Modelling and Identification of Individual Stage Contributions in an Industrial Pharmaceutical Process by Multiblock PLS

Lígia P. Brás1, João A. Lopes1, Carlos R. Santos2, Joaquim P. Cardoso2, José C. Menezes1,*

1 Centre for Biological and Chemical Engineering (bsel@ist.utl.pt)
Technical University of Lisbon, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
2 CIPAN SA, Vala do Carregado, P-2601-906 Alenquer, Portugal

Abstract
In this paper, multiblock partial least squares (MBPLS) was applied to model a multi-stage pharmaceutical process producing an active pharmaceutical ingredient (API) by fermentation. The aim is to model the final API’s chromatographic purity in terms of each operating variable, establishing the relative contribution of each process stage. A balanced bootstrap algorithm was used to choose the best set of variables from each stage to obtain a model with the highest predictive capability. Seven variables from the original twenty-eight were selected and a two-latent variable MBPLS model was found to minimize the prediction residuals ($Q^2 = 64.3\%$). The multiblock methodology revealed that the fermentation stage has a prominent role in the description of the final API’s chromatographic purity.

Keywords: multivariate data modelling, multiblock PLS, pharmaceutical production

1. Introduction
The majority of biochemical processes consist of a sequence of unit operations that are usually run in a defined order in batch, semi-batch, or continuous mode. For economic reasons, industrial pharmaceutical processes are usually batch operated. Active pharmaceutical ingredients (APIs) are often produced by fermentation (Vandamme, 1984). Recently, Lopes et al. (2002) investigated the relative contributions of the inoculum production and fermentation stages in the industrial production of an API by fermentation of a *Streptomyces* strain. Using the concept of multiblock partial least squares (MBPLS) proposed by Weterhuis et al. (1998), Lopes et al. separated the information from the two production stages into different blocks that were used in modelling the process. They pointed out that inoculum’s quality variables were highly correlated with the API’s production.

Taking into account additional information about the process, MBPLS methodology divides the descriptors X into meaningful blocks of variables ($X_1, X_2, \ldots, X_n$) that are then related to the response block. As a result, MBPLS methods can improve models’
interpretability and facilitate the description of complex multi-stage systems. Thus, MBPLS methods are potentially very useful in modelling and monitoring a variety of chemical processes (Weterhuis et al., 1998; MacGregor et al., 1994).

In this paper, we extend the application of the MBPLS approach to an industrial pharmaceutical process by considering not only the API’s production stages (inoculum production and fermentation), but also the API’s isolation stages (filtration, extraction and precipitation). The objective is to model the chromatographic purity of the API at the end of the multi-stage process and to evaluate the relative contribution of each process stage for the obtained purity. In order to optimise model predictions, a balance bootstrap resampling strategy was used to estimate the optimal set of predictors. For confidentiality reasons the API and process variables names and absolute values cannot be disclosed.

2. Methods

2.1 Data

Figure 1 illustrates the stages involved in the API’s production phase (inoculum production and fermentation) and the API’s isolation phase (filtration, extraction and precipitation). A total of ninety-five industrial sequential batches were considered. These batches were operated under nominal operating conditions. Consumption of raw materials, operation yields, quality parameters of intermediate operation streams and other relevant process parameters measured at the end of each batch (twenty-eight variables) were separated into meaningful blocks in terms of process engineering knowledge. The response variable was API’s chromatographic purity measured at the end of the precipitation stage, since this quality parameter influences subsequent pharmaceutical process performance.

2.2 Multiblock PLS

Multiblock PLS is an extension of the PLS method, a class of regression models motivated by the attempt to find the relationship between explanatory and response variables by assuming that they are generated by a common set of underlying factors (Gerlach et al., 1979; Martens and Naes, 1989).

![Figure 1. Schematic representation of the initial stages of the industrial pharmaceutical process considered.](image-url)
In MBPLS, the predictors are separated into subsets or blocks, accordingly to a meaningful criterion or process knowledge. For prediction purposes the variables should be combined in a single block, as in conventional PLS (Geladi, 1988). The method employed in this work was proposed by Westerhuis and Coenegracht (1997). As in PLS the following equations apply to MBPLS method:

\[
X = T \ P^T + E \tag{1}
\]

\[
y = T \ q^T + f \tag{2}
\]

In the previous equations, X is the matrix of predictors, y is the response vector, E and f are the residuals of X and y, respectively, T are the super-scores and P and q are the independent and dependent data loadings, respectively. The super-scores T contain the block-scores associated with each independent block, therefore summarizing the information present in all blocks. Predictions (ŷ) are obtained merging the data blocks into a single X matrix, applying the calibration scaling factors and finally the regression vector (b_{PLS}):

\[
\hat{y} = X \ b_{PLS} \tag{3}
\]

Equation 4 provides a measure of the fitting between model predicted responses and experimental data. To investigate the optimum number of latent variables to use in each model, a strategy based on the “leave-one-out” cross validation method was applied.

\[
Q^2_Y = 1 - \frac{\text{trace} \ ((y - \hat{y})^T (y - \hat{y}))}{\text{trace} \ (y^T y)} \tag{4}
\]

**2.3 Model optimization**

A resampling strategy based on balanced bootstrap was used to estimate the best set of explanatory variables (Manly, 1998). Balanced bootstrap is an improvement of the simple bootstrapping resampling where it is ensured that each observation in the original sample occurs the same number of times in the bootstrapped samples. Five-thousand bootstraps were used to generate the distribution of each PLS model regression coefficient (b_{PLS,i}). A 95% confidence interval for the true value of each b_{PLS,i} was estimated by the two values that encompass the central 95% of the generated distribution. If the zero value was included in the confidence interval determined for a given predictor, then that variable was discarded from the model.

**3. Results**

To evaluate the importance of each stage in the description of the response variable, different PLS models were built by sequentially adding the corresponding operation predictors to the global X-block. This strategy was proposed in order to simulate the knowledge that becomes available as each of the unit operations are computed in the multi-stage process. Figure 2.a) summarizes the results obtained. The most significant increase in model’s performance takes place when the fermentation stage (block 2) is introduced. The inclusion of the other blocks provides minor performance increases.
with the exception of the extraction stage (block 4), which has no relevant role in the final purity, under the present nominal operating conditions used in this stage. This is not surprising since the fermentation stage dictates the conditions for the subsequent separation steps, which in turn will have repercussions on the final API’s quality.

The balanced bootstrapping methodology was applied to each model obtained after sequentially adding an operation block. Results are depicted in Figure 2.b) showing the improvements in models’ predictions. Regarding to process stage contributions, a similar pattern was observed before and after variable selection.

Figure 3.a) represents the 95% confidence limits determined for the regression vector of the five-block PLS model obtained using the original twenty-eight predictors. For many variables, the confidence interval calculated contains the zero value, which means that these variables are not significant to describe the variance in \( y \). Therefore, they were eliminated and a new model was constructed, and again submitted to variable selection by bootstrapping. This procedure was repeated until all variables were considered significant.

Following the variable selection strategy, the final MBPLS model was found using seven of the original twenty-eight process variables and two-latent variables (Figure 2.b). The entire extraction stage (block 4) was excluded by the selection procedure, which was supported by the insignificant contribution of this block, as noticed above. The regression coefficients for the final model and the respective 95% confidence intervals are depicted in Figure 3.b). The highest regression coefficient was observed for variable \( x_5 \) (inorganic nitrogen concentration), which belongs to the fermentation stage. Block-weights calculated for the MBPLS model provide the individual contribution of each process stage (Table 1). The fermentation stage has a prominent role in the description of \( y \) (with strong contributions in both latent variables). This result is consistency with the evidence in Figure 2.

![Figure 2](image-url)
Figure 3. Regression coefficients determined for the PLS model comprising all the original predictors (a) and for the final two-latent MBPLS model (b). The line bars correspond to the 95% confidence limits determined by balanced bootstrapping.

Table 1. Individual process stage contribution to the overall performance of the final optimised two-latent variable MBPLS model.

<table>
<thead>
<tr>
<th>Block</th>
<th>Process Stage</th>
<th>MBPLS super-weights (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LV 1</td>
</tr>
<tr>
<td>1</td>
<td>Inoculum production</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>Fermentation</td>
<td>38</td>
</tr>
<tr>
<td>3</td>
<td>Filtration</td>
<td>26</td>
</tr>
<tr>
<td>5</td>
<td>Separation</td>
<td>20</td>
</tr>
</tbody>
</table>

Figure 4. Experimental (open squares) and predicted API's chromatographic purity (filled circles) from the final two-variables MBPLS model. Line bars are 95% prediction intervals determined by balanced bootstrapping.
The predicted variance of the API’s chromatographic purity with the proposed model is only 64.3% (obtained by leave-one-out cross validation). However, the model is able to accurately describe the observed trend and be accurate within confidence limits. Figure 4 compares the experimental and predicted responses. The line bars correspond to the 95% predictions intervals determined by balance bootstrapping constructed from the leave-one-out cross validated predictions of five thousand MBPLS models.

4. Conclusions

The multiblock methodology revealed that the fermentation stage has a prominent role in the description of API’s chromatographic purity measured at the end of a five-stage industrial pharmaceutical process. All the other considered process stages have a minor but relevant contribution to the model, with the exception of the extraction step which influence can be neglected. By identifying the more influential stages and parameters of the industrial process with respect to a key quality parameter, this approach enabled a rapid characterization of the process. The knowledge obtained with this methodology can now be used to enhance process’ performance (increasing process quality, productivity and reducing variability) by enabling the process engineers to focus on specific process stages and variables.

Acknowledgements

The authors would like to thank Companhia Industrial Produtora de Antibióticos SA (CIPAN) in Portugal for providing the data and for permission to publish these results. Ligia Brás thanks Foundation for Science and Technology in Portugal for financial support (POCTI BD/10302/2002).

References

Vandamme E., 1984, Biotechnology of Industrial Antibiotics, Marcel Dekker, New York.