A new algorithm for selecting the unfolding method and the sub-models in batch process modelling with PCA

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

When modelling batch process data with Principal Component Analysis (PCA), one of the most critical decisions is how to arrange the three-way data in two dimensions. Rather than using the same modelling approach always, evaluating which arrangement of the data is appropriate for a specific process may be more advantageous. The aim of this poster is to define a general mechanism to compare PCA models calibrated from different arrangements of the data.

Some Notation

Let us define $\underline{\mathbf{X}}(I \times J \times K)$ as the process data matrix collected from a batch process and aligned, which contains the values of J variables at K sampling times in I batches. To apply PCA, these data have to be rearranged in two dimensions.

Unfolding

The batch dynamic unfolding can be expressed as:

$$\mathbf{X} = unfold(\underline{\mathbf{X}}, \kappa) \equiv \underline{\mathbf{X}}^{(\kappa)} \quad (1)$$

where κ stands for the number of lagged measurement vectors (LMVs) and:

$$\kappa = \{k-1 \ : \ k \ \epsilon \ \{1, 2, .., K\}\} \quad (2)$$

Therefore, the batch-wise unfolding is:

$$\mathbf{X} = \underline{\mathbf{X}}^{(K-1)} \quad (3)$$

and the variable-wise unfolding:

$$\mathbf{X} = \underline{\mathbf{X}}^{(0)} \quad (4)$$

Number of sub-matrices

Let $\underline{\mathbf{X}}_{k_i:k_e}$ contain the data of $\underline{\mathbf{X}}$ from sampling time k_i to k_e. One way to arrange $\underline{\mathbf{X}}$ in two dimensions is to divide the data in K Local sub-matrices:

$$\mathbf{X} = \{\mathbf{X}_k : k = 1, ..., K\} \quad (5)$$

Combining the unfolding with the division in several sub-matrices, many other approaches can be specified. For instance, the Evolving approach:

$$\mathbf{X} = \{\underline{\mathbf{X}}_{1:k}^{(k-1)} : k = 1, ..., K\} \quad (6)$$

or the Moving Window approach:

$$\mathbf{X} = \{\underline{\mathbf{X}}_{k-d:k}^{(d)} : k = d+1, ..., K\} \quad (7)$$

Generalized PCA model of batch data

A generalized arrangement in two dimensions can be defined as:

$$\mathbf{X} = \{\underline{\mathbf{X}}_{\phi_l}^{(\kappa_l)} : l = 1, ..., L\} \quad (8)$$

with:

$$\phi_l = k_{il} : k_{el}$$

$$\kappa_l = \{k-1 \ : \ k \ \epsilon \ \{1, 2, .., k_{el}\}\} \quad (9)$$

$$s.t. \ k_{il} \leq k_{el}, \ k_{i1} = 1, \ k_{eL} = K$$

Finally, a generalized PCA model of a batch process is completely specified by:

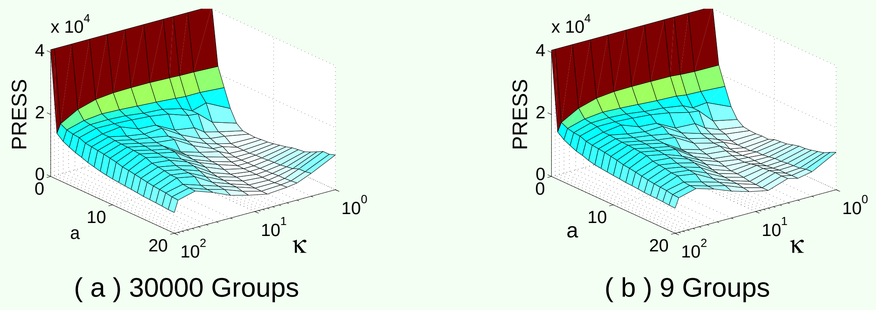
$$M = \{PCA(\underline{\mathbf{X}}_{\phi_l}^{(\kappa_l)}, a_l) : l = 1, ..., L\} \quad (10)$$

where a_l is the number of PCs of sub-model l.

Evaluation Algorithm

For each sub-model ($l = 1...L$)
  $\mathbf{X} = \underline{\mathbf{X}}_{\phi_l}^{(\kappa_l)}$
  Calibrate a PCA model of a_l PCs from $\mathbf{X}$, obtaining $\mathbf{P}_{a_l}$ and $\mathbf{T}_{a_l}$
  For each group of batches ($b = 1...B$)
    Form $\mathbf{X}^*$ and $\mathbf{T}^*_{a_l}$ with data from all groups but b
    Form $\mathbf{X}^\#$ and $\mathbf{T}^\#_{a_l}$ with data from b
    $\mathbf{X}^*_{aug} = [\mathbf{X}^*, \mathbf{T}^*_{a_l}]$, remember not to scale $\mathbf{T}^*_{a_l}$
    Calibrate a PCA model of a_l PCs from $\mathbf{X}^*_{aug}$, obtaining $\mathbf{P}^*_{aug,a_l}$ and $\mathbf{T}^*_{aug,a_l}$
    For each sample group of variables ($h = 1...H$)
      Set $\mathbf{X}^\#_h = 0$
      $\mathbf{X}^\#_{aug} = [\mathbf{X}^\#, \mathbf{T}^\#_{a_l}]$
      $\mathbf{T}^\#_{aug,a_l} = \mathbf{X}^\#_{aug} \cdot \mathbf{P}^*_{aug,a_l}$
      $\hat{\mathbf{X}}^\#_{aug} = \mathbf{T}^\#_{aug,a_l} \cdot \mathbf{P}^{*t}_{aug,a_l}$
      Restore its actual value to $\mathbf{X}^\#_h$
      $\mathbf{E}_{b,h} = \mathbf{X}^\#_h - \hat{\mathbf{X}}^\#_h$
    end
  end
  Fold back matrix $\mathbf{E}$ yielding $\underline{\mathbf{E}}$
  $PRESS_l = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=k_{il}}^{k_{el}} e^2_{i,j,k}$
end

Improving the computational efficiency



(a) 30000 Groups (b) 9 Groups

Saccharomyces Cerevisiae Cultivation.

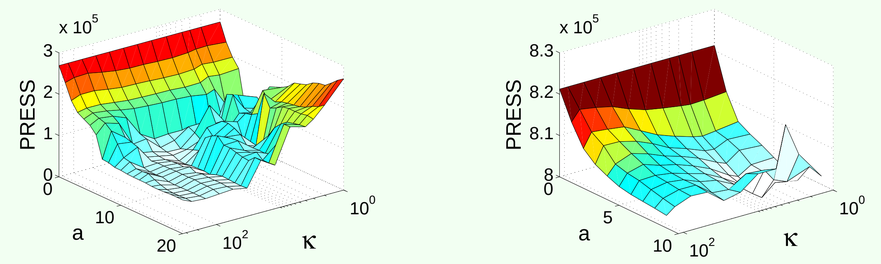
In order to make the algorithm efficient, it is necessary to reduce the number of iterations. This is accomplished by reducing the number of groups of batches (B) and of variables (H). The elements -batches or variables- belonging to a group should be chosen randomly. It was observed that **if this random selection is maintained for all the models which are compared, B and H can be reduced to very low values (from 3 to 7) with almost no loss of comparison performance.**

On the left, an example with the data of a simulated process, the cultivation of Saccharomyces Cerevisiae, is shown. Single models with different number of LMVs (κ) and PCs (a) are compared using the algorithm presented. The figure on the left (a) shows the result of a left-one-out approach. The figure on the right (b) is computed for $B = 3$ and $H = 3$. Both shapes are very similar, but the latter has been computed more than 10 times faster.

Comparing models with different unfolding methods

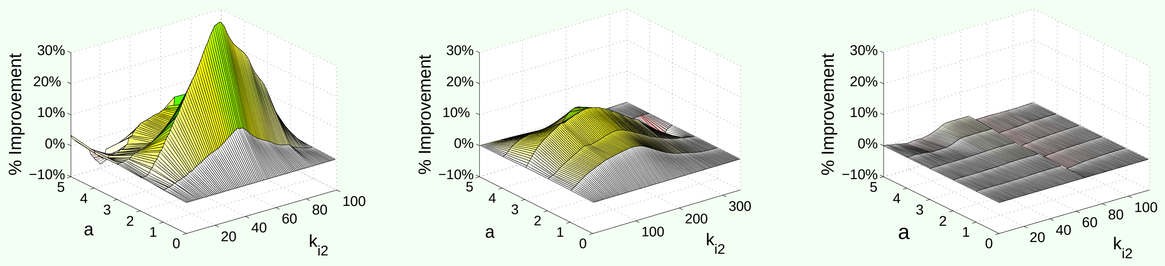
The folding operation of the error matrix is of upmost importance. If $0 < \kappa < K$, several error values corresponding to the same measurement-vector appear. **Only the first appearance -in the first raw of the two-way matrix- should be taken into account.**

On the right, the data from two real processes is studied. Once again, single models with different numbers of LMVs (κ) and PCs (a) are compared using the algorithm presented and $B = 3$ and $H = 3$. Since the shape of the PRESS is different for the three processes under study, it can be concluded that **the optimum model structure (in terms of the number of LMVs and PCs) is dependent on the nature of the process.** For instance, for the waste-water treatment process (a), a single model should present a high κ value, contrarily to what happens with the Nylon 6'6 polymerization process (b).



(a) Waste-water treatment (b) Nylon 6'6 polymerization

Division in sub-models



(a) Saccha. Cerev. Cult. (b) Waste-water treatment (c) Nylon 6'6 polymerization

When modelling with several sub-models, the PRESS of the joint model can be computed by simply adding the PRESS computed for the sub-models with the algorithm.

On the left, the percentage of reduction of PRESS due the division in two of a variable-wise model ($\kappa = 0$) is studied. The PRESS is shown for different PCs (a) and location of the 2 sub-models -i.e., different values of k_{i2}. In (a) the improvement exceeds the 40%, in (b) it only reaches a 10% and in (c) it is negligible. Also, the optimum location is different for the processes.

Therefore, **The optimum model structure (in terms of number of sub-models and location of these) is dependent on the nature of the process.**

The optimum arrangement in two dimensions to apply PCA to batch data is dependent on the process under analysis. The algorithm proposed in this poster can be successfully used to evaluate and compare different arrangements of a specific data set.