New Cross-Validation Methods in Principal Component Analysis

Cross-validation has become one of the most used methods to identify the number of significant principal components (PCs) in Principal Components Analysis (PCA) models. This is, to a large extent, thanks to the contribution of two notorious papers, namely (Wold, 1978) and (Eastment and Krzanowski, 1982). The approaches presented in both papers are based on the definition of a cross-validatory algorithm to compute the sum of squares of prediction error (PRESS) together with the definition of a statistical index. The number of significant PCs is detected when this index exceeds or falls below a certain threshold value.

An alternative approach was also suggested by Wold (1978). This approach presents an attractive feature: the number of significant PCs is detected when the PRESS reaches its minimum value. Therefore, no additional statistical index needs to be defined. This is, in principle, a more intuitive behavior for the prediction error: decrease as the addition of PCs improves the prediction performance of the model, and increase when this addition is noisy. Nonetheless, this approach has one drawback: PCs modelling independent variables do not reduce the PRESS. Therefore, these PCs are not recognized as significant, although they are -if these PCs are not included in the PCA model, the independent variables are simply not modelled-.

In this poster, two novel cross-validation algorithms, named fast corrected-leave-n-samples-out (fast-CLnSO) and corrected-leave-n-samples-out (CLnSO) (Camacho et al., 2007) are presented. These algorithms are based on the alternative approach of Wold (1978), overcoming its limitation to detect significant PCs modelling independent variables. These novel algorithms outperform the other well-known approaches, yielding a 100% of effectiveness in determining the correct number of significative PCs in all the simulated data sets studied, for measurement noise levels up to a 30% and 40%. The algorithms are also tested with a real data set used in (Wold, 1978) and (Eastment and Krzanowski, 1982).

References:

