Leave-n-Samples-Out Cross-validation in PCA for Missing Data Recovery and Robustness to Measurement Noise

J. Camacho, J. Picó
Instituto de Automática e Informática Industrial.
A. Ferrer
Departamento de Estadística e Investigación Operativa Aplicadas y Calidad.
Universidad Politécnica de Valencia
Cno. de Vera s/n, 46022, Valencia, Spain

One of the most used algorithms for PCA cross-validation is the leave-n-samples-out (LnSO) algorithm. The curve of PRedictive Error Sum-of-Squares (PRESS) computed by LnSO has the nice property that the minimum value signals the optimum number of PCs in terms of predictive error. In this poster, the theoretical properties of the LnSO method are stated for the first time (section 1). Furthermore, some modifications of the algorithm to improve its performance for the imputation of missing values in the model exploitation (section 2) and to enhance its robustness to independent measurement noise (section 3) are studied.

1. Analysis of the Leave-n-Samples-Out (LnSO) method

1.1 Algorithm

The inner loop is highlighted in yellow color and the core in red color:

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
    Fit a PCA model from $\mathbf{X}^*$, obtaining $\mathbf{P}_a^*$ and $\mathbf{T}_a^*$
    For each group of variables ($h = 1...H$)
      Set $\mathbf{X}_h^\# = 0$ (*)
      $\mathbf{T}_a^\# = \mathbf{X}^\# \cdot \mathbf{P}_a^*$
      $\hat{\mathbf{X}}^\# = \mathbf{T}_a^\# \cdot \mathbf{P}_a^{*t}$
      Restore its actual value to $\mathbf{X}_h^\#$
      $\mathbf{E}_{g,h} = \mathbf{X}_h^\# - \hat{\mathbf{X}}_h^\#$
    end
  end
  $PRESS_a = \sum_{n=1}^{N} \sum_{m=1}^{M} e_{n,m}^2$
end

(*) Initial estimation

1.2 Properties

In linear PCA, an observable variable is understood as the sum of: **Redundant information**, which can be found in another observable variable, and **Non-redundant information** - linearly independent information, which is not found in any other observable variable.

P1 *The error of estimation in LnSO of a variable containing only redundant information for a PCA model with $A = Rank(\mathbf{X})$ depends on the initial estimation.*

P2 *The error of estimation in LnSO of a variable with any content of non-redundant information for a PCA model with $A = Rank(\mathbf{X})$ is equal to the error in the initial estimation.*

P3 *The PRESS according to LnSO attainable by any PCA model is lower bounded by the sum of squares of the non-redundant information in the data.*

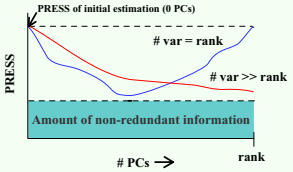


Figure 1: *Typical examples of PRESS by LnSO. For $\#var = rank$, all variables satisfy P2 and the PRESS for both $\#PCs = 0$ (the initial est.) and $\#PCs = rank$ coincide. For $\#var >> rank$, most variables satisfy P1 instead of P2. In any case, the curves remain above the amount of non-redundant information (P3).*

2. Imputation of missing data (model exploitation)

Aim: To minimize the error of estimation of future missing data.

2.1 Leave-n-Samples-Out 2 (LnSO2) method

According to **P1**: **For full rank, the error of estimation with LnSO of a variable containing only redundant information -which can be recovered from the other variables- is not null as it should.** To overcome this limitation, the core of the LnSO algorithm should be repeated till convergence (LnSO2 approach).

2.1.1 Core

Repeat until $\mathbf{X}_h^\#$ converges
  $\mathbf{T}_a^\# = \mathbf{X}^\# \cdot \mathbf{P}_a^*$
  $\hat{\mathbf{X}}^\# = \mathbf{T}_a^\# \cdot \mathbf{P}_a^{*t}$
  $\mathbf{X}_h^\# = \hat{\mathbf{X}}_h^\#$
end

2.1.2 Properties

P1b *The error of estimation in L1SO2 of a variable containing only redundant information for a PCA model with $A = Rank(\mathbf{X})$ is equal to 0.*

P1b can be generalized to LnSO2 by properly selecting the groups of variables left out at the same time. P2 and P3 are true for LnSO2.

2.2 Imputation methods

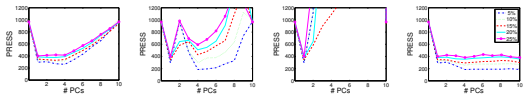
On the other hand, to improve the performance of LnSO, more sophisticated imputation methods may be used. Here Projection to Model Plane (PMP) [1] and Trimmed Score Regression (TSR) [2] have been considered.

2.3 Results

Table 1: *Minimum PRESS in the imputation of missing data for different approaches. The number of latent and observed variables of the simulated examples in parenthesis.*

Noise	First example (4/10)				Second example (12/27)				Third example (15/50)			
	LnSO	LnSO2	PMP	TSR	LnSO	LnSO2	PMP	TSR	LnSO	LnSO2	PMP	TSR
5%	262	186	291	**185**	884	878	878	**873**	814	784	784	**692**
10%	313	297	329	**256**	**1.017**	1.037	1.037	1.030	1.083	1.076	1.076	**981**
15%	328	346	346	**294**	**1.111**	1.147	1.147	1.138	1.351	1.356	1.356	**1.297**
20%	375	383	383	**355**	**1.264**	1.355	1.355	1.310	1.590	1.653	1.653	**1.574**
25%	401	400	400	**382**	**1.314**	1.461	1.461	1.408	**1.756**	1.835	1.835	1.760

First example (4/10) PRESS with LnSO, LnSO2, PMP-LnSO and TSR-LnSO.



Conclusion: The correction of LnSO in LnSO2 does not improve the imputation of missing data. Nonetheless, the use of TSR yields good outcomes (Table 1) and the best suited PRESS curve (Figure above on the right). If the number of PCs is overestimated, the imputation of missing data for TSR is not worsened in an important degree. This is not true for the other approaches, where the PRESS tends to increase fast under this situation.

3. Robustness to measurement noise

Aim: To identify the optimum number of PCs when data are corrupted with measurement noise.

3.1 Cross-validation Corrected Leave-n-Samples-Out (CCLnSO) method

Neither LnSO nor LnSO2 take into account the non-redundant information (P3). **What happens if important information is only contained in one observable variable???** This information will have no effect in the PRESS curve of LnSO or LnSO2, leading to an underestimated number of PCs. To correct this limitation, X is augmented with redundant information (CCLnSO approach):

3.1.1 Inner loop

$\mathbf{T}_a^\# = \mathbf{X}^\# \cdot \mathbf{P}_a^*$
$\mathbf{X}_{aug,a}^* = [\mathbf{X}^*, \mathbf{T}_a^*]$, remember not to scale $\mathbf{T}_a^*$
Fit a PCA model from $\mathbf{X}_{aug,a}^*$, obtaining $\mathbf{P}_{aug,a}^*$ and $\mathbf{T}_{aug,a}^*$
For each group of variables ($h = 1...H$)
  Set $\mathbf{X}_h^\# = 0$
  $\mathbf{X}_{aug,a}^\# = [\mathbf{X}^\#, \mathbf{T}_a^\#]$
  $\mathbf{T}_{aug,a}^\# = \mathbf{X}_{aug}^\# \cdot \mathbf{P}_{aug,a}^*$
  $\hat{\mathbf{X}}_{aug,a}^\# = \mathbf{T}_{aug,a}^\# \cdot \mathbf{P}_{aug,a}^{*t}$
  Restore its actual value to $\mathbf{X}_h^\#$
  $\mathbf{E}_{a,g,h} = \mathbf{X}_h^\# - \hat{\mathbf{X}}_{a,h}^\#$
end

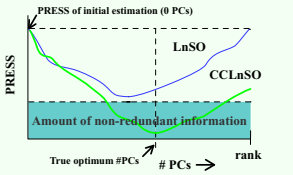


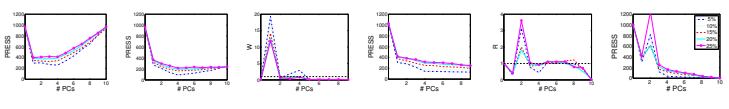
Figure 3: *Example of PRESS computed by LnSO and CCLnSO. The $\#PCs$ of minimum PRESS change since CCLnSO takes into account the non-redundant information.*

3.2 Results

Table 2: *Number of PCs detected by different approaches. R stands for the R-statistic of [3] and W stands for the W-statistic of [4]. The number of latent and observed variables in parenthesis.*

Noise	First example (4/10)				Second example (12/27)				Third example (15/50)			
	R	W	LnSO	CCLnSO	R	W	LnSO	CCLnSO	R	W	LnSO	CCLnSO
5%	1	**4**	**4**	**4**	6	6	**12**	**12**	10	12	13	**15**
10%	1	**4**	3	**4**	6	6	**12**	**12**	10	12	13	**15**
15%	1	1	3	**4**	6	6	**12**	**12**	10	12	13	16
20%	1	1	3	**4**	6	6	**12**	**12**	10	12	13	16
25%	1	1	1	**4**	6	6	**12**	**12**	10	12	13	17

First example (4/10) PRESS with LnSO, CCLnSO, W-statistic and PRESS in [4], R-statistic and PRESS in [3].



Conclusion: The CCLnSO method outperforms the other approaches (Table 2) in determining the optimum number of PCs for data corrupted with independent measurement noise, since it takes into account the non-redundant information. In the examples studied, the other approaches tend to underestimate the true number of PCs.

[1] Nomikos, P., and MacGregor, J.F. (1995), Multivariate SPC Charts for Monitoring Batch Processes, Technometrics, 37, 41-59.
[2] Arteaga, F. and Ferrer, A. (2002), Dealing with missing data in MSPC: several methods, different interpretations, some examples, Journal of Chemometrics, 16, 408-418.
[3] Wold, S. (1978), Cross-Validatory Estimation of the Number of Components in Factor and Principal Components, Technometrics, 20, 397-405.
[4] 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.

Simulations

Three simulated data matrices are used as experimental data. The three have different number of latent variables (LVs) and observable variables (OVs). The LVs are generated independently at random following a normal distribution of 0 mean and standard deviation 1. Each OV is obtained from one LV alone or as a linear combination of two or more LVs. All OVs are computed so that they have zero mean and unit variance. Notice that the data sets are composed of OVs alone.

Measurement noise is generated independently for each OV and at random, following a normal distribution of 0 mean. The standard deviation used depends on the percentage of noise chosen to be added to the data sets:

$$x'_i = (x_i + (\sqrt{\sigma_n}) \cdot n)/(\sqrt{1+\sigma_n})$$

where x'_i is the corrupted variable, x_i the noise-free variable, σ_n the standard deviation of the noise and n the noise generated. The data sets are corrupted with noise from 5% to 25%, where this percentage is computed so that the lowest standard deviation of a LV is the 100%.

For the **first data set**, 4 LVs were generated and from them, a total of 10 OVs in the following way:

$$x_i = (\sqrt{i/5}) \cdot lv_1 + (\sqrt{1-i/5}) \cdot lv_2, i\epsilon\{1,..,5\}$$
$$x_i = (\sqrt{0.5}) \cdot lv_1 + (\sqrt{i/10 - 0.5}) \cdot lv_2 + (\sqrt{1-i/10}) \cdot lv_3, i\epsilon\{6,..,9\}$$
$$x_{10} = ((\sqrt{0.01}) \cdot lv_1 + (\sqrt{0.01}) \cdot lv_2 + (\sqrt{0.01}) \cdot lv_3 + lv_4)/\sqrt{1.03}$$

where x_i stands for the i-th OV and lv_j stands for the j-th LV. This data set has been designed to present a first LV of very high variance. Also, it includes an OV composed almost completely of non-redundant information.

A simple analysis can be performed to observe how the PCA subspace is affected by the noise. A 4 PCs model is computed from the data sets corrupted with noise from 5% to 25%, obtaining loadings matrices $\{P_5...P_{25}\}$. A 4 PCs model is also computed from the noise-free original data, obtaining the eigenvectors $\{p_1, p_2, p_3, p_4\}$. Then, each of the latter are projected on the subspace spanned by each of $\{P_5...P_{25}\}$ and the percentage of explained variance is computed, so that 100% means perfect matching and 0% means non-correlation at all -i.e. the LV is not identified. The results are shown in the following table, where it can be seen that the PC subspace remains more or less the same in all the cases, since the biggest amount of variance lost in a LV is less than a 7%.

Eig.vec.	Eig.val.	5%	10%	15%	20%	25%
p_1	732.7	99.94	99.94	99.88	99.68	99.81
p_2	101.7	99.96	99.83	98.35	98.94	98.60
p_3	64.8	99.75	99.64	98.06	93.25	94.47
p_4	41.9	99.09	97.64	95.10	94.85	97.13

For the **second data set**, 12 LVs were generated and from them, a total of 27 OVs:

$$x_i = lv_j, i\epsilon\{1,..,12\}, j\epsilon\{1,..,12\}$$
$$x_i = (\sqrt{0.5}) \cdot lv_j + (\sqrt{0.5}) \cdot lv_k, i\epsilon\{13,..,27\}, j \neq k\epsilon\{1,..,6\}$$

This data set contains 6 OVs completely composed of non-redundant information. Again, an analysis to observe the effect of the noise in the PCA subspace is carried out in the following table. The results show the amount of variance lost per LV is not high (less than a 10%).

Eig.vec.	Eig.val.	5%	10%	15%	20%	25%
p_1	627.7	99.93	99.87	99.64	99.69	99.35
p_2	398.5	99.86	99.68	99.41	99.43	98.86
p_3	346.6	99.83	99.71	99.33	98.61	99.07
p_4	321.1	99.80	99.59	99.68	98.05	98.28
p_5	252.6	99.55	99.48	99.09	98.39	98.74
p_6	209.4	99.67	99.67	99.13	98.36	98.29
p_7	135.0	99.53	99.43	98.23	95.95	98.24
p_8	99.9	98.68	98.63	97.22	96.72	94.91
p_9	83.0	98.54	98.81	97.66	96.88	91.62
p_{10}	78.7	98.79	96.83	97.94	94.31	93.07
p_{11}	69.9	99.35	97.74	96.20	97.07	92.53
p_{12}	64.2	98.82	98.56	94.23	95.19	91.44

For the **third data set**, 15 LVs were generated and from them, a total of 50 OVs:

$$x_i = (\sqrt{0.5}) \cdot lv_j + (\sqrt{0.5}) \cdot lv_k, i\epsilon\{1,..,45\}, j \neq k\epsilon\{1,..,10\}$$
$$x_i = lv_j, i\epsilon\{46,47\}, j\epsilon\{11,12\}$$
$$x_{48} = (\sqrt{0.5}) \cdot lv_{11} + (\sqrt{0.5}) \cdot lv_{13}$$
$$x_{48} = (\sqrt{0.5}) \cdot lv_{12} + (\sqrt{0.5}) \cdot lv_{14}$$
$$x_{50} = lv_{15}$$

This data set contains 1 OV completely composed of non-redundant information and 2 OVs with half of its content non-redundant. An analysis of the effect of the noise in the PCA subspace is carried out in the following table. The results show that the amount of variance lost per LV is high for the last LV. Therefore, part of the structured information is located in the residuals. This may be partly the reason why the CCLnSO method is detecting a higher number of PCs than 15 for noise percentages form 15% to 25% (see the poster).

Eig.vec.	Eig.val.	5%	10%	15%	20%	25%
p_1	932.7	99.87	99.44	99.34	99.35	98.87
p_2	686.9	99.72	99.52	99.02	99.00	99.02
p_3	566.4	99.81	99.35	99.24	98.76	97.63
p_4	537.1	99.60	99.44	98.74	98.72	97.24
p_5	384.0	99.59	98.95	99.06	97.94	96.92
p_6	324.4	99.64	98.87	98.37	98.23	97.36
p_7	302.3	99.14	98.26	97.68	98.39	97.50
p_8	280.8	99.44	98.76	98.09	97.57	96.72
p_9	259.8	99.41	98.78	97.43	96.93	94.96
p_{10}	208.0	99.18	98.39	98.08	95.97	96.83
p_{11}	132.9	98.97	97.28	94.57	92.83	92.31
p_{12}	106.0	98.09	96.01	95.27	92.10	93.64
p_{13}	77.0	97.27	95.50	90.06	**86.34**	93.27
p_{14}	27.8	94.37	**81.98**	**77.03**	**69.35**	**72.39**
p_{15}	18.2	90.95	**78.51**	**82.06**	**29.44**	**47.40**