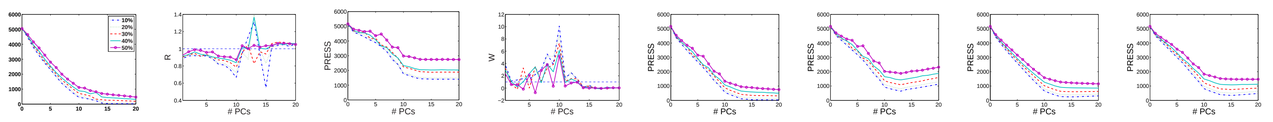
% Noise	R	W	L1SO	fast-CL1SO	CL1SO
10%	11	13	13	**15**	**15**
20%	12	13	13	**15**	**15**
30%	11	13	13	**15**	**15**
40%	10	13	13	20	**15**
50%	10	10	13	20	16

3.4 McReynolds Data

The data set from (McReynolds, 1970). The data was analyzed with and without outliers (a total of 13 outliers are found by Wold and Andersson (1973)).

	R	W	L1SO	fast-CL1SO	CL1SO
Full	2	4	1	1	1
Reduced	5	3	1	1	1

Third simulated data set From left to right.: PRESS and R-statistic [3], PRESS and W-statistic [1], PRESS with L1OO, L1SO, fast-CL1SO and CL1SO.



4. Conclusions

a) Both the R-statistic and the W-statistic follow heuristical laws, more or less theoretically justified. Although they have proven to be useful when they are visually inspected, it is not possible to define a hard threshold -like 1- which works for the general case.

b) The L1SO approach presents problems in the selection of the number of PCs when the eigenvalues corresponding to the PCs are very different.

c) Both fast-CL1SO and CL1SO determined correctly the number of PCs when data is corrupted with up to a 30% and a 40% of measurement noise, respectively, for the simulated data studied. Nonetheless, all these results correspond to data generated following the PCA structure and, thus, nothing can be said about any other type of data.

[1] Eastment, H.T., and Krzanowski, W.J. (1982), Cross-Validatory Choice of the Number of Components From a Principal Component Analysis, Technometrics, 24, 73-77.

[2] McReynolds (1970), Characterization of Some Liquid Phases, Journal of Chromatography Science, 8, 685-691.

[3] Wold, S. (1978), Cross-Validatory Estimation of the Number of Components in Factor and Principal Components, Technometrics, 20, 397-405.

[4] Wold, S., and Andersson, K. (1973), Major Components Influencing Retention Indices in Gas Chromatography, Journal of Chromatography, 80, 43-59.