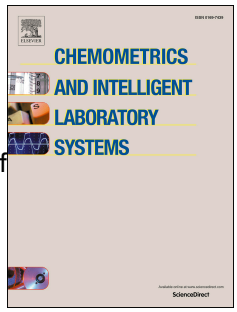
# Accepted Manuscript




A computational approach to partial least squares model inversion in the framework of the process analytical technology and quality by design initiatives

S. Ruiz, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez

PII: S0169-7439(17)30496-3

DOI: 10.1016/j.chemolab.2018.08.014

Reference: CHEMOM 3674

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 26 July 2017

Revised Date: 26 August 2018

Accepted Date: 27 August 2018

Please cite this article as: S. Ruiz, M.C. Ortiz, L.A. Sarabia, M.S. Sánchez, A computational approach to partial least squares model inversion in the framework of the process analytical technology and quality by design initiatives, *Chemometrics and Intelligent Laboratory Systems* (2018), doi: 10.1016/j.chemolab.2018.08.014.

# A COMPUTATIONAL APPROACH TO PARTIAL LEAST SQUARES MODEL INVERSION IN THE FRAMEWORK OF THE PROCESS ANALYTICAL TECHNOLOGY AND QUALITY BY DESIGN INITIATIVES

S. Ruiz[1], M.C. Ortiz[2*], L.A. Sarabia[1], M.S. Sánchez[1]

[1]Dpt. Mathematics and Computation, Faculty of Sciences, Universidad de Burgos

[2]Dpt. Chemistry, Faculty of Sciences, Universidad de Burgos

Plaza Misael Bañuelos s/n, 09001 Burgos, Spain

**Abstract**

In the context of the paradigms founding the Quality by Design and Process Analytical Technology initiatives, the work herein presents a computational approach to support the decision-making process, in particular, about the feasibility of a product defined for some a priori given quality characteristics.

The approach is based on the computation of the pareto-optimal front when simultaneously minimizing the expected differences between the predicted and the desired characteristics. Thus, the feasibility is tackle as an optimization problem with the novelty of doing so simultaneously for all the characteristics, preserving the correlation structure, but by handling each individual characteristic separately.

With data from a low-density polyethylene production process, with fourteen process variables and five measured characteristics of the final polyethylene, solutions are found to define the Design Space for targeted quality characteristics on the product, and without the need of explicitly inverting the PLS (Partial Least Squares) prediction model fitted to the process.

**Keywords**: Process Analytical Technology; Quality by Design; Partial Least Squares; Pareto optimality; process decision making; industrial processes

## 1. Introduction

The need of more organised approaches to process and product development with the aim of consistently guaranteeing quality and value to processes and products has been formally identified [1] by the pharmaceutical industry as the Quality by Design (QbD) initiative, but the concept can be extended to any industrial process, in particular, chemical processes.

In any case, the implementation of these concepts requires the quantitative characterization of process and product performance and, also, some modelling techniques that support decision-making, usually, multivariate models that relate the input variables (process parameters, material properties, etc.) and product quality (measured characteristics of the final product), which is the setting of the so-called Process Analytical Technology (PAT). In fact, in [2] it is already stated that the identification of optimal operating conditions for large

* Corresponding author: mcortiz@ubu.es

and complex chemical processes is done using advanced mathematical tools largely developed by the process systems engineering community.

The fitting of such prediction models is usually implemented via latent variable modelling techniques by exploitation of historical databases from experiments, current manufacturing process or historical products. The relevance of such a model, in the words of Ottavian et al. [3], is that coupling the information provided by the model to engineering knowledge about the process represents a formidable tool to advance the knowledge on the underlying physical and chemical phenomenon occurring in the process, therefore significantly contributing to improve the overall process understanding.

Besides, such latent variable models, specifically a PLS (Partial Least Squares) model is able to capture the correlation among quality characteristics, and also the correlation across process variables and quality characteristics. An additional advantage is that we can use the residual Q and Hotelling's $T^2$ statistics to assess the adequacy of a sample before predicting its expected quality characteristics.

Once a validated model to predict product characteristics is available, the focus turns to the desired quality characteristics, those that define a target product. Consequently, the issue now is that these characteristics should be somehow reverted into the necessary process variables, if any, that would produce the target product. Here, 'process variable' is used in general terms, referring to controllable factors but also to other factors that influence the process such as raw material properties, process parameters, processing conditions, etc., which may not be directly controllable but, in general, are controlled.

In this context, the procedure of looking for the process variables that will produce a product with desired characteristics is known as latent variable model inversion (LVMI). Much of the effort invested in this regard involves the study of the feasibility of a given product by studying the conditions for mathematically inverting the system, which is related to the 'effective' dimension of both the latent space and the responses space [4], and ensuring that the picked solution matches the correlation structure among quality characteristics [5]. A detailed explanation as well as a good summary of procedures and examples in the pharmaceutical field can be found in [3] and a summary of references (commented) up to 2012 in the supplementary material in [6].

In the direct mathematical matrix inversion scenario, it is not possible to add constraints in the desired quality characteristics. With that purpose, the problem can be re-defined as an optimization problem. For example, looking for the scores that minimize the quadratic weighted difference between the desired and the predicted characteristics, provided a limiting value of Hotelling $T^2$ is not surpassed in [7]. In [6] a new statistic is proposed to check the ability of the model to reconstruct the regressor vector.

In this optimization context, the problem is then posed as a least squares problem, linear if the predicting models is linear, like in [5,8], or nonlinear in [9] where GP (Genetic Programming) nonlinear models relating latent vectors (PCA –Principal Component Analysis scores) for **X** and **Y** blocks are used as predicting models, solving also a weighted least squares problem.

The alternative approach presented herein is to pose the study of the feasibility of a given product also as an optimization problem, by directly looking for the process variables that predict final product characteristics close enough to the desired ones. However, this is done

simultaneously for all the characteristics (the product is unique) but by handling each individual characteristic separately.

Both goals can be achieved when estimating the pareto-optimal front (see for instance [10]) that contains the best possible solutions for each variable independently and trade-off solutions among characteristics when the product is unfeasible. As a consequence, the PLS regression model is computationally inverted which, as far as the authors know, is done for the first time.

The proposed procedure is explained in section 2 trying to keep the theory to the minimum necessary, and then section 4 shows some of its characteristics. The first two cases, sections 4.1 and 4.2, address the computational inversion from the latent variable space or from the process variable space, respectively. Section 4.3 tackles the feasibility of a new product, and section 4.4 shows some other possibilities of the methodology, as far as they can be posed as optimization problems.

## 2. Model-base optimal selection of process variables

### 2.1 Optimization engine

It has already been said that the procedure is mainly an optimization procedure for looking for process variables that predict a final product with some previously defined quality characteristics. To clearly state the procedure, we divide it into the four necessary components [2], namely, a predictive model, that is, a mathematical expression that relates process variables with product characteristics, which is usually fitted with a process-representative data set; an objective function, which is a quantitative measure of the adequacy of a candidate solution; some constraints related to both process variables and product characteristics, mainly the former; and an optimization engine to search among the candidate solutions to find the optimal ones according to the defined objective function, while adhering to the constraints.

In more precise terms, let **Y** be a $(n \times q)$ matrix that contains the set of $q$ response (output) variables, that is, $q$ measured properties on the final product that define its quality characteristics. Matrix **X** $(n \times p)$ on the other hand gathers the available set of $p$ predictor input variables, what we called the process variables i.e., the variables that describe the situation of the process that gave rise to the product whose characteristics are recorded in **Y**.

These data matrices are used to fit a multivariate prediction model, let it be denoted as $f$, to predict **Y** from **X**. It should be noted that $f$ is a functional mapping between a $p$-dimensional space and a $q$-dimensional space, that is, a vector multivariate function, defined in some precise domain, *D,* the one determined by the controlled process variables. Notice that defining *D* is in fact a form of defining constraints on the process variables.

Then, given a target vector $\mathbf{y_t}$ that describes some desired characteristic, the feasibility of such product must be studied and, if this is so, the predictor vector $\mathbf{x} \in D$ with the corresponding process variables should be estimated. In other words, it is about studying the existence of some viable $\mathbf{x}$ such that

$$\hat{\mathbf{y}} = f(\mathbf{x}) \tag{1}$$

is, in practice, equal to the given $\mathbf{y_t}$. Mathematically speaking, this is $\mathbf{x} = f^{-1}(\mathbf{y}_t)$, provided $f$ is invertible and $f^{-1}$ denotes its inverse.

As an alternative to the explicit inversion of the prediction model $f$, the approach presented here is a computational approach that consists on taking into account its optimization nature, that is, to look for $\mathbf{x} \in D$ such that the predicted $\hat{\mathbf{y}}$ (Eq. (1)) and the target $\mathbf{y_t}$ are similar enough.

Given that we have more than one quality characteristic, the mentioned similarity is, in fact, between two vectors. In the present case, it is computed as the absolute difference between each coordinate, that is, each characteristic is individually considered and compared to the corresponding characteristics in the predicted vector. Consequently, we look for $\mathbf{x} \in D$ such that the predicted value $\hat{\mathbf{y}} = f(\mathbf{x})$ minimizes all these differences. More formally, the problem can be written as

$$\underset{\mathbf{x}}{\text{minimize}}\{d_i, i = 1,...,q\} = \underset{\mathbf{x}}{\text{minimize}}\left\{\left|\hat{y}_i - y_{t,i}\right|, \quad i = 1,...,q\right\} \tag{2}$$

where $\hat{\mathbf{y}} = f(\mathbf{x})$ for a given $\mathbf{x} \in D$ and multivariate prediction model *f*.

Eq. (2) defines the function to be minimized, which is a vector function. This is so because there are *q* differences (one per coordinate) to be simultaneously minimized. In a general context, it is said that there are multiple individual objectives to be met in the optimization problem, or multiple criteria to be optimized simultaneously, and as such, a problem like this one is known as a multi-objective or multicriteria minimization problem.

However, a multidimensional space, such as the space of the individual differences in Eq. (2), is not well ordered (see, for instance [11]) and, consequently, the optimization problem is no longer well defined in the sense that, in general, there might not be a solution that simultaneously provides the minimum values attainable for all the objectives (i.e., all the vector coordinates in Eq. (2)) at once.

This approach to the optimization problem, posing it as a vector function to be optimized, already makes a huge difference with the alternative of transforming the multiple different criteria into a single-criterion (for example, with a weighted sum), which is then optimized.

In any case, to tackle the problem, a partial order useful in this vector optimization context is the Pareto order which is defined, for two *q*-dimensional vectors, as

$$(d_1, d_2, ..., d_q) \prec (d'_1, d'_2, ..., d'_q) \quad \text{when } d_i \le d'_i \ \ \forall i = 1, 2, ..., q \tag{3}$$

It seems clear that any pair of vectors are not necessarily comparable with this relation, but, when the vectors are comparable, the preferred solution for a minimization problem is the one that gives $(d_1, d_2, ..., d_q)$.

For a finite set of vectors, those that are not comparable to one another with the Pareto order constitute the set of non-dominated solutions. When considering the set of non-dominated solutions of the entire feasible space we have the so-called Pareto optimal front. In a multicriteria optimization context the Pareto-optimal front thus contains the Pareto-optimal

solutions for each given criterion, that is, when moving along the front necessarily one of the criteria is improved though at least another one is worsened. In particular, it contains the solutions that provide the minimum attainable value in each individual criterion.

The optimization engine to look for the Pareto-optimal front is a computational approach by using an evolutionary algorithm, primarily because we need to maintain a population of solutions. Contrary to the usual need of convergence of all the members in a population towards a single optimum value, in this case, we need to estimate the whole pareto-optimal front thus the population must converge to different solutions spread along this front, so non-dominance among solutions and maintenance of the dispersion are the two main characteristics to be taken into account along evolution.

Hence, the population is made up by points in the domain of the process variables, *D*, evaluated (vector fitness function) according to the individual proximity to $\mathbf{y}_t$ and it evolves by moving these points inside *D* looking for non-dominated vectors (in the *q*-dimensional space) according to the criteria to be optimized, in this case, the minimization problem as defined in Eq. (2), with the partial order among differences defined in Eq. (3).

In each generation of the evolution, a new population is created, as usual, by selection, crossover, and mutation; and the new elements are evaluated according to the criteria to obtain the corresponding fitness value. Then, by considering the enlarged population formed by joining the current population (parents) and the new generated offsprings, the update is made by firstly sorting all of them in sequential levels of non-dominance and selecting among these levels in order until completion of the new population. If in the last level we need to arrive to fill the population there is more solutions than needed, only the most dispersal ones are selected according to the crowding distance [12]. More details about the implementation can be found in [13], which is in the context of experimental design. In this field, a similar methodology allowed the computation of ad-hoc experimental designs with competing criteria [14,15].

### 2.2 Prediction model and constraints

As for the regression method, we will use a PLS2 model between predictor matrix **X** with values of *p* process variables and response matrix **Y** that contains the measured *q* quality characteristics.

This adds constraints for the optimization problem posed in eq. (2) further to the constraint already written ($\mathbf{x}$ should be in $D$). To check consistency with the latent space defined by the PLS2 model, critical values for a given confidence limit (0.95 along the present work) for both Q and $T^2$ statistics cannot be surpassed for any candidate vector of process variables $\mathbf{x}$.

The computation of these statistics are as follows:

Given a PLS2 model with *k* latent variables, the relation between the model and **X** via loadings $\mathbf{P}_k$ and scores $\mathbf{T}_k$, is described by the following equation (as usual, upper t means transposing)

$$\mathbf{X} = \mathbf{T}_k \mathbf{P}_k^t + \mathbf{R} \qquad (4)$$

If $\mathbf{r}_i$ is the i-th row of residual matrix **R** in eq. (4), the $Q = \mathbf{r}_i \mathbf{r}_i^t$ residual index indicates the difference or residual between the value of the i-th sample and its projection on the subspace

of the model. Its critical value at an α confidence level is obtained by the Jackson-Mudholkar method [16].

The $T_i^2$ Hotelling statistic for the i-th sample is the sum of the normalized squared scores, defined as $T_i^2 = (\mathbf{t}_i\, \mathbf{t}_i^t) / \lambda_i$ where $\mathbf{t}_i$ is the i-th row of $\mathbf{T}_k$ in eq. (4) and $\lambda_i$ the eigenvalue corresponding to the i-th latent variable. $T_i^2$ is a measure of the Mahalanobis distance from each sample to the centroid, measured in the projection plane (hyperplane) of the PLS2 model considered. At a confidence level α the threshold value is computed as $(k\,(n-1) / (n-k))\, F_{k,n-k,\alpha}$, where $F_{k,n-k,\alpha}$ is the critical value of a F distribution with $k$ degrees of freedom in the numerator and $n-k$ in the denominator.

Finally, if the quality characteristics are unequally important, the procedure can be also applied with weighted responses, by using weights defined by the user according to the different relevance of each response.

**3. Software**

The main routines implementing the evolutionary algorithm to estimate the pareto-optimal front are home-made programs written in Matlab® (The Mathworks, Inc., Natick, MA, USA). The PLS_Toolbox (Eigenvector Research, Inc., Manson, WA, USA) for Matlab has been used for all the calculations with PLS models. Also, Statgraphics Centurion XVII (Statpoint Technologies Inc., The Plains, VA, USA) has been occasionally used.

**4. Results and discussion**

To illustrate the methodology, data [17, 18] from a low-density polyethylene production process is used. They correspond to two tubular reactors connected in series, so that the process depends on 14 (process) variables. Data from 50 products are recorded, with the corresponding fourteen process variables and five characteristics measured in each produced polyethylene to qualify it.

Table 1 summarizes the process variables, as well as their mean and standard deviation values. Also, minima and maxima are in Table 1 because they define the searching domain $D$, that is, the viable settings of the process variables. The last rows in Table 1 contain the same information for the five quality characteristics.

The differences in magnitude and variability are noticeable for both the process variables in $\mathbf{X}$ ($50 \times 14$) and the product characteristics in $\mathbf{Y}$ ($50 \times 5$).

Now, the focus is on the quality characteristics, which in this case are defined for each product in the five-dimensional space. In this space, the variation of each variable in the dataset is small (which is not surprising for a controlled process), the maximum and minimum values of each characteristic in $\mathbf{Y}$ in Table 1 show that the range of each response variable is no larger than 9.7% of the corresponding mean.

Besides, some of them are highly correlated as can be observed in the lower triangular part of the correlation matrices written in Table 2. An opposite behaviour of $Y_2$ with respect to both $Y_4$ and $Y_5$ is expected due to the high negative correlation coefficients, namely -0.80 and -0.92, respectively, that appear in the second column in Table 2, contrary to the situation with

$Y_1$ which is positively correlated with them (0.88 and 0.74 are their respective correlation coefficients) and, thus, negatively correlated with $Y_2$, $r = -0.71$.

To start the procedure, a prediction model that relates process variables in **X** with product characteristics in **Y** is needed. We fitted a PLS (Partial Least Squares) model by using autoscaled values in **X** as regressors to predict the five responses in **Y**, also autoscaled. By using venetian blinds (with 7 splits and 1 sample per split) as cross-validation method and taking into account the variance explained in both **X** and **Y,** and the prediction of the five responses in the training set, four latent variables were selected.

The four latent variables PLS2 model explains 65.97% of the variance in regressors (**X**-block) and 92.63% of the variance in the **Y**-block. The fitted PLS2 model provides values of $R^2$ equal to 0.92, 0.97, 0.77, 0.98 and 0.99 in fitting and 0.88, 0.94, 0.65, 0.95 and 0.98 in crossvalidation for response variables $Y_1$, $Y_2$, $Y_3$, $Y_4$ and $Y_5$, respectively. The stability of the model is remarkable since the values in fitting and crossvalidation are very similar.

Also, the PLS2 model has been evaluated by performing permutation test where the **Y**-block is shuffled allowing the calculation of the probability that the results obtained with the unperturbed **Y**-block are significant or not. The self-predicted and cross-validated residuals of each permutation are compared to the original residuals using the Wilcoxon, Sign and Randomized t-tests. The three tests (in fitting and in crossvalidation) ensure that the model is significant with a confidence level greater than 0.995 for each of the five variables.

### 4.1 **Producing a product with average quality**

The first case serves to explain the steps when looking for the settings of process variables to guarantee a given product, typically the product we are producing, i.e, the one whose quality characteristics are defined by the mean value of each variable in **Y**. According to Table 1, the 'average' product, which is now the target or desired vector that contains the required characteristics, is

$$\begin{aligned}\mathbf{y}_t &= mean(\mathbf{Y}) \\ &= (0.1330,\ 27320.4,\ 164054.9,\ 0.796,\ \ 26.11) \\ &= (y_{t,1}, y_{t,2}, y_{t,3}, y_{t,4}, y_{t,5})\end{aligned} \tag{5}$$

Notice that the PLS model was fitted to the process with autoscaled **X** and autoscaled **Y**, thus, the mean value of **X** is projected into the origin of the latent space (all the scores are null) and consequently, the PLS model predicts null responses, which correspond to the mean value of **Y.** That means that there is at least one vector in the process variables space that guarantees the quality characteristics sought.

For looking for the process variables to achieve a product with the characteristics in $\mathbf{y}_t$, the procedure is applied with a population size of 150 chromosomes, that is, one hundred and fifty 14-dimensional vectors selected inside the hyper-cube defined by the range of the process variables in **X,** detailed in the last column of Table 1, and that constitutes the space of allowable process variables, *D*.

Then, offsprings are generated with uniform selection among the elements in the current population in pairs, simple crossover (with uniform selection of the crossing point) and 20% of the coordinates are randomly modified (that is, the probability of mutation is 0.2) also with uniform probability.

To evaluate the members of the population and compute fitness, the fitted PLS should be applied, provided it is applicable to the 'potential' process variables. To check whether the generated vectors are in the same latent space as the one used to fit the model, we use Q and $T^2$ statistics so that offsprings with any of them larger than the corresponding critical values at 95% confidence level are directly discarded.

In this way, predictions of quality characteristics are obtained with the PLS model and these predictions are compared, individually, with the target ones. After 1000 generations, the non-dominated solutions of the final population constitute the estimate of the Pareto-optimal front, 116 solutions in this case.

The first remark is that one might have thought that the pareto-optimal front should have consisted of a single point (no conflict among the five objectives, so the front reduces itself to a single point) since the product is feasible. The variation observed in the estimated front is due to the stopping criteria. It has been already stated one of the characteristics that make evolutionary algorithms suitable for this problem, the fact that they maintain a population of solutions. Another characteristic is that the mutation guarantees that every point in *D* can be reached during evolution, although there is little chance to exactly pick a specific exact value because the search space is an infinite set. The expected behavior of the method is to converge towards this point, but, in practice, as any other iterative method, the evolution should be stopped at some point. We have used the number of generations as stopping criteria.

In practical terms, to take into account the whole variability before deciding about the solutions obtained, we need to state which differences are, in practice, null differences. Let's assume that individual differences below, say, the 20% of the range of **Y** are undistinguishable from the practical point of view. This means we have established a threshold value,

$$\mathbf{d}_0 = (0.0015, 229, 2492, 0.015, 0.12) \qquad (6)$$

After rounding the values to the precision of the corresponding units and by imposing this threshold (which is an additional constraint to the optimization problem), we discard those solutions among the front that surpass this threshold in at least one of the coordinates. The 82 remaining ones correspond to process variables that would produce the target characteristic in the resulting product.

These eighty-two 14-dimensional vectors containing the process variables in the pareto-optimal front are written in full in the supplementary material (Sheet 'case1' of the file pofs.xlsx), along with the fitness values $d_i$, *i=1, …, 5*, and some global measures.

In detail, the first fourteen columns are for the process variables in the pareto-optimal front, in the order of Table 1. Because it is known that the mean value of X, $\overline{\mathbf{x}} = (\overline{x}_1, \overline{x}_2, ..., \overline{x}_{14})$, constitutes a set of process variables to achieve $\mathbf{y}_t$, for each setting of process variables $(x_1, x_2, ..., x_{14})$ in the pareto-optimal front, the difference between it and the theoretical

solution is computed as a root squared error, $\text{RSE}_X = \sqrt{\sum_{i=1}^{14} \left(\overline{x}_i - x_i\right)^2}$ and added in the table after the X's values. An analogous calculation in the responses space (i.e., for the expected differences $d_i$) would be $\text{RSE}_Y = \sqrt{\sum_{i=1}^{5} \left(y_{t,i} - \hat{y}_i\right)^2} = \sqrt{\sum_{i=1}^{5} d_i^2}$ . These values are also written as the last column in the spread sheet and used to sort the different solutions in the pareto-optimal front, in increasing order of $\text{RSE}_Y$.

However, this is just a reduction of all the differences into a single number. The individual differences when comparing the predicted and target responses are the ones that qualify the performance of each solution in terms of pareto optimality. For instance, three of the solutions in the pareto-optimal front have the differences and $\text{RSE}_Y$ of Table 3.

Despite the fact that the $\text{RSE}_Y$ in the second row is twice the one of the first row, globally the second is a better solution for all the responses but $Y_3$. The $\text{RSE}_Y$ in the third row is almost 40 times greater than that of the second and more than 75 times than the one of the first but it is again only a consequence of the value achievable for $Y_3$ because this last solution is a much better solution for $Y_1$ and $Y_5$ and better for $Y_2$.

This also illustrates the kind of information we have in the trade-offs found in the pareto-optimal front. For moving from the first to the second row in Table 3, it is necessary to let $Y_3$ be separated (not much) from its desired value to have the other four characteristics closer to their desired values. Notice that the differences are reduced at least in a 30% comparing to the ones in the first row, except for $d_4$ that is the same. Looking at the last row, the target value for $Y_5$ can be achieved ($d_5 = 0$) simultaneously reducing $d_1$ and $d_2$ as long as we let $Y_4$ and $Y_3$ move further away from the desired values.

The first two columns in Table 4 reproduce the minimum and maximum of each column in the spreadsheet so that we see that, in fact, the minimum value of each difference $d_i$, last rows in Table 4, is 0. That means that there are settings for the process variables to guarantee the exact value of, at least, each individual characteristic (each coordinate of $\mathbf{y}_t$). Again, these solutions have interest since they are somehow the 'extremes' of the pareto-optimal front, though these are now theoretical properties because, in the context established with $\mathbf{d}_0$, we see that all the maxima are below the threshold, meaning that all the predicted values are, in practice, equal to $\mathbf{y}_t$.

Table 4 also shows the variation of $\text{RSE}_X$, from 4.82 to 29.41 among the computed solutions, and $\text{RSE}_Y$ in [29.66, 2490.89]. More interestingly, the first rows show the individual variation of each coordinate in the process variables. Comparing the individual intervals with those in Table 1, we see that the pareto-optimal solutions are, indeed, around the theoretical value (the mean) but do not spread over the whole domain, but in shorter intervals, specifically for $X_1$, $X_2$, and from $X_8$ to $X_{11}$ and, to a lesser extent, $X_3$ and $X_6$.

Although we are not as interested in each coordinate independently as in the whole vectors that define the set of process variables, it is clear that these settings in the pareto-optimal front lie in a sub-region inside the domain. In this region, these solutions show some of the variability of the settings of the process variables that we can afford without affecting the quality defined for the product, $\mathbf{y}_t$. Consequently, they are elements of the design space for $\mathbf{y}_t$.

Although all the settings in the pareto-optimal front are equal in relation to the intended quality, they might be not so in terms of the settings of the process variables. To describe these settings in the pareto-optimal front, a principal component analysis (PCA) is performed with the process variables. Fig. 1a) depicts their scores in the plane made up by the first two principal components. Six points are highlighted, numbered according to the row they have in the supplementary material, and representative of different score's behaviour.

To give an indication about the differences among these setting for the process variables, Table 5 identifies the solution (among these six) in which the minimum and maximum values of each process variable are attained. For example, solution number 69 has the maximum values of $X_1$, $X_2$, $X_3$, $X_8$, and $X_9$ and the minimum of $X_4$, $X_5$, $X_7$, $X_{12}$, and $X_{13}$, whereas the maximum of $X_4$ and $X_{12}$ are achieved in solution 52 together with the minimum of $X_2$, $X_3$, $X_6$, and $X_{14}$. The inspection of Table 5 shows that the six vectors contain settings that are maximum or minimum in at least one process variable, which is different in each one. Last column in Table 5 is the percentage of the range of these six solutions as against the whole range in **X** (Table 1). For instance, the range of $X_1$ among the six highlighted points (the maximum achieved in solution 69 minus the minimum in solution 22) is the 29% of the range this variable has in the historical data in **X**. Except for four of the process variables, this percentage is larger than 50%, that is, the six highlighted settings for the process variables are not only different from each other but they are well separated in the process variables space.

The corresponding quality characteristics for the settings in the pareto-optimal front are visualized in the biplot in Fig. 1b), corresponding to a new PCA now in the 5-dimensional output space. The points highlighted are the corresponding to the six settings in Fig. 1a) and Table 5. It is observed how solutions 22 and 52 maximize $Y_2$ and $Y_4$, respectively, while solution 57, despite its central position in Fig. 1a), maximizes $Y_3$ and $Y_1$. Finally, solution 69 minimizes $Y_3$ and $Y_1$.

In relation to the theoretical solution in this case (the mean value of process variables), in general longer evolutions, usually combined with larger populations, would be needed to get closer to it. In our case, with a population of three thousand vectors of process variables in *D* evolving for ten thousand generations, the population converges to a single point, the mean in **X** corresponding to the zero in the latent spaces (scores).

### 4.2 **Further exploring the design space**

For the purposes here, we take the quality characteristics of one of the products already in matrix **Y**, for example the first one, so that there is no doubt the product is feasible. On the other hand, the settings in **X** provide a direct solution to the problem (though there is some variability both in the process variables and in the predictions computed with the PLS model).

To explore the effect of this variability, a run of ten thousand generations with a population of 500 potential solutions moving towards spread and non-dominated solutions, provides a final population with an estimate of the pareto-optimal front made by 477 points, 318 of which with all the differences below $\mathbf{d}_0$ (Eq. 6)

The values of the members of the front that, in practice, predict the same quality characteristics along with the differences obtained (fitness values $d_1$-$d_5$) are in sheet 'case2' in the supplementary material. Like in the previous case, the extremes (maximum and minimum)

are noted in the first rows, and also the corresponding values of $RSE_X$ ($15^{th}$ column) and $RSE_Y$ (last column) are included.

When projecting these 318 settings of the process variables into the model of two principal components built with **X**, all the scores of the computed pareto-optimal solutions are around the scores of the 'real' one. Likewise, the scores on the first two PCs of the predicted responses are around the desired one in the plane corresponding to the model built with **Y** (figures not shown here).

More interestingly, to explore the pareto-optimal solutions, a PCA model with the process variables in the pareto-optimal front was computed. Fig. 2a), which depicts the scores on the first two principal components, shows three clear groups, marked with different symbols and colours. To identify the points, the same symbols and colours are used in Fig. 2b) that is the biplot of a new PCA made up with the predicted quality characteristics (**Y**-space) to see the effect of the corresponding process variables (**X**-space in Fig. 2a).

The groups in Fig. 2a) are not so clear in Fig. 2b), though it is identifiable that the process variables depicted as green squares in Fig. 2a) gives larger values of $Y_4$ and lesser values of $Y_2$, but always inside the range of acceptable values, according to Eq. (6), which is the property that allows users to guide the analysis of these optimal solutions according to their specific needs.

Also, except for $Y_3$, notice the similarity between the positions of variables $Y_i$ projected in the biplot in Fig. 1b) (case 4.1) and their position in Fig. 2b), both showing the relation among characteristics when exploring pareto-optimal solutions. This arrangement is different from the one when decomposing matrix **Y**, Fig. 3a).

### 4.3 **Feasibility of a product with given characteristics**

In sections 4.1 and 4.2 we do not directly address the feasibility of a product. Instead, we know that the product is feasible, either theoretically by looking at the latent variable space in section 4.1 or an actual product in the process variable itself in section 4.2. The two situations illustrate the use of the methodology to somehow estimate the variability that should be expected and to explore the design space for given characteristics.

So, let us consider the desired characteristics for a new product given in Eq. (7)

$$\mathbf{y}_{des} = \left(0.13,\ 26800,\ 160000,\ 0.82,\ 25.9\right) \tag{7}$$

All of them, individually, are inside the values already achieved in matrix **Y,** see Table 1, but with different structure. This difference can be seen, for instance, in Fig. 3 that depicts the characteristics of the projection of $\mathbf{y}_{des}$ into the two-component PCA model built with autoscaled **Y,** Figure 3a) is the biplot, Fig. 3b) contains the values of the Q and $T^2$ statistics.

Although $\mathbf{y}_{des}$ is projected well inside the model, in terms of $T^2$, we can see that it has a very large Q-residual, well above the limit (95% confidence level, dashed line in Fig. 3b). This fact, by itself, implies that the product is unfeasible [8] because it does not respect the structure of the quality characteristics.

However, it can be worthy to explore the extent of the conflict among the desired properties. To that end, several runs of the algorithm have been conducted to minimize all the

differences (in absolute value) between the predicted quality characteristics and the new target characteristics defined in Eq. (7). In the different runs, population size and probability of mutation have been varied with longer evolution times (more generations) to study the stability of the estimated pareto-optimal front. This is not a systematic study about the effect of the different metaparameters of the evolutionary algorithm on the estimated Pareto-optimal front. Rather, the different runs in different conditions (a couple of minutes each, more or less) converging to similar final populations was taken as a sign of reproducibility and stability of the estimated fronts.

One of these typical runs, after rounding to the corresponding units and extracting the non-dominated solutions, provide 121 settings of process variables with their corresponding fitness (the vector of individual differences, in absolute value). The minima and maxima of the process variables in the pareto optimal front and of the expected differences, in absolute value, are in the last two columns of Table 4. In spite of having process variables that exactly predict each individual characteristic (the minimum difference in all the coordinates is 0, the already mentioned theoretical extremes of the pareto-optimal front), the maximum values in each coordinate are much larger than in the first two columns of Table 4 to such an extent that no solution verifies that all the differences $d_1$ to $d_5$ are below the threshold value stated in $\mathbf{d}_0$, Eq. (6). In fact, we also see that the range of the process variables is practically the same as in Table 1, which is another indication that the algorithm explored the whole domain trying to comply the desired values but 'fail' to do so. That means that, in practice, we would have to change some of the desired characteristics to obtain a set of viable process variables; otherwise, as we already knew from Fig. 3, the product is unfeasible with the present process.

Nevertheless, the obtained estimate of the pareto-optimal front is already a description of the feasible solutions near the desired one. In that sense, it can help the user in deciding about the target characteristics and the way they can be redefined, if it were the case. In that sense, the approach presented here is different from the situation if we were mathematically inverting the matrix model, where possibly the user would have to start a sequential procedure trying different possibilities or combinations of responses to guarantee the values of one or several of the desired characteristics. In its present form, the own nature of the pareto-optimal front states that the user will have settings that would allow to reach, at least, each individual value of the desired characteristics, limited to the values of the others.

If some other alternatives must be explored, that is, the individual zeros (constrained by the values of the others) are not good enough, then the optimization algorithm can be guided in a different way by defining the optimization criteria accordingly. To illustrate this property, a new run has been conducted to explore the possibility of guaranteeing some of the desired values in $\mathbf{y}_{des}$ (at least two) without imposing the remaining ones.

The 449 solutions in the pareto optimal front extracted from the enlarged population (the one with the old and new solutions) is in the sheet 'case3' of the file in the supplementary material. Table 6 shows the differences, in absolute value, as against all the values in $\mathbf{y}_{des}$, from some solutions extracted from the front.

The solution number 1, whose differences in achieving $\mathbf{y}_{des}$ are in the first row of Table 6, has zeros in the columns corresponding to responses (quality characteristics) $Y_2$ and $Y_3$, so we have settings for the process variables (in the supplementary material) to guarantee the desired values for both $Y_2$ and $Y_3$.

Maintaining the desired value of $Y_3$, and with larger differences for $Y_1$, $Y_5$, and especially $Y_2$, the settings linked to solution #2 guarantee also $Y_4$; larger distance to $Y_2$ should be allowed to reduce the distance in $Y_5$ and achieve the desired value for $Y_1$ in solution #3 together with the desired value of $Y_3$.

Analogously, solution #4, much larger distance to the desired value of $Y_2$ allows obtaining the value in $Y_5$, still achieving the value in $Y_3$, and maintaining the distance in $Y_4$; whereas some more distance in both $Y_2$ and $Y_3$ are needed to guarantee the desired values for $Y_1$ and $Y_5$ in solution #5.

If, additionally, the practical differences defined by the threshold in $\mathbf{d}_0$ are taken into account, then there are in fact 122 possible settings to guarantee one of the targeted values, 257 for guaranteeing two of the characteristics, and 70 for simultaneously achieving three of the desired values. No solution is found with more than three.

Similar systematic approaches can be conducted, depending on the goal of the study, which in all the cases can give insights about what characteristics could be optimized at the expense of what others.

### 4.4 **Maximize all the quality characteristics**

This section extends the possibilities of the method being exposed, which is indeed an optimization engine, so the procedure can also be used to handle optimization problems expressed in its usual way. For example, we may wonder if it is possible to maximize all the quality characteristics, or maximize two of them minimizing the other three, or any other of the different situations that the person responsible of the process might want to explore.

Then, let us suppose that we want to study if, with the actual process, there are process variables that simultaneously maximize all the responses. In this particular case-study, we know that the problem does not have a unique solution due to the correlation among quality characteristics, the one observed in the lower part of Table 2: strong negative correlation of $Y_2$ as against the other three ($Y_1$, $Y_4$ and $Y_5$) evidences their opposite behaviour whereas these three, on their part, show positive correlation and can be simultaneously maximized. However, we need to know the extent of the conflict, and the maximal value attainable for each characteristic.

Contrary to the previous sections, there is not a specified target value because we do not know how much each quality characteristic can be improved (increased) when the process variables are in *D*. In other words, the direct matrix inversion is not well defined for a problem like this one that we can handle by re-defining the objective in Eq. (2) as:

$$\text{maximize}\{\hat{y}_i, i = 1, 2, ..., q\} \tag{8}$$

With the restrictions already stated, i.e., maintaining the defined domain for the process variables, *D*, and the restriction of being projected into the latent space defined by the PLS2 prediction model.

The modified setting of the problem forces to change the fitness values, which are no longer differences, but expected quality characteristics. Adapting the fitness accordingly and applying the algorithm for a thousand generations with population size 1000 and probability of

mutation 0.15, the resulting pareto-optimal front has 734 elements, sheet 'case 4' in the supplementary material, where also the expected quality characteristics are written.

It is worth repeating that now the fitness values are not differences but expected values of the quality characteristics. The correlation among the achievable values of the quality characteristics when they are guided towards larger values, upper triangular part in Table 2, reflects that the achieved solutions maintain the correlation among the quality characteristics while also keeping their optimal character.

In that sense, the extremes of the pareto-optimal front would correspond to the individual maxima attainable for each quality characteristic. The intervals of variation of the process variables in the Pareto-optimal front and the corresponding predicted quality characteristics are written in Table 7, where we can see that, indeed, all the process variables are inside *D,* covering the whole allowable space.

Also, comparing to the maximum values we already have in **Y** (Table 1) each individual characteristic $Y_i$ can be slightly improved with the present state of the process but at the cost of obtaining the minimum values already achieved in **Y** for some other characteristic. This last property cannot be deduced from Table 7, but following the different solutions in sheet 'Case4' in the supplementary material.

Table 8 shows the whole set of simultaneously attainable quality characteristics for the 'extreme' solutions in the pareto-optimal front, that is, the ones with the maximum value in one characteristic. They are identified by the row number in the file in supplementary material.

Thus, Table 8 shows that to predict the maximum attainable value for $Y_2$ (28163 in the second row in Table 8 that corresponds to row 541 in the supplementary material), values of 0.1294 should be assumed for $Y_1$ that is almost its minimum value (0.1289 in Table 1), middle values for $Y_3$ (among the achievable ones in the pareto-optimal front, from 160135 to 172930, Table 7), and again almost the minimum values in the pareto-optimal front of 0.747 for $Y_4$ and 25.67 for $Y_5$. The situation with $Y_3$ is completely different; its best value can be achieved with smaller 'lost' in the other four characteristics. $Y_1$, $Y_3$ and $Y_4$ on their part have their corresponding individual maxima around the largest values simultaneously attainable for the three characteristics, linked to middle values of $Y_3$ and among the worst values for $Y_2$.

## 5. Conclusions

The possibilities of a computational approach to help in the decision-making process have been shown. It allows tackling the feasibility of a product with desired final characteristics as an alternative to the explicit inversion of the mathematical prediction model, which is usually a latent variable model needed to predict the quality characteristics as a function of the process variables. To be precise, a PLS2 model is computationally inverted here for the first time.

By posing the issues to be tackled (the day-to-day decision making) as a multicriteria optimization problem, the proposed method helps in the study of an ongoing process in the framework of QbD. The procedure allows constraints on the process variables, on the quality characteristics or even in the latent variable space.

The fact of having the pareto-optimal front for each given problem facilitates the study of the feasibility of products jointly with the needed process variables and their allowable variation.

In that sense, it allows the definition of the design space. Even if the product is unfeasible, the estimate of the pareto-optimal front is still useful for exploring the conflict and decide about possible modifications of some of the desired characteristics.

The flexibility of the proposed procedure is shown as an alternative for studying the feasibility or unfeasibility of a given product that is characterised by defining a vector of targeted quality characteristics; or as a mean to handle other situation such as the possibility of optimizing (maximize, minimize, maintain a given value) one or several quality characteristics in a simultaneous way.

There are some open questions to be further studied. One of them refers to the precision of the conclusions attainable that, as usual when working with prediction methods, are limited for the goodness of the model when predicting. Besides, the effect on the design space of this uncertainty, and of the one due to the projection of **X** into the latent space, needs to be systematically addressed.

**Acknowledgements**

Authors acknowledge the financial support of Junta de Castilla y León (BU012P17), and also Spanish MINECO and Agencia Estatal de Investigación under research projects CTQ2014-53157-R, and CTQ2017-88894-R, respectively. All research projects were co-financed with European FEDER funds.

**References**

[1] E. Tomba, P. Facco, F. Bezzo, M. Barolo, Latent variable modeling to assist the implementation of Quality-by-design paradigms in pharmaceutical development and manufacturing: A review, International Journal of Pharmaceutics 457 (2013) 283-297.
[2] S. García-Muñoz, J. Mercado, Optimal selection of raw materials for pharmaceutical drug product design and manufacture using mixed integer nonlinear programming and multivariate latent variable regression models, Ind. Eng. Chem. Res. 52 (2013) 5934-5942.
[3] M. Ottavian, E. Tomba, M. Barolo, Advanced process decision making using multivariate latent variable methods, in: Ierapetritou, Ramachandran (Eds), Process simulation and data modeling in solid oral drug development and manufacture, Humana Press, Springer, New York, 2016.
[4] C.M. Jaeckle, J.F. MacGregor, Product design through multivariate statistical analysis of process data, AIChE Journal 44 (1998) 1105-1118
[5] E. Tomba, P. Facco, F. Bezzo, S. García-Muñoz, Exploiting historical databases to design the target quality profile for a new product, Ind. Eng. Chem. Res. 52 (2013) 8260-8271
[6] E. Tomba, M. Barolo, S. García-Muñoz, General framework for latent variable model inversion for the design and manufacturing of new products, Ind. Eng. Chem. Res. 51 (2012) 12886-12900.
[7] F. Yacoub, J.F. MacGregor, Product optimization and control in the latent variable space of nonlinear PLS models, Chemom. Intell. Lab. Syst. 70 (2004) 63– 74.
[8] S. García-Muñoz, T. Kourti, J.F. MacGregor, F. Apruzzese, M. Champagne, Optimization of Batch Operating Policies. Part I. Handling Multiple Solutions, Ind. Eng. Chem. Res. 45 (2006), 7856-7866.
[9] S. Lakshminarayanan, H.Fujii, B. Grosman, E. Dassau. D.R. Lewin, New product design via analysis of historical databases, Computers and chemical engineering 24 (2000) 671-676.

[10] M.C. Ortiz, L. Sarabia, A. Herrero, M.S. Sánchez, Vectorial optimization as a methodogical alternative to desirability function, Chemom. and Intell. Lab.Systems 83 (2006) 157–168.
[11] L.A. Sarabia, M.S. Sánchez, M.C. Ortiz, Introduction to ranking methods, in: Manuela Pavan and Roberto Todeschini (Eds.), Scientific Data Ranking Methods: theory and applications, Data Handling in Science and Technology 27. Elsevier, 2008, pp. 1-50.
[12] K. Deb, Multi-objective optimization using evolutionary algorithms, Wiley, New York, 2001.
[13] C. Reguera, M.S. Sánchez, M.C. Ortiz, L.A. Sarabia, Pareto-optimal front as a tool to study the behaviour of experimental factors in multi-response analytical procedures, Anal. Chim. Acta 624 (2008) 210-222.
[14] M.S. Sánchez, L.A. Sarabia, M.C. Ortiz, On the construction of experimental designs for a given task by jointly optimizing several quality criteria: Pareto-optimal experimental designs, Analytica Chimica Acta 754 (2012) 39– 46.
[15] M.S. Sánchez, M.C. Ortiz, L.A. Sarabia, A useful tool for computation and interpretation of trading-off solutions through pareto-optimal front in the field of experimental designs for Mixtures, Chemometrics and Intelligent Laboratory Systems 158 (2016) 210–217.
[16] J.E. Jackson, A User's Guide to Principal Components, p. 36, John Wiley & Sons, New York, NY, 1991.
[*dataset*] [17] Kevin Dunn, LDPE, OpenMV.net Datasets, 2011, http://openmv.net/info/ldpe (accessed 24/7/2017).
[18] J. F. MacGregor, C. Jaeckle, C. Kiparissides, M. Koutoudi, Process monitoring and diagnosis by multiblock PLS methods. AIChE J. 40 (1994) 826–838.

**Figure captions**

Figure 1. Case 4.1. Principal components analyses computed with the solutions in the pareto-optimal front: a) Scores of process variables; b) Biplot of quality characteristics. The number that identifies some of them is the number that corresponds to their position in the supplementary material (Sheet 'case 1' of pofs.xlsx)

Figure 2. Case 4.2. Separate principal components analyses computed with the solutions in the pareto-optimal front for a) process variables, and b) quality characteristics. Different symbols are used to synchronise the process variables in Fig. 2a) with the expected quality characteristics in the biplot of Fig. 2b).

Figure 3. Projection of $y_{des}$ in Eq. 7, red square, into the two principal components model computed for **Y**. Biplot in a); values of Q and $T^2$ statistics in b) where the dashed lines mark the limits computed at 95% confidence level.

Table 1. Process variables and product characteristics measured for a low-density polyethylene. The last columns contain mean, standard deviation, minimum and maximum of each variable.

| | Name | Explanation | Mean (Standard deviation) | Interval of variation |
|---|---|---|---|---|
| Process variables in **X** | $X_1$ | inlet temperature to zone 1 of the reactor | 206.94 K (1.61) | [204.13, 209.88] |
| | $X_2$ | maximum temperature along zone 1 | 296.45 K (1.85) | [292.96, 300.19] |
| | $X_3$ | outlet temperature from zone 1 | 232.71 K (4.20) | [225.61, 240.25] |
| | $X_4$ | maximum temperature along zone 2 | 284.21 K (2.24) | [279.28, 288.74] |
| | $X_5$ | outlet temperature from zone 2 | 242.44 K (2.76) | [235.76, 246.76] |
| | $X_6$ | temperature of inlet coolant to zone | 116.94 K (0.56) | [115.91, 118.00] |
| | $X_7$ | temperature of inlet coolant to zone 2 | 117.82 K (0.499) | [116.90, 118.58] |
| | $X_8$ | percentage along zone 1 where Tmax1 occurs | 0.0295 % (0.0019) | [0.026, 0.033] |
| | $X_9$ | percentage along zone 2 where Tmax2 occurs | 0.5796 % (0.0053) | [0.569, 0.592] |
| | $X_{10}$ | flow rate of initiators to zone 1 | 0.4631 g $s^{-1}$ (0.0269) | [0.4089, 0.5053] |
| | $X_{11}$ | flow rate of initiators to zone 2 | 0.4648 g $s^{-1}$ (0.0282) | [0.4081, 0.5091] |
| | $X_{12}$ | flow rate of solvent to zone 1 [% of ethylene] | 665.77 % (6.14) | [655.025, 676.97] |
| | $X_{13}$ | flow rate of solvent to zone 2 [% of ethylene] | 246.32 % (2.29) | [242.07, 250.10] |
| | $X_{14}$ | pressure in the reactor | 3004 atm (19) | [2970, 3033] |
| Product (quality) characteristics in **Y** | $Y_1$ | cumulative conversion | 0.1330 (0.00177) | [0.1289, 0.1366] |
| | $Y_2$ | number average molecular weight | 27320 (262) | [26758, 27904] |
| | $Y_3$ | weight average molecular weight | 164055 (3063) | [157505, 169964] |
| | $Y_4$ | long chain branching per 1000 C atoms | 0.7964 (0.0173) | [0.752, 0.829] |
| | $Y_5$ | short chain branching per 1000 C atoms | 26.11 (0.15) | [25.84, 26.43] |

Table 2. Pearson's correlation coefficients for the variables in **Y**, lower triangular part; and for the predicted **Y** in case 4.4, upper triangular part. The correlation coefficients of 1, between a variable and itself, are omitted.

| | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ |
|---|---|---|---|---|---|
| $Y_1$ | | -0.94 | 0.20 | 0.98 | 0.91 |
| $Y_2$ | -0.71 | | 0.10 | -0.91 | -0.98 |
| $Y_3$ | 0.37 | 0.18 | | 0.32 | -0.22 |
| $Y_4$ | 0.88 | -0.80 | 0.42 | | 0.83 |
| $Y_5$ | 0.74 | -0.92 | -0.14 | 0.71 | |

Table 3. Case 4.1. Fitness values ($d_1$ to $d_5$) for three pareto-optimal solutions. The last column refers to the root squared error (RSE) when comparing the predicted and target characteristics.

| $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ | $RSE_Y$ |
|---|---|---|---|---|---|
| 0.00112 | 12.2 | 31 | 0.0002 | 0.06 | 33.3143 |
| 0.00072 | 8.9 | 66 | 0.0002 | 0.03 | 66.5974 |
| $9.0\ 10^{-5}$ | 2.1 | 2486.8 | 0.0048 | 0 | 2486.8009 |

Table 4. Intervals of variation of the solutions in the Pareto-optimal fronts for cases 4.1 (first two columns) and 4.3 (last two columns). $\mathbf{y}_t$ is defined in Eq. (5) and $\mathbf{y}_{des}$ in Eq. (7).

| | | Case 4.1 | | Case 4.3 | |
|---|---|---|---|---|---|
| | Name | Minimum | Maximum | Minimum | Maximum |
| Process variables | $X_1$ | 205.17 | 207.87 | 204.13 | 209.88 |
| | $X_2$ | 293.27 | 298.63 | 292.96 | 300.19 |
| | $X_3$ | 227.97 | 240.25 | 225.81 | 240.25 |
| | $X_4$ | 279.28 | 288.53 | 279.59 | 288.74 |
| | $X_5$ | 235.76 | 246.08 | 236.53 | 246.69 |
| | $X_6$ | 116.10 | 117.59 | 115.91 | 118.00 |
| | $X_7$ | 116.90 | 118.49 | 117.01 | 118.58 |
| | $X_8$ | 0.026 | 0.031 | 0.026 | 0.033 |
| | $X_9$ | 0.574 | 0.587 | 0.569 | 0.592 |
| | $X_{10}$ | 0.4268 | 0.4847 | 0.4089 | 0.5053 |
| | $X_{11}$ | 0.4269 | 0.5021 | 0.4243 | 0.5091 |
| | $X_{12}$ | 656.24 | 674.86 | 658.78 | 676.97 |
| | $X_{13}$ | 242.07 | 249.21 | 242.07 | 247.52 |
| | $X_{14}$ | 2977 | 3033 | 2970 | 3010 |
| $RSE_X$ | | 4.8242 | 29.4130 | | |
| Differences in absolute value | $d_1$ | 0 | 0.0014 | 0 | 0.0056 |
| | $d_2$ | 0 | 203 | 0 | 945. |
| | $d_3$ | 0 | 2489 | 0 | 11516. |
| | $d_4$ | 0 | 0.014 | 0 | 0.066 |
| | $d_5$ | 0 | 0.11 | 0 | 0.48 |
| $RSE_Y$ | | 29.6626 | 2490.8930 | | |

Table 5. Case 4.1. Comparison of the pareto-optimal solutions marked in Fig. 1. See text for details.

| Variable | minimum | maximum | % of range |
|---|---|---|---|
| $X_1$ | 22 | 69 | 29 |
| $X_2$ | 52 | 69 | 74 |
| $X_3$ | 52 | 69 | 77 |
| $X_4$ | 69 | 52 | 98 |
| $X_5$ | 69 | 2 and 36 | 55 |
| $X_6$ | 52 | 2 and 36 | 49 |
| $X_7$ | 36 and 69 | 2 | 91 |
| $X_8$ | 2 | 69 | 64 |
| $X_9$ | 2 | 69 | 42 |
| $X_{10}$ | 22 | 36 | 57 |
| $X_{11}$ | 52 | 57 | 38 |
| $X_{12}$ | 69 | 52 | 60 |
| $X_{13}$ | 69 | 22 | 89 |
| $X_{14}$ | 52 | 22 | 60 |

Table 6. Case 4.3. Fitness values $d_i$ when predicting the desired values in Eq. (7) without imposing all the values at once.

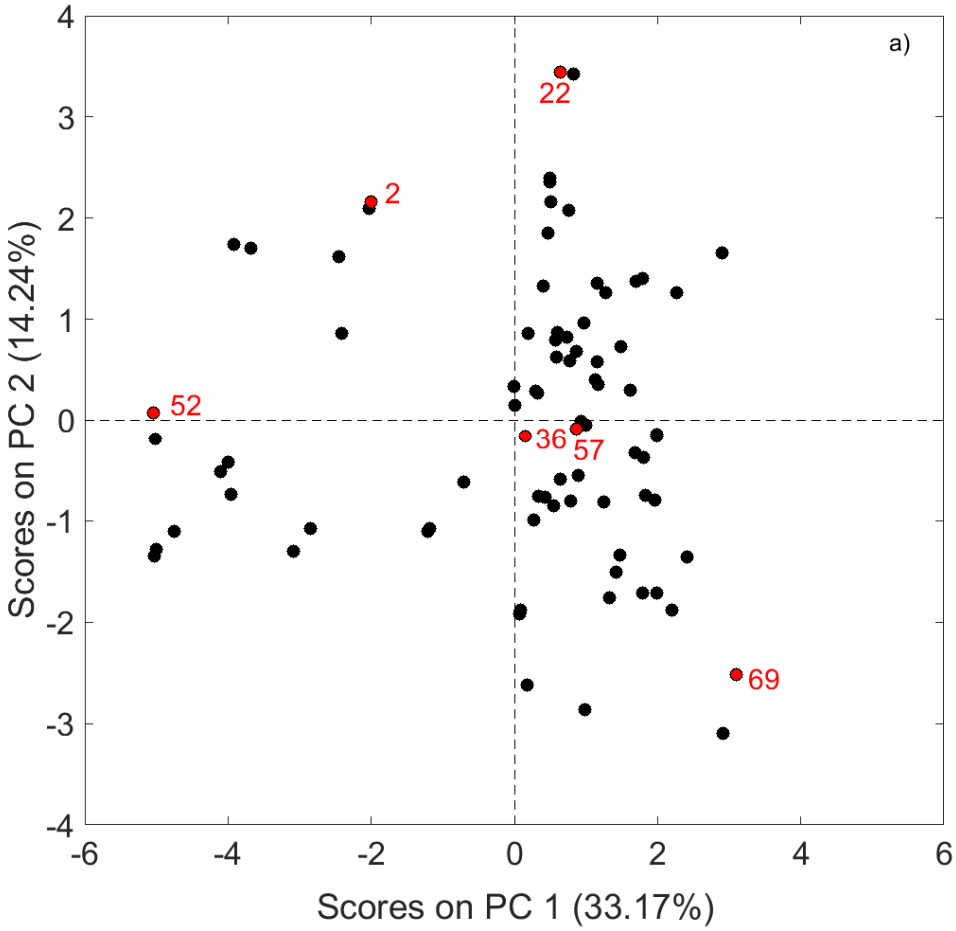
| Number | $d_1$ | $d_2$ | $d_3$ | $d_4$ | $d_5$ |
|---|---|---|---|---|---|
| 1 | 0.00381 | 0 | 0 | 0.0052 | 0.419 |
| 2 | 0.00422 | 64 | 0 | 0 | 0.438 |
| 3 | 0 | 673.4 | 0 | 0.0465 | 0.073 |
| 4 | 0.00026 | 741.2 | 0 | 0.0472 | 0 |
| 5 | 0 | 876.4 | 2247 | 0.0518 | 0 |

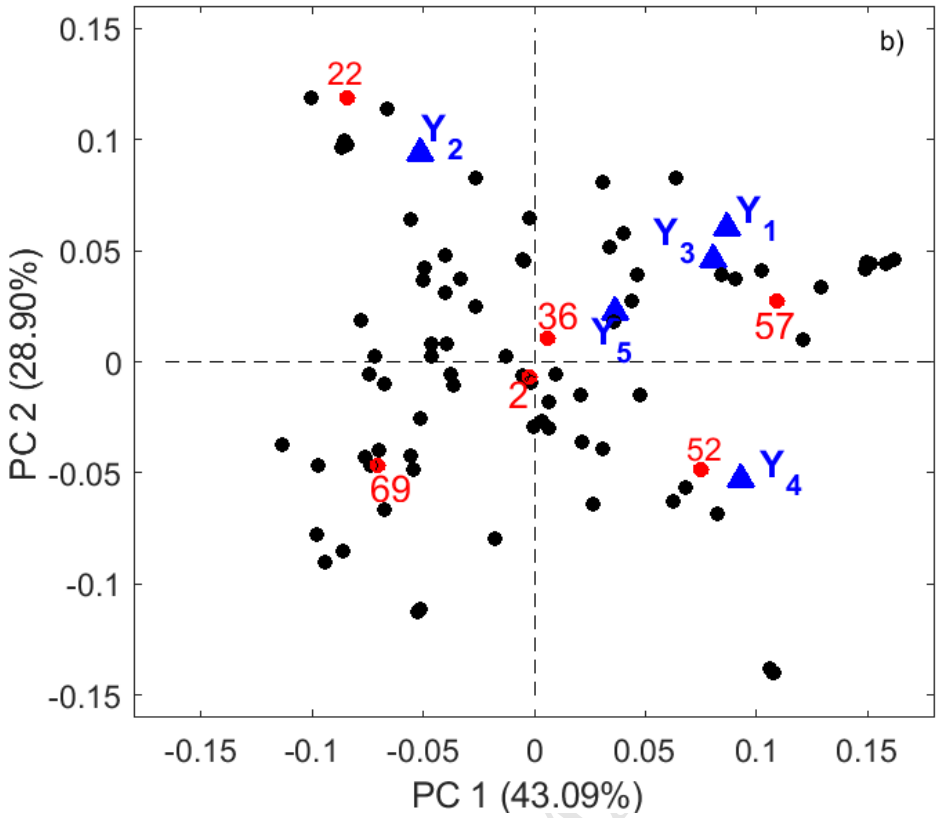
Table7. Case 4.4. Maximum and minimum values of the process variables ($X_i$) and the predicted characteristics ( $\hat{y}_i$ ) obtaining when trying to maximize all of them

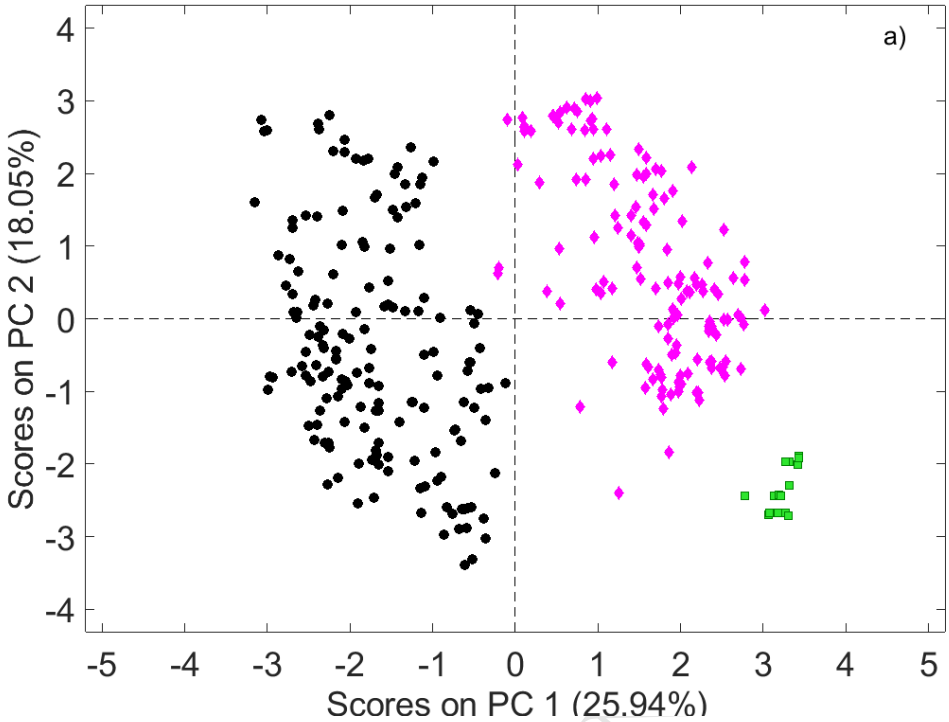
| Variable | minimum | maximum |
|---|---|---|
| $X_1$ | 204.13 | 209.88 |
| $X_2$ | 292.96 | 300.19 |
| $X_3$ | 225.61 | 240.25 |
| $X_4$ | 279.28 | 288.74 |
| $X_5$ | 235.76 | 246.76 |
| $X_6$ | 115.91 | 118 |
| $X_7$ | 116.90 | 118.58 |
| $X_8$ | 0.026 | 0.033 |
| $X_9$ | 0.569 | 0.592 |
| $X_{10}$ | 0.4089 | 0.5053 |
| $X_{11}$ | 0.4081 | 0.5091 |
| $X_{12}$ | 655.02 | 663.56 |
| $X_{13}$ | 242.07 | 250.10 |
| $X_{14}$ | 2970 | 3033 |
| $\hat{y}_1$ | 0.1290 | 0.1386 |
| $\hat{y}_2$ | 26603 | 28163 |
| $\hat{y}_3$ | 160135 | 172930 |
| $\hat{y}_4$ | 0.7469 | 0.8515 |
| $\hat{y}_5$ | 25.64 | 26.55 |

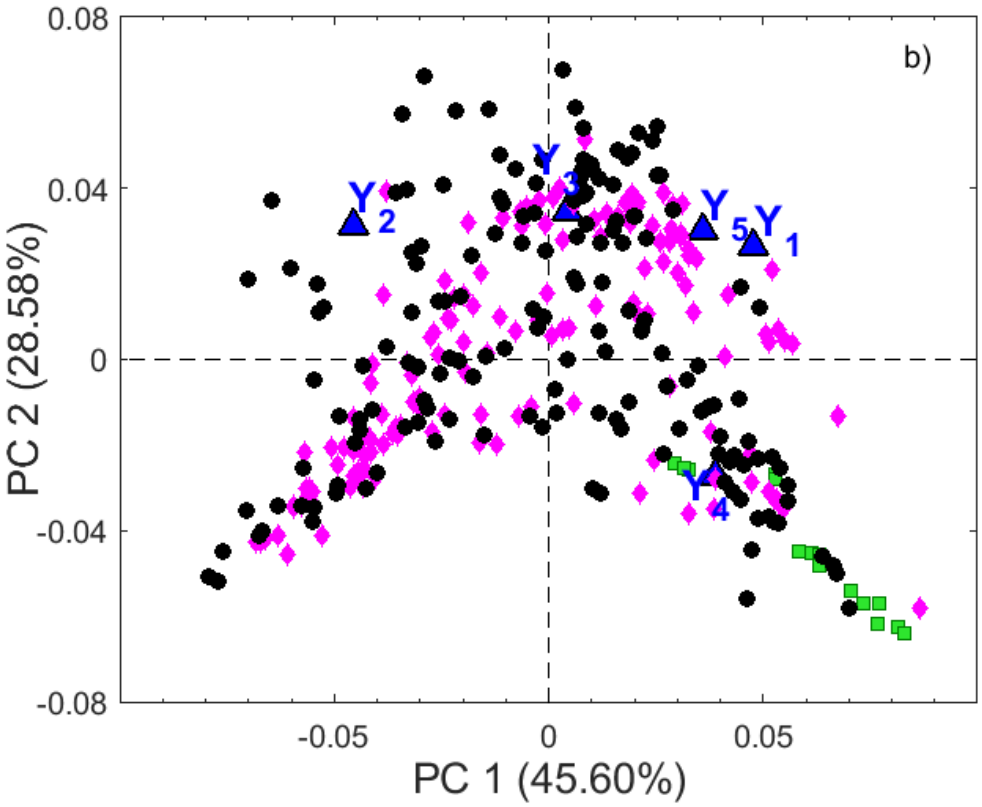
Table 8. Values of the quality variables for the solutions in the pareto-optimal front that reach the maximum in each individual characteristic (in bold in the main diagonal).

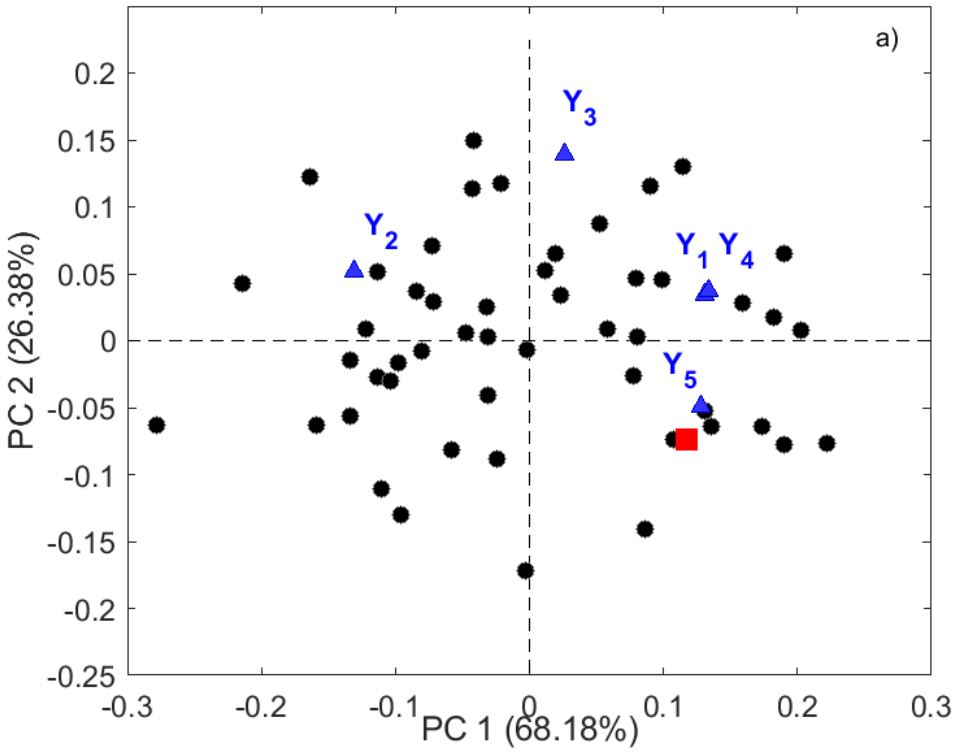
| Row in sheet 'Case 4' | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | $Y_5$ |
|---|---|---|---|---|---|
| 541 | **0.1386** | 26742 | 168070 | 0.845 | 26.47 |
| 16 | 0.1294 | **28163** | 166273 | 0.754 | 25.67 |
| 285 | 0.1357 | 27409 | **172930** | 0.824 | 26.05 |
| 601 | 0.1381 | 26679 | 167982 | **0.852** | 26.41 |
| 715 | 0.1375 | 26677 | 163387 | 0.830 | **26.55** |

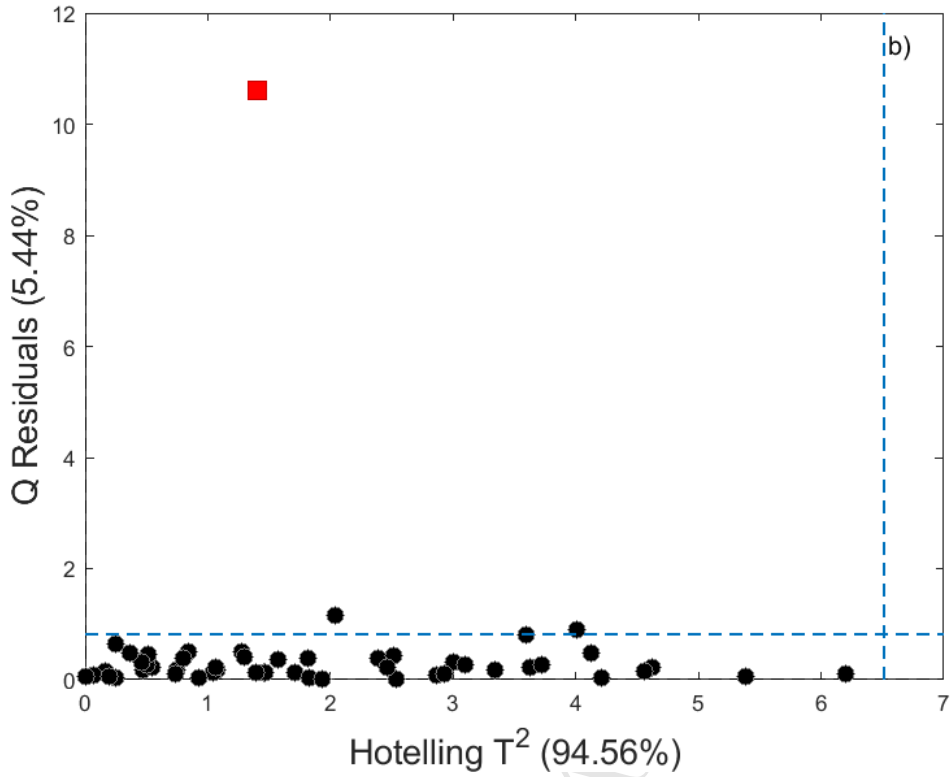

a)
22
2
52
36
57
69
Scores on PC 2 (14.24%)
Scores on PC 1 (33.17%)
4
3
2
1
0
-1
-2
-3
-4
-6
-4
-2
0
2
4
6








a)
Y3
Y2
Y1
Y4
Y5
PC 2 (26.38%)
PC 1 (68.18%)
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.2
-0.1
0
0.1
0.2
0.3



Computational method for Latent Variable Model Inversion, desired in process control

Help in the decision-making, avoiding the computation of matrix inverses

Exploration of potential solutions in both the process and latent variables spaces

A new procedure to evaluate design space in Quality by Design

An approach to partial least squares model inversion in Process Analytical Technology