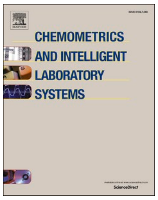
Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics



Defining multivariate raw material specifications via SMB-PLS

Joan Borràs-Ferrís [a,*], Carl Duchesne [b], Alberto Ferrer [a]

[a] Multivariate Statistical Engineering Group, Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València, València, Spain
[b] Chemical Engineering Department, Laval University, Quebec, Canada

ARTICLE INFO

Keywords:
Multivariate specifications
Raw materials
Design space
Industry 4.0
SMB-PLS

ABSTRACT

The Sequential Multi-Block Partial Least Squares (SMB-PLS) model inversion is applied for defining analytically the multivariate raw material region providing assurance of quality with a certain confidence level for the critical to quality attributes (CQA). The SMB-PLS algorithm does identify the variation in process conditions uncorrelated with raw material properties and known disturbances, which is crucial to implement an effective process control system attenuating most raw material variations. This allows expanding the specification region and, hence, one may potentially be able to accept lower cost raw materials that will yield products with perfectly satisfactory quality properties. The methodology can be used with historical/happenstance data, typical in Industry 4.0. This is illustrated using simulated data from an industrial case study.

1. Introduction

Raw materials properties are usually considered as Critical Input Parameters (CIPs) because their variability has an impact on Critical Quality Attributes (CQAs) of the final product. Despite their importance, specifications are usually defined in an arbitrary way based mostly on subjective past experience, instead of using a quantitative objective description of their impact on CQAs. Furthermore, in many cases, univariate specifications on each property are designated, with the implicit assumption that these properties are independent from one another. However, multivariate specifications provide much insight into what constitutes acceptable raw material batches when their properties are correlated (as usually happens) [1]. To cope with this correlation several authors suggest using multivariate approaches, such as Partial Least Squares (PLS) regression. Two approaches emerge from the literature when using PLS [2]. The first is based on a direct mapping in the latent space whereas the second defines specifications by the PLS model inversion.

The first systematic study was reported by De Smet [1] based on a direct mapping of good quality final product and associated lots of raw materials in the latent space, followed by selection of boundaries that best balance type I and II risks. The resulting region is then used to decide whether a new incoming lot of raw materials should be accepted or rejected. The key assumption of this method is that variability in the CQAs results exclusively from variations in the raw materials properties. Duchesne and MacGregor [3] generalized this method by assuming that both variations in raw materials properties and in process operating conditions are responsible for CQAs variations. Later on, García-Muñoz [4] extended the Duchesne-MacGregor method to combine data from multiple scales (e.g., lab or pilot scale and commercial scale) with different processing conditions and control strategies. These approaches, however, focused on defining multivariate specification regions on the multiple properties of a single raw material. To overcome this limitation, MacGregor et al. [5] extended them to determine the acceptability of new raw materials from multiple suppliers and with multiple measured properties, as well as to assess the suitability of combining specific batches of raw materials currently in inventory to minimize the risk of manufacturing a poor quality product. Finally, Azari et al. [6] proposed a Sequential Multi-block PLS (SMB-PLS) algorithm instead of PLS. The SMB-PLS imposes a sequential pathway between the regressor blocks according to the process flowsheet (e.g., raw material properties and process operating conditions), and then uses orthogonalization to separate correlated information between the blocks from orthogonal variations. Hence, the SMB-PLS captures the impact of variations in raw material properties on the process and on CQAs in the first block of latent variables. This allows identifying feedback/feedforward control actions made to compensate for variations in raw material properties. Then, the second block of latent variables captures process variations that are independent from raw material properties and also affect CQAs, e.g., certain (unplanned) excitations due to small

[*] Corresponding author. Department of Applied Statistics and Operational Research and Quality, Universitat Politècnica de València, Camino De Vera s/n, 7A, 46022, València, Spain.
E-mail address: joaborfe@eio.upv.es (J. Borràs-Ferrís).

https://doi.org/10.1016/j.chemolab.2023.104912
Received 6 April 2023; Received in revised form 7 June 2023; Accepted 3 July 2023
Available online 5 July 2023

changes in the process conditions during their daily operation. For that reason, the SMB-PLS is more efficient to establish the multivariate specifications when raw material properties and process conditions are correlated as it better sorts the contribution of both on the CQA variations.

In the aforementioned references, based on the direct mapping approach, the general shape (e.g., an ellipsoid or a straight line) and locus of the boundaries is decided by the user, trying to best balance out the type I and type II risks as commented. In contrast to this, García-Muñoz, Dolph, and Ward [7] emphasized the use of mathematical and statistical models as an objective way to define such specifications by linking them with a desired set of CQAs. In this sense, the PLS model inversion is of interest because it allows predicting an appropriate set of raw materials linked to the specification limits for CQAs. Moreover, when inverting PLS models, their prediction uncertainty is also back-propagated [8,9] and, hence, Borràs-Ferrís et al. [10] presented a methodology for defining analytically the raw material specification region in the latent space where the prediction uncertainty is considered. Thus, this region is expected to provide assurance of quality with a certain confidence level for the CQAs. In this regard, such region refers to the estimation of the so-called raw material Design Space (DS), which is defined as the multidimensional combination and interaction of inputs variables (e.g., raw material properties) that have been demonstrated to provide assurance of quality [11].

Since not only raw material properties influence the quality of the final product, but also the process conditions do, it is reasonable to consider the possibility to modify process conditions to compensate for raw material properties variations. Thus, wider raw materials specifications could be used if an effective process control system attenuating most raw material variations is implemented. In this sense, García-Muñoz, Dolph, and Ward [7] already proposed a feed-forward controller based on the PLS model inversion. However, this approach requires solving an optimization problem by a non-linear programming method, where raw material properties are fixed to hard constraints reducing the degrees of freedom to only process conditions. Thus, once a new raw material batch is received, the controller is executed in order to calculate the combination of the best process conditions, based on the desired CQAs, for such raw material batch. Note that, if too many constraints are specified for raw material properties, the model inversion solution may be forced to move away from the latent model [12]. Besides, this approach makes no attempt to differentiate between correlated and uncorrelated variations in process conditions with raw material properties and, hence, its proposed feed-forward controller does not identify properly the control actions from the past.

The purpose of this work is to develop a novel methodology taking advantage of the SMB-PLS model already discussed in the direct mapping approach but applied into the PLS model inversion approach. Thus, by means of the SMB-PLS model inversion, this methodology allows defining analytically such specifications by considering the possibility to modify process conditions prior to selecting a new raw material batch and, hence, it does not require solving an optimization problem each time a new raw material batch is received. In addition to that, unlike PLS, the SMB-PLS model does identify the variation in process conditions uncorrelated with both raw material properties and known disturbances, which is crucial as the modification of process conditions only must be inferred from such variations.

2. Data requirements

The data required for developing raw materials multivariate specifications following the methodology proposed in this paper involves three data blocks: **Z**, **X** and **Y**. **Z** ($N \times M$) is a matrix of inputs which includes a total of M measurements characterizing the properties of each of the N batches of a particular raw material, **X** ($N \times K$) is a matrix of inputs which includes a total of K process conditions used to process each one of the N batches of a particular raw material. In this work, it is assumed that process conditions refer to process manipulated variables. The **Y** ($N \times L$) output matrix consists of L measurements of the CQAs of the final product obtained for each one of the N corresponding batches. Finally, batches of raw materials are typically large, and it is assumed that the process will run for a long period at steady state on each batch. Thus, the three data blocks are collected in steady state.

3. Latent variable regression model

The latent variable regression models are tools specifically designed to analyze large data sets of highly correlated data, that find the main driving forces (i.e., latent variables) on the input space that are most related to the output space, being both spaces projected into a common latent space [12]. Thus, it is used not only to model the inner relationships between the matrix of inputs (**Z** and **X**) and the matrix of output variables (**Y**), but also to provide a model for both. This fact gives them a very nice property: uniqueness and causality in the reduced latent space no matter if the data come either from a Design of Experiments (DOE) or daily production process (historical/happenstance data) typical in Industry 4.0 [13].

The SMB-PLS is a multi-block latent regression model that combines the strengths of Multi-block PLS (MB-PLS) and those of the Sequential Orthogonal PLS (SO-PLS) methods [14]. Indeed, the SMB-PLS improves interpretability of between block relationships over the traditional MB-PLS methods by imposing a sequential ordering of the blocks (pathway) and applying stepwise block orthogonalization. Besides, as opposed to the SO-PLS, it models both the orthogonal and correlated information between blocks. In this section, the SMB-PLS is described as in Ref. [6], followed by the analytical definition of its inversion.

3.1. SMB-PLS regression model

The pseudocode of the SMB-PLS is presented in Appendix A and the algorithm is also shown schematically in Fig. 1, similarly as in Ref. [14]. The algorithm in Fig. 1 is presented for the two-blocks case (**Z** and **X**) for simplicity, but it can be extended to any number of regressor blocks as it is shown in Appendix A.

Fig. 1 shows that the SMB-PLS uses a hierarchical structure where the input blocks are ordered according to the process flowsheet with the first

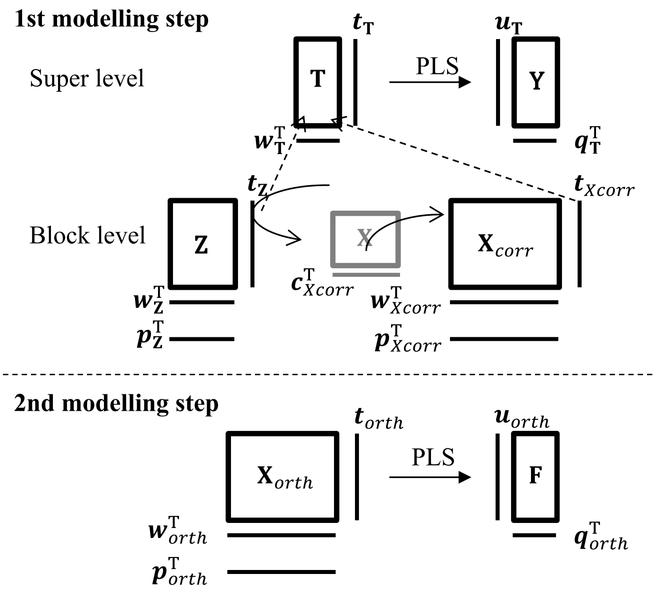


Fig. 1. Scheme of SMB-PLS algorithm for two input blocks.

block **Z** containing incoming raw material properties, and process data in the second block **X**. The algorithm computes the block weights and scores from the first block **Z**. The subsequent block **X** is then regressed onto the first block scores to extract the information that is correlated with **Z**, and their block weights and scores are then calculated. All block scores are combined in the super level score matrix **T** and a PLS model is built between **Y** and **T** to obtain the super weights and super scores. Upon convergence, super-score deflation is applied to the input blocks, **Z** and **X**, and the output block, **Y**, ensuring that the next component will extract orthogonal information to the first one. The procedure is repeated for computing the next component using the residual of all data blocks. It continues to extract components from the first regressor block in the sequence until it has modeled all relevant information from **Y**. When all relevant information from **Z** is extracted in the first modelling step, a regular PLS model is fitted to the **X** and **Y** residuals (i.e., $\mathbf{X}_{orth}$ and **F**, respectively) in the second modelling step. Thus, the SMB-PLS latent space can be expressed, similarly as PLS, but as two blocks of latent variables (Eq. (1)).

$$\mathbf{Y} = [\mathbf{T}_{\mathbf{T}}\ \mathbf{T}_{orth}] \cdot [\mathbf{Q}_{\mathbf{T}}\ \mathbf{Q}_{orth}]^{\mathrm{T}} + \mathbf{F}^{*} = \mathbf{T} \cdot \mathbf{Q}^{\mathrm{T}} + \mathbf{F}^{*} \tag{1}$$

where $\mathbf{F}^{*}$ are the residuals of **Y** after extracting the last SMB-PLS component. SMB-PLS captures the impact of variation in raw material properties on the process and on **Y** in the first modelling step represented by the first block of latent variables, $\mathbf{T}_{\mathbf{T}}$, referring to $[\mathbf{Z}\ \mathbf{X}_{corr}]$. These latent variables allow identifying past operating procedures, and control actions from the past (i.e., feedback/feedforward control) implemented to compensate for raw material properties variations.

Note that as already commented, process data is collected in steady state, and hence, dynamics are not considered. Besides, if the feedforward or feedback controllers remove the disturbances completely (perfectly), no deviation in **Y** in steady state will be captured after a raw material disturbance occurred. In such a case, there will be a correlation between **Z** and the manipulated variable in the control loop **X**, but that information should not be captured by any latent variable since there will be no correlation with **Y**.

However, in the case of feedforward control on raw material properties, an ideal controller would compensate any raw material disturbance completely only if it would know the "true" model, which is never the case. In the case of a feedback control, if the controller transfer function includes an integrating element (e.g., the I mode in PID controller that seeks to eliminate the residual error according to the historic cumulative error), and if the manipulated variable does not reach an upper or lower bound (i.e., saturation), the impact of the disturbance on **Y** should not be captured if the data is collected truly in steady state (i.e., perfect controller). Note that, feedforward controllers are never ideal, nor feedback controllers are perfect. Therefore, these controllers do not compensate perfectly (i.e., there will be a residual effect on **Y**). In addition to that, regarding the feedback control, the correlations between the manipulated and controlled variable of the control loop are not causal but anti-causal, that is, these correlations capture the reciprocal of the control transfer functions, leading to the negative inverse of the controller gain for steady state data. Finally, note that control loops are known, and hence, when interpreting the SMB-PLS model, these correlations are not a surprise, but something expected.

In both feedforward/feedback control, the correlation between **Z** and the manipulated variables in the control loops **X**, related to the residual effect on **Y***,* will be captured by some latent variables in the first block. Anyway, in both feedback/feedforward control, the purpose is not to interpret these relationships as causal (which would be wrong), but to account for them in the first block of latent variables. Thus, in the second modelling step, the second block of latent variables, $\mathbf{T}_{orth}$ referring to $\mathbf{X}_{orth}$, is expected to capture only process variations that are independent from raw materials and also affect **Y** (e.g., certain (unplanned) excitations). The main aim of this study is to take advantage of the information captured in this second block to improve the control actions from the past in a feedforward control strategy.

In order to evaluate the model performance of an observation, the Hotelling T^2 in the latent space and the Squared Prediction Error *SPE* are calculated [15]. The Hotelling T^2 statistic of an observation is the estimated squared Mahalanobis distance from the center of the latent subspace to the projection of such observation onto this subspace. The SPE statistic referring to the X-space gives a measure of how close (in an Euclidean way) such observation is from the *A*-dimensional latent space. Upper confidence limits (with a specified confidence level) for both statistics can be calculated based on theoretical distributions [16,17] or they can be obtained from distribution free methods by repeated sampling [18]. In the following sections, *SPE* and T^2 99% confidence limits are calculated from theoretical distributions.

3.2. *SMB-PLS model regression inversion and the null space*

The objective of model inversion is to find (predict) a window of inputs (raw materials properties, process conditions, etc.) for a desired product quality. Jaeckle and MacGregor [13] proposed a framework for the inversion of latent variable regression models using historical data available on the process operating conditions and on the corresponding product quality. Using standard regression models (e.g., linear regression or artificial neural network), the inversion is inadequate because those models do not model the regressors space and, consequently, the inversion solution of the model almost certainly does not respect previous structural relationships in the regressors space, leading to unfeasible solutions. By contrast, when inverting a latent variable regression model the inversion solution belongs to the latent space (defined by the latent variables) and, therefore, such solution is constrained to be physically feasible and consistent with the sets of process conditions and correlation structure from the past. In this respect, the latent variable regression model inversion has been demonstrated to be a valid tool to support the development of new products and their manufacturing conditions using historical data in several case studies [12,19–23].

When considering the inversion of a SMB-PLS model, the set of input variables (column vector $\begin{bmatrix} \boldsymbol{z}^{new} \\ \boldsymbol{x}^{new} \end{bmatrix}$) that will yield the desired set of CQAs (column vector $\boldsymbol{y}^{des}$) are obtained by solving the following system of linear equations:

$$\boldsymbol{y}^{des} = \mathbf{Q} \cdot \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}}^{new} \\ \boldsymbol{\tau}_{ortho}^{new} \end{bmatrix} = \mathbf{Q} \cdot \boldsymbol{\tau}^{new} \tag{2}$$

where $\boldsymbol{\tau}^{new}$ is the vector of scores corresponding to the observation $\begin{bmatrix} \boldsymbol{z}^{new} \\ \boldsymbol{x}^{new} \end{bmatrix}$. The way to calculate $\boldsymbol{z}^{new}$ and $\boldsymbol{x}^{new}$ from $\boldsymbol{\tau}^{new}$ is explained in Section 5. Notice that the SMB-PLS model inversion involves solving a system of linear equations represented in a matrix form (Eq. (2)), where there are as many linear independent equations as the rank of **Y** (r_Y), and the number of unknown variables corresponds to the dimensionality of the latent space (*A*). Commonly, r_Y is lower than *A* and, hence, Eq. (2) corresponds to an underdetermined system of linear equations. The multiple solutions $\boldsymbol{\tau}^{\mathrm{new}}$ fall into a $(A - r_Y)$- dimensional hyper-plane of the *A*-dimensional space, that theoretically yields the same desired set of CQAs [9]. This hyper-plane is so-called Null Space (NS).

4. High-Confidence Design Space

In this section, a brief overview of the DS is shown based on Borràs-Ferrís et al. [10], but by considering the process conditions by means of the SMB-PLS model instead of applying PLS as in Ref. [10]. This is possible as the SMB-PLS latent space (Eq. (1)) is expressed similarly as PLS. Thus, the DS refers to the multidimensional combination and interaction, not only of raw materials properties but also process conditions, that have been demonstrated to provide assurance of quality. If there is no prediction uncertainty, the DS must be defined as a region in the latent space associated with raw materials properties and process conditions such that they yield an expected value of CQAs within their specification limits. Besides, since SMB-PLS is an empirical model based on historical data, any new set of raw materials properties must respect the *correlation structure* and *range* of those historical data [19]. Regarding the *correlation structure*, since the DS is defined in the latent space, it ensures new observations to behave in the same way as the ones used to create the model, in the sense that the correlation structure of the model is respected. Regarding the historical *range*, when considering the Hotelling T^2 confidence limit as a raw material specification limit, the new set of raw material properties are constrained to be within historical *ranges* in a multivariate sense. Additionally, historical univariate *ranges* for each property (and other constraints) might be included.

In this study, we initially focus on the l-th CQA and, hence, vector $\boldsymbol{y}^{\mathrm{des}}$ degenerates to scalar y^{des}, and matrix $\mathbf{Q}$ degenerates to vector $\boldsymbol{q}_l^{\mathrm{T}}$ (l-th row of matrix $\mathbf{Q}$). Thus, if a specific value of the l-th CQA is required ($y_l = y_l^{des}$), the desired specific value for the l-th CQA yields a (*A-1*)-dimensional NS and the DS is defined by the intersection of this NS and the Hotelling's T^2 confidence region. In the same way, if the desired l-th CQA refers to both lower and upper specifications limits (y_l^{LSL} and y_l^{USL}, respectively), the DS in the latent space is defined by the intersection of the scores fulfilling the specifications' NSs and the Hotelling T^2 confidence region.

Until now, the DS has been defined without taking into account the prediction uncertainty. However, since empirical models are subject to uncertainty, when a latent variable regression model is inverted, the uncertainty is back-propagated to the calculated inputs (i.e., the DS calculation is probabilistic) [8,9]. For that reason, even though working in the NS associated with the specification limit leads to a predicted value between specifications, it might yield out of specifications values for the l-th CQA due to prediction uncertainties. In this sense, Borràs--Ferrís et al. [10] proposed to define the DS in the latent space as the region where any combination of raw properties and process conditions results in a prediction interval inside specifications. When calculating the confidence limit for the multiple solutions along the NS of y_l^{LSL} and y_l^{USL}, a non-linear boundary is obtained for each specification: Lower Specification Confidence Limit (LSCL) and Upper Specification Confidence Limit (USCL), respectively. Appendix A of Borràs-Ferrís et al. [10] shows the analytical expression, which allows calculating the score belonging to both the lower and upper specification confidence limits given its respective score in the NS for the l-th CQA.

The intersection regions delimited by the LSCL, USCL and the Hotelling T^2 confidence ellipsoid, delimits the so-called High-Confidence Design Space (HC DS). This is illustrated in a two-dimensional latent space, and the focus is on the l-th CQA (Fig. 2).

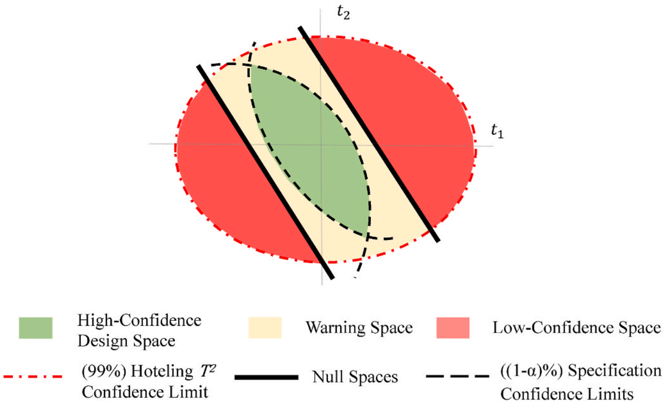

t_2
t_1
High-Confidence Design Space
Warning Space
Low-Confidence Space
(99%) Hoteling T^2 Confidence Limit
Null Spaces
((1-α)%) Specification Confidence Limits

Fig. 2. Defining the High-Confidence Design Space, Warning Space and Low-Confidence Space in the two-dimensional latent space.

From a frequentist probabilistic interpretation, batches belonging to the HC DS in Fig. 2 are expected to produce product with CQAs within specification limits with a confidence level equal or higher than $1-\alpha$. Additionally, the intersection between the region bounded by the two NSs corresponding to the y_l^{LSL} and y_l^{USL}, and the Hotelling's T^2 confidence region, but outside the High-Confidence DS, defines the so-called Warning Space (WS) (Fig. 2). Note that, although this space does not belong to the HC DS as defined, it does not necessarily imply the rejection of batches. In fact, batches lying within the WS lead to predicted values between specifications, but they result in prediction intervals for the CQA partially outside of specifications given the predefined confidence level $1-\alpha$. Finally, the Low-Confidence Space (Fig. 2) leads to predicted values outside specifications.

5. Multivariate raw material specification region

The HC DS, defined by the SMB-PLS model, simultaneously considers the raw material properties and process conditions. At this point, one could use such model to define the multivariate raw material specification region (i.e., the Raw Material HC DS) according to two strategies: without or under improved control.

5.1. Without improved control

In this section, it is assumed that process variations, correlated with raw material properties will remain in place in the future without any improvement. Thus, establishing specifications in raw material properties aims at penalizing those combinations that are not compensated for by the current control schemes.

A priori, in this strategy, there is no need to consider the orthogonal variations in process conditions and, hence, Raw Material DS refers to the HC DS of the SMB-PLS for $[\mathbf{Z}\ \mathbf{X}_{corr}]$. Thus, given a new raw material batch, z^{new}, its corresponding $\mathbf{Z}$ scores, $\boldsymbol{\tau}_{\mathbf{z}}^{new}$, the expected process conditions according to the control actions from the past, $\boldsymbol{x}_{corr}^{new}$, and its corresponding $\mathbf{X}_{corr}$ scores, $\boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new}$, are calculated according to Eq. (3).

$$\begin{aligned}\boldsymbol{\tau}_{\mathbf{z}}^{new} &= {\mathbf{W}^{*}}_{\mathbf{Z}}^{\mathrm{T}} \cdot z^{new}\\ \boldsymbol{x}_{corr}^{new} &= \mathbf{C}_{\mathbf{X}_{corr}} \cdot \boldsymbol{\tau}_{\mathbf{z}}^{new}\\ \boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new} &= {\mathbf{W}^{*}}_{\mathbf{X}_{corr}}^{\mathrm{T}} \cdot \boldsymbol{x}_{corr}^{new}\end{aligned} \tag{3}$$

where $\mathbf{W}_{\mathbf{Z}}^{*}$ is the $\mathbf{Z}$ block weights transformed to be independent between components, $\mathbf{C}_{Xcorr}$ is the correlation coefficient matrix calculated in the first modelling step which directly relate $\boldsymbol{\tau}_{\mathbf{z}}^{new}$ to $\boldsymbol{x}_{corr}^{new}$, and $\mathbf{W}_{\mathbf{X}_{corr}}^{*}$ is the $\mathbf{X}_{corr}$ block weights transformed to be independent between components. Both, $\mathbf{W}_{\mathbf{Z}}^{*}$ and $\mathbf{W}_{\mathbf{X}_{corr}}^{*}$, are calculated in the first modelling step as it is shown in Appendix B. The corresponding projection into the first block of latent variables, $\boldsymbol{\tau}_{\mathbf{T}}^{new}$, are obtained in the super level score matrix (Eq. (4)).

$$\boldsymbol{\tau}_{\mathbf{T}}^{new} = diag\left(\left[\boldsymbol{\tau}_{\mathbf{z}}^{new}\ \boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new}\right] \cdot \mathbf{W}_{\mathbf{T}}\right) \tag{4}$$

where $[\boldsymbol{\tau}_{\mathbf{z}}^{new}\ \boldsymbol{\tau}_{\mathbf{X}_{corr}}^{new}]$ refers to the matrix of concatenated score vectors ($A \times 2$), $\mathbf{W}_{\mathbf{T}}$ is the super weight matrix containing the super weight vectors organized by columns ($2 \times A$), and *diag* is the matrix-to-vector diagonal operator. Then, if any point, $\boldsymbol{\tau}_{\mathbf{T}}^{new}$ is within the HC DS, one would expect good quality with a certain confidence level for such z^{new}. Hence, the Raw Material HC DS (i.e., RM HC DS) can be defined as Eq. (5).

$$\text{RM HC DS} := \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in \text{HC DS}\} \tag{5}$$

In the case of considering also $\mathbf{X}_{orth}$ in the second modelling step, the Raw Material HC DS would refer to the space defined in Eq. (6).

$$\begin{aligned}&\text{Ort} := \left\{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{orth}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathbf{T}}}, \boldsymbol{\tau}_{orth} = \boldsymbol{\tau}_{orth}^{new}\right\}\\ &RM\ \text{HC DS} := \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in \text{HC DS} \cap \text{Ort}\}\end{aligned} \tag{6}$$

Note that, the Raw Material HC DS defined in Eq. (6) a priori requires that the vector of scores referring to orthogonal variations in process conditions, $\boldsymbol{\tau}_{orth}^{new}$, is known beforehand. If this is not the case, it is assumed that $\boldsymbol{\tau}_{orth}^{new}$ will remain on average with respect to the past (i.e., $\boldsymbol{\tau}_{orth}^{new} = \mathbf{0}_{A_{orth}}$ where $\mathbf{0}_{A_{orth}}$ is a zero vector of size A_{orth}). However, as it is unknown, the confidence limits must be calculated disregarding the orthogonal latent space. In other words, the prediction uncertainty, back-propagated in the definition of the specification confidence limits, must be estimated assuming that the $\boldsymbol{\tau}_{orth}^{new}$ remain on average with respect to the past.

5.2. Under improved control

Several works have already emphasized the control actions from the past could be improved in order to compensate for some of the raw materials variability [6,7,24]. Hence, wider raw materials specifications can be used if an effective process control system attenuating most raw material variations is implemented. In this sense, the SMB-PLS is particularly useful approaching this strategy as it models the orthogonal variations in process conditions in a second block of latent variables being orthogonal to the first one. Thus, one can infer causality interpretations in the reduced latent space of the second block. This information offers an effective way of manipulating the process conditions, with respect to the control actions from the past, for compensating raw material variations.

In this strategy, given a new raw material batch, z^{new}, the expected process conditions according to the control actions from the past, $\boldsymbol{x}_{corr}^{new}$, and the first block of latent variables, $\boldsymbol{\tau}_{\mathbf{T}}^{new}$, are obtained as above. Then, any raw material batch, resulting in $\boldsymbol{\tau}_{\mathbf{T}}^{new}$, is expected to have good quality with a certain confidence level by modifying process conditions (i.e., it belongs to the Raw material HC DS), if and only if there is any $\boldsymbol{\tau}_{\mathrm{orth}} = \boldsymbol{\tau}_{orth}^{new}$ such that $\boldsymbol{\tau}^{new} = \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}}^{new} \\ \boldsymbol{\tau}_{\mathrm{orth}}^{new} \end{bmatrix}$ belongs to the HC DS, where $\boldsymbol{\tau}_{orth}^{new}$ is the score values of the second block of latent variables. From $\boldsymbol{\tau}_{orth}^{new}$, one can figure out how to manipulate the process conditions to compensate for raw material variations according to Eq. (7).

$$\boldsymbol{x}^{new} = \boldsymbol{x}_{corr}^{new} + \boldsymbol{x}_{orth}^{new} = \boldsymbol{x}_{corr}^{new} + \mathbf{P}_{orth} \cdot \boldsymbol{\tau}_{orth}^{new} \tag{7}$$

where $\mathbf{P}_{orth}$ is the loading matrix of the second latent block. Note that, $\boldsymbol{\tau}_{orth}^{new}$ represents the locus of the $\boldsymbol{x}_{orth}^{new}$ projections within the HC DS given a new raw material batch. Therefore, if it exists, the control actions could be improved in different ways without leaving the DS, which provides operational flexibility in process improvement.

Finally, we can define analytically the Raw Material HC DS, prior to selecting a new raw material, as the projection of the HC DS onto the space defined by the first block of latent variables as Eq. (8).

$$\text{RM HC DS} := \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} = T_{\mathbf{P}_{\mathbf{T}}}[\boldsymbol{\tau}], \forall \boldsymbol{\tau} \in \text{HC DS}\} \tag{8}$$

where $T_{\mathbf{P}_{\mathbf{T}}}$ is the linear transformation that projects from $\mathbb{R}^{A_{\mathrm{T}}+A_{orth}}$ to $\mathbb{R}^{A_{\mathrm{T}}}$ defined by the matrix $\mathbf{P}_{\mathbf{T}} = [\mathbf{I}_{A_{\mathrm{T}}}\ \mathbf{0}_{A_{\mathrm{T}},A_{orth}}]$, $\mathbf{I}_{A_{\mathrm{T}}}$ is the identity matrix of size $A_{\mathbf{T}}$, $\mathbf{0}_{A_{\mathrm{T}},A_{orth}}$ is a zero matrix of size $A_{\mathbf{T}} \times A_{orth}$, and $A_{\mathbf{T}}$ and A_{orth} are the latent dimensionality of the first and second block, respectively.

Note that saturation in the control actions caused by the actuators limits can be accounted for in the model inversion procedure [9]. This is feasible because the latent variable regression models allow projecting constraints on process conditions onto the latent space in order to delimit a portion of the HC DS within which products of the desired quality may still be produced while meeting saturation constraints.

6. Presence of known disturbances affecting control actions

Until now, we have assumed that orthogonal process variations to raw material properties and related to CQAs are due to certain (unplanned) excitations. However, process conditions could present variations due to feedforward compensation for some known disturbances.

This issue needs special attention as if one decides to ignore the known disturbance for not being manipulatable, the SMB-PLS could model, in the orthogonal block, variations in process conditions that may be related to such disturbance. The fact that the correlation between process conditions and the known disturbance could still explain variations in CQAs is because the control adjustment may not be perfect (i.e., the effect of the known disturbances is not removed completely). This will yield misleading causality relations in the reduced latent space. Therefore, we suggest adding an intermediate block $\mathbf{D}$ ($N \times O$) being a matrix of inputs which includes a total of O known disturbances measured in each one of the N batches of a particular raw material. Thus, the SMB-PLS algorithm includes an intermediate modelling step that captures the impact of variation in disturbances orthogonal to $\mathbf{Z}$ (i.e., $\mathbf{D}_{orth}$) on the process and on $\mathbf{Y}$, represented by latent variables $\mathbf{T}_{\mathbf{D}}$. This intermediate block of latent variables allows identifying control actions from the past implemented to compensate for disturbances not related to raw material properties. This ensures that the last modelling step only model certain (unplanned) excitations in process conditions, $\mathbf{X}_{orth}$, from which causality can be inferred.

In the same way as Subsection 5.1 but including the disturbance space, the Raw Material HC DS without improved control would refer to the space defined in Eq. (9).

$$\begin{aligned} Dis &:= \left\{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}, \boldsymbol{\tau}_{orth}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathrm{T}}}, \boldsymbol{\tau}_{\mathbf{D}} = \boldsymbol{\tau}_{\mathbf{D}}^{new}, \boldsymbol{\tau}_{orth} \in \mathbb{R}^{A_{orth}} \right\} \\ Ort &:= \left\{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}, \boldsymbol{\tau}_{orth}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathrm{T}}}, \boldsymbol{\tau}_{\mathbf{D}} \in \mathbb{R}^{A_{\mathrm{D}}}, \boldsymbol{\tau}_{orth} = \boldsymbol{\tau}_{orth}^{new} \right\} \\ RM\ HC\ DS &:= \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in HC\ DS \cap Ort \cap Dis\} \end{aligned} \tag{9}$$

Eq. (9) assumes that both the disturbance and the orthogonal space are not manipulatable and, hence, they must be defined as constraints, Dis and Ort respectively, that intersect with the HC DS. However, if control actions can be improved by means of the orthogonal space, such space must be projected onto the remaining space in the same way as Subsection 5.2. Thus, the Raw Material HC DS, by considering the possibility to modify process conditions prior to selecting a new raw material batch, can be defined analytically as the intersection between the projection of the HC DS onto the first and second block of latent variables (i.e., Pr), and the subspace defined by $\boldsymbol{\tau}_{\mathbf{D}}^{new}$ (i.e., Dis), as it is shown in Eq. (10).

$$\begin{aligned} \text{Dis} &:= \left\{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}) : \boldsymbol{\tau}_{\mathbf{T}} \in \mathbb{R}^{A_{\mathrm{T}}}, \boldsymbol{\tau}_{\mathbf{D}} = \boldsymbol{\tau}_{\mathbf{D}}^{new} \right\} \\ Pr &:= \left\{(\boldsymbol{\tau}_{\mathbf{T}}, \boldsymbol{\tau}_{\mathbf{D}}) : \begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}} \\ \boldsymbol{\tau}_{\mathbf{D}} \end{bmatrix} = T_{\mathbf{P}_{\mathbf{TD}}}[\boldsymbol{\tau}], \forall \boldsymbol{\tau} \in \text{HC DS} \right\} \\ \text{RM HC DS} &:= \{(\boldsymbol{\tau}_{\mathbf{T}}) : \boldsymbol{\tau}_{\mathbf{T}} \in \text{Dis} \cap Pr\} \end{aligned} \tag{10}$$

where $T_{\mathbf{P}_{\mathbf{TD}}}$ is the linear transformation that projects from $\mathbb{R}^{A_{\mathrm{T}}+A_{\mathrm{D}}+A_{orth}}$ to $\mathbb{R}^{A_{\mathrm{T}}+A_{\mathrm{D}}}$ defined by the matrix $\mathbf{P}_{\mathbf{TD}} = [\mathbf{I}_{A_{\mathrm{T}}+A_{\mathrm{D}}}\ \mathbf{0}_{A_{\mathrm{T}}+A_{\mathrm{D}},A_{orth}}]$, $\mathbf{I}_{A_{\mathrm{T}}+A_{\mathrm{D}}}$ is the identity matrix of size $A_{\mathbf{T}} + A_{\mathbf{D}}$, $\mathbf{0}_{A_{\mathrm{T}}+A_{\mathrm{D}},A_{orth}}$ is a zero matrix of size $A_{\mathbf{T}} + A_{\mathbf{D}} \times A_{orth}$, and $A_{\mathbf{D}}$ is the latent dimensionality of the disturbance block.

Note that, the Raw Material HC DS defined in Eq. (9) and Eq. (10) a priori requires that the vector of scores referring to orthogonal variations in disturbances, $\boldsymbol{\tau}_{\mathbf{D}}^{new}$, is known beforehand. If this is not the case, it is assumed that $\boldsymbol{\tau}_{\mathbf{D}}^{new}$ will remain on average with respect to the past (i.e., $\boldsymbol{\tau}_{\mathbf{D}}^{new} = \mathbf{0}_{A_{\mathrm{D}}}$ where $\mathbf{0}_{A_{\mathrm{D}}}$ is a zero vector of size $A_{\mathbf{D}}$), and the confidence limits must be calculated disregarding the disturbance latent space.

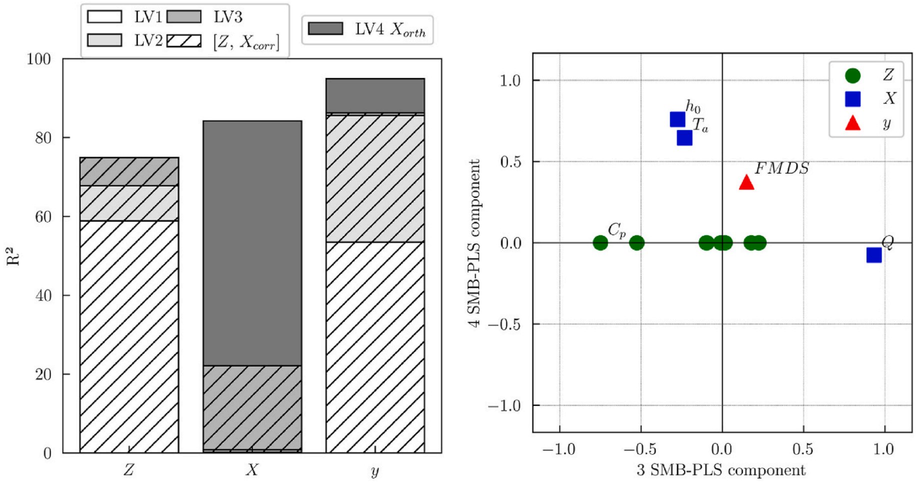


Fig. 3. Explained **Z**, **X** and **y** variability for the SMB-PLS model depending on either the number of latent variables (LVs) or the two blocks of latent variables (LV1-LV3 explain the first block $[\mathbf{Z}\ \mathbf{X}_{\text{corr}}]$, and LV4 explains the second block $\mathbf{X}_{\text{orth}}$) (left), and bi-plot of the block weights and **y** loadings for last two components (right).

7. Industrial simulated case study

7.1. Description of the dataset

A simulated polymer extrusion film blowing process was used to generate data in order to illustrate how to define multivariate specification regions for incoming raw materials [3,14]. The dataset consists of two regressor blocks (**Z** and **X**) and a response block (**Y**). The raw material block (**Z**) contains the following polymer resin properties: ten temperature dependent viscosities (η), heat capacity (C_p), and density (ρ). The second block (**X**) contains 3 process conditions, namely the air temperature (T_a), the polymer flow rate (Q) and the cooling air flow rate represented by the maximum local heat transfer coefficient along the film bubble (h_0). A priori, these process conditions are assumed to be manipulated variables, but in Section 7.6, T_a is considered a known process disturbance. The response block (***y***) is characterized by one quality attribute of the film, which is the full stress in the machine direction (FMDS), with a lower specification defined as its average.

The dataset was simulated in two steps. First, variability was introduced in raw material properties and process conditions in such a way that both regressor blocks affect ***y***, but variations in **Z** and **X** are uncorrelated to each other (initially blocks are orthogonal). This was achieved by introducing random variations in raw material properties (**Z**) and processing conditions (**X**) to simulate their effect on product quality. However, the variables within each block are collinear to a certain extent. Regarding **Z**, correlation is due to viscosities measured at different temperatures, and for **X**, h_0 must be adjusted to compensate for variations in T_a. In a second step, similar uncorrelated variations were again implemented in both regressor blocks, but between block correlations were introduced by a feedforward controller, added to attenuate variations caused by raw material properties. This controller corrects for some of the variability in the polymer heat capacity C_p by adjusting the flow rate Q. The processing of 50 raw materials batches were simulated. The simulated data that support the findings of this study are available on request from the authors.

7.2. Building the SMB-PLS model

Three components were found sufficient to capture the impact of raw material properties (and correlated process variations) on ***y*** in the first modelling step. One additional component was also needed in the second modelling step to model the effect of orthogonal variations in process conditions on the remaining variations in ***y***. The goodness of fit, R^2 (i.e., variability percentage explained by the model) for each one of the input blocks, **Z** and **X**, and the output block, ***y***, and each component, is presented in Fig. 3a.

Fig. 3a shows that the first three components of the first modelling stage explain 74.89% of the information in **Z** and 22.10% of the information in **X** that was correlated with **Z**, to explain a great percentage of the response variability (86.22%). Component 4 (the unique component of the second modelling stage) shows that the 62.12% of the variation in **X**, not related to **Z**, is able to explain 8.65% of the response variability. Since the last two components explain the greatest variation in **X** (Fig. 3a) and b shows the bi-plot of the block weights and ***y*** loadings for these components to understand the behavior of process conditions. Fig. 3b reveals that the explained variation in the polymer flow rate Q seems to be related to raw material properties according to the third component. In fact, Q is strongly negatively correlated with the heat capacity C_p because when C_p increases, Q is reduced (as a result of the feedforward controller) to mitigate its impact on quality product. However, component 4 shows that Q barely presents orthogonal variations to raw material properties related to ***y***. By contrast, the air temperature T_a and the cooling air flow rate, which effect is represented by a change in h_0, present orthogonal variations to raw material properties highly correlated with each other, from which one can infer causality in the reduced latent space. In other words, for any active change in the process conditions of T_a and h_0, being consistent with the correlation structure modeled by the latent orthogonal space, the SMB-PLS model will reliably predict the changes in ***y***.

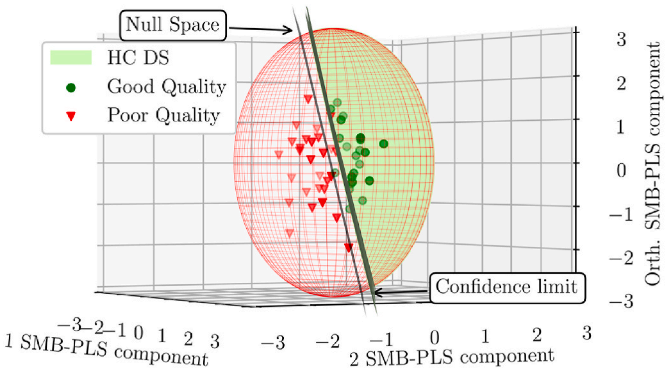


Fig. 4. Graphical definition of the High-Confidence Design Space by showing calibration data.

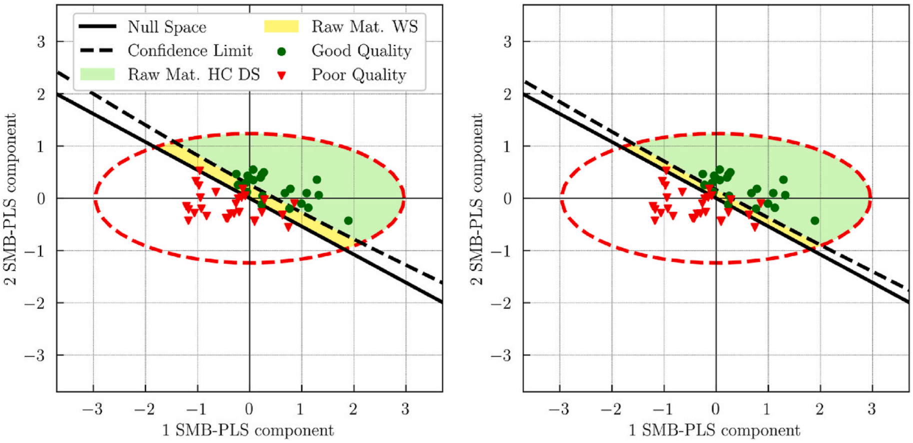


Fig. 5. Graphical definition of the Raw Material High-confidence Design Space (Multivariate Raw Material Specification region) and Raw Material Warning Space when considering only $[\mathbf{Z}\ \mathbf{X}_{\text{corr}}]$ (left) and also $\mathbf{X}_{\text{orth}}$ assuming that orthogonal variations remain at the average value (right).

7.3. Defining the High-Confidence Design Space

The HC DS is defined with at least a 90% confidence level of obtaining superior or equal FMDS values to the average of calibration data (lower specification limit). Fig. 4 shows the HC DS by showing the calibration data for the first two components of the first modelling step $[\mathbf{Z}\ \mathbf{X}_{corr}]$ and the orthogonal one. The third $[\mathbf{Z}\ \mathbf{X}_{corr}]$ component from the first modelling step is omitted.

One would expect that, of the batches lying within the HC DS, 90% or more would be acceptable batches. Indeed, the negative predictive value[1] is 95%. On the other hand, the HC DS leads to 3.57% type I risk and 13.64% type II risk. This means that if only batches lying within the HC DS are accepted, 13.64% of unacceptable batches of raw materials had been accepted at the expense of rejecting 3.57% of acceptable batches.

7.4. Multivariate raw material specification region without improved control

In this section, it is assumed that process variations, correlated with raw material properties due to control actions through manipulated variables, will remain in place in the future without any improvement. In such a case, a priori there is no need to consider process conditions to establish the specification region associated with the raw material properties and, hence, one could define this region by the PLS model inversion by considering only raw material properties as in Ref. [10]. By contrast, without improved control, we propose to define the raw material HC DS as the HC DS of the first block of latent variables, referring to $[\mathbf{Z}\ \mathbf{X}_{corr}]$ of the SMB-PLS. The amount of information/variability contained in the first input block depends on $\mathbf{Z}$ as $\mathbf{X}_{corr}$ does not provide a new source of variability. Therefore, the predictive power of both, PLS for $\mathbf{Z}$ with three components and SMB-PLS for only $[\mathbf{Z}\ \mathbf{X}_{corr}]$ with three components, are the same. Consequently, a priori, the classification performance of new raw material batches is expected to be equivalent. However, incorporating process data by means of the SMB-PLS presents some advantages with respect to PLS as we will see below.

As Azari et al. [6] discussed, the SMB-PLS provides great insights in agreement with process knowledge for the effects of material variations and correlated process conditions (control schemes mainly). Firstly, since the SMB-PLS also can model the orthogonal variations in process conditions by the second block of latent variables, it provides a great capability for diagnosing assignable causes of such variations. In fact, by interrogating the underlying SMB-PLS model, one can extract diagnostic or contribution plots which reveal the group of process conditions making the greatest contributions to the deviations in the squared prediction errors, and the scores [15,25]. In addition to that, the second block of latent variables provides a better understanding of the response variability with respect to both PLS and SMB-PLS for only $[\mathbf{Z}\ \mathbf{X}_{corr}]$ (this increases the response variance percentage up to 95.87%). The latter results in less prediction uncertainty, and this affects the definition of the Raw Material HC DS. Fig. 5 shows the graphical definition of such space for the SMB-PLS depending on whether the $\mathbf{X}_{orth}$ is considered or not.

Fig. 5 shows graphically that, as expected, the Raw Material DS without uncertainty (i.e., the union of the Raw Material HC DS and the Raw Material WS) are equal regardless of whether $\mathbf{X}_{orth}$ is considered or not. In addition to that, the less uncertainty there is, the more similar the Raw Material HC DS and the Raw Material DS without uncertainty are. For that reason, the SMB-PLS Raw Material HC DS becomes wider when incorporating the $\mathbf{X}_{orth}$ block as can see in Fig. 5. Therefore, it can be concluded that, for model building, the SMB-PLS provides useful information in order to achieve a higher level of process understanding when considering the $\mathbf{X}_{orth}$. However, it is crucial to bear in mind that for exploiting the model, Fig. 5b requires that orthogonal variations are known beforehand. Indeed, the Raw Material HC DS shown in Fig. 5b arises from the assumption that the orthogonal variation in process conditions will remain at the average value with respect to the past. If they are not known beforehand, it is assumed that $\boldsymbol{\tau}_{\mathbf{D}}^{new}$ will remain on average with respect to the past and, hence, the confidence limits must be calculated disregarding the orthogonal block yielding Fig. 5a.

7.5. Multivariate raw material specification region under improved control

Let us considerer the HC DS defined previously (see Fig. 4). Then, a new raw material batch is considered prior to the manufacturing process (i.e., only raw material properties are known). Thus, the red triangle in Fig. 6 refers to the projection onto the latent space assuming that the control actions of process conditions remain in place, and the orthogonal variation in process conditions remain at the average value with respect to the past (i.e., the orthogonal component is null): $\begin{bmatrix} \boldsymbol{\tau}_{\mathbf{T}}^{new} \\ 0 \end{bmatrix}$.

In such a case, as shown in Fig. 6, this batch would be outside the specification region. However, if the orthogonal component is modified orthogonally, such batch can become part of the specification region

[1] The negative predictive is the proportion of batches that actually result in a good product out of all those within the HC DS.

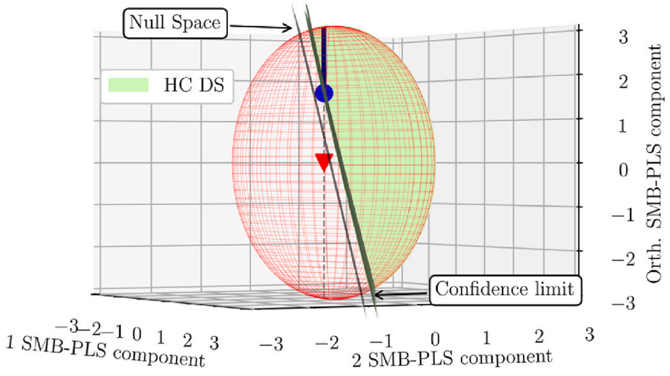


Fig. 6. Graphical definition of the High-Confidence Design Space by showing the projection of the new raw material batch when: i) orthogonal variation in process conditions remain at the average value with respect to the past (red triangle), and ii) control actions are improved (blue circle). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

(deep blue solid line). This is a batch that, a priori, would give place to a film with an unacceptable response value (FMDS), but that by improving the control actions it would yield a film with an acceptable response value (FMDS). As commented, since the control actions could be improved in different ways without leaving the HC DS, it provides operational flexibility in process improvement. As an example, the blue circle is selected among all process conditions yielding the score: $\begin{bmatrix}\tau_{\mathbf{T}}^{new}\\ \tau_{orth}^{new}\end{bmatrix}$. This solution belongs to the latent space and, therefore, it behaves in the same way as the ones used to create the model, in the sense that the correlation structure of the model is respected. A logical question then arises: how to manipulate the process conditions to get this solution? The answer is applying Eq. (7). This is shown graphically in Fig. 7.

Fig. 7 shows the time series of manipulated variables with their historical limits. The red triangles refer to the expected process conditions due to the control actions from the past, $\boldsymbol{x}_{corr}^{new}$, while the blue circles show the final conditions after improving such control for compensating raw material variations. The latter arises from adding the orthogonal variation, $\boldsymbol{x}_{orth}^{new}$, which is obtained as $\mathbf{P}_{orth}\cdot\tau_{orth}^{new}$. As expected, the flow rate Q is barely modified with respect to the expected control actions because, as it is shown in Fig. 3b, this process condition does not present a significant amount of orthogonal variation related to y. By contrast,

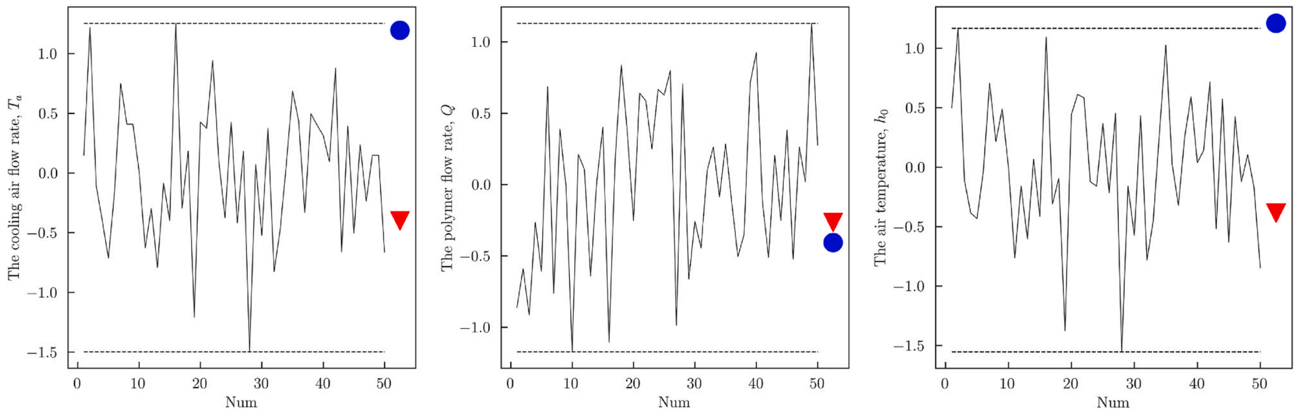


Fig. 7. Time series of process conditions, T_a, Q and h_0, and new setpoints in two scenarios: no improved control (red triangle) and under improved control (blue circle). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

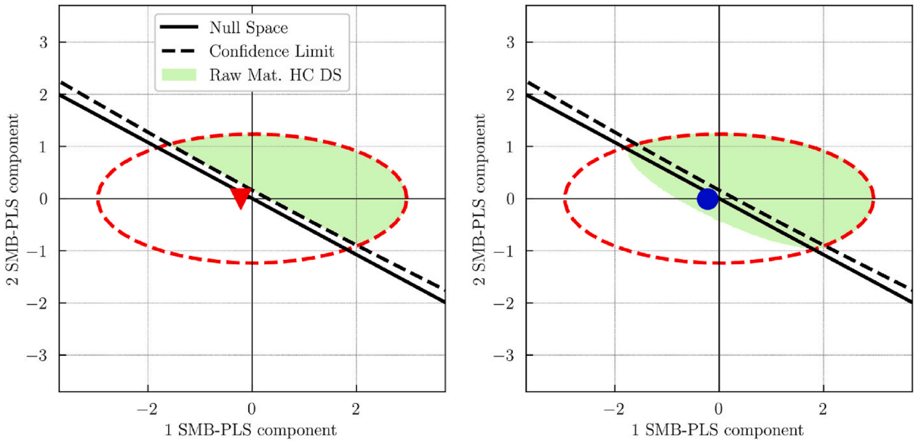


Fig. 8. Raw Material High-Confidence Design Space without improved control by showing the projection of the new raw material batch as a red triangle (left), and under improved control by showing the projection of the new raw material batch as a blue circle (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

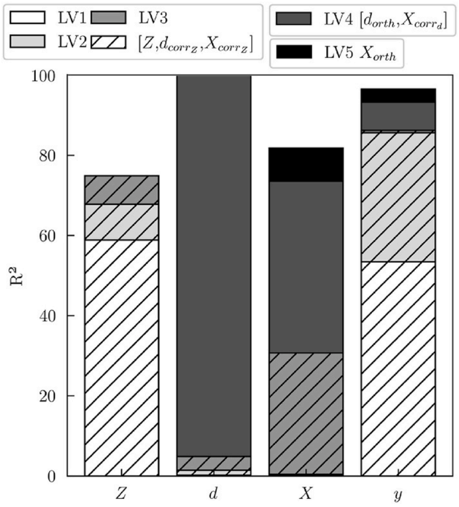


Fig. 9. Explained **Z**, **d**, **X** and **y** variability for the SMB-PLS model depending on either the number of latent variables (LVs) or the three blocks of latent variables (LV1-LV3 explain the first block $[\mathbf{Z}\ \mathbf{d}_{\mathrm{corr_Z}}\ \mathbf{X}_{\mathrm{corr_Z}}]$, LV4 explains the second block, $[\mathbf{d}_{\mathrm{orth}}\ \mathbf{X}_{\mathrm{corr_d}}]$, and LV5 explains the last block $\mathbf{X}_{\mathrm{orth}}$).

the air temperature T_a and the cooling air flow rate h_0 do and, hence, one can infer causality in the reduced latent space in order to attenuate most raw material variations. Note that, since causality is inferred in the reduced latent space, process conditions are manipulated being consistent with the latent orthogonal space shown in Fig. 3b.

Finally, the Raw Material HC DS, by considering the possibility of modifying process conditions prior to selecting a new raw material batch, can be defined analytically as the projection of the HC DS onto the space defined by the first block of latent variables, according to Eq. (8) (see Fig. 8b).

Fig. 8 shows that Raw Material HC DS is expanded when considering the possibility to modify process conditions for compensating raw material variations. Thus, one may be able to accept raw materials that will yield products with perfectly satisfactory quality properties as a consequence of the process conditions modification, as in the considered new raw material batch.

7.6. *Presence of known disturbances affecting control actions*

Process conditions could present variations due to feedforward compensation for some known disturbances. In fact, in the simulated polymer extrusion film blowing process, the air temperature T_a refers to the air ambient temperature. In such a case, this process condition cannot be manipulated but it is a major process known disturbance affecting cooling conditions and hence, quality properties. In addition to that, the cooling air flow rate h_0 is manipulated by a feedforward controller to compensate for some variations in the ambient air temperature T_a. To identify these variations as explained in Section 6, an intermediate block **D** must be added. In this case, since there is only one known disturbance, the intermediate block is defined as vector $\boldsymbol{d}$. The goodness of fit for the SMB-PLS model, R^2, for each one of the input blocks, **Z**, $\boldsymbol{d}$ and **X**, and the output block, $\boldsymbol{y}$, and each component, is presented in Fig. 9.

Fig. 9 shows that three components were found sufficient to capture the impact of raw material properties (and correlated disturbances and process variations) on $\boldsymbol{y}$ in the first modelling step explaining 74.89% of the information in **Z**, 4.89% of the information in $\boldsymbol{d}$ and 30.70% of the information in **X**, to explain a high percentage of the response variability (86.22%). One additional component was also needed in the second modelling step to model the effect of orthogonal variations in disturbances (and correlated process variations) on the remaining variations in $\boldsymbol{y}$. This component shows that the 95.11% of the variation in $\boldsymbol{d}$, not related to **Z**, is able to explain 42.83% of **X** and 7.02% of the response variability. The latter represents variations in $\boldsymbol{d}$ affecting $\boldsymbol{y}$, but not compensated by the controller. Finally, one component was used to capture the orthogonal variations in process variations on the remaining variations in $\boldsymbol{y}$ showing that the 8.28% of variation in **X**, not related to **Z** and $\boldsymbol{d}$, is able to explain 3.27% percentage of the response variability.

The most common case is that the ambient air temperature T_a is not known when receiving a raw material batch. Therefore, it is assumed that, for exploiting the model, this disturbance remains on average with

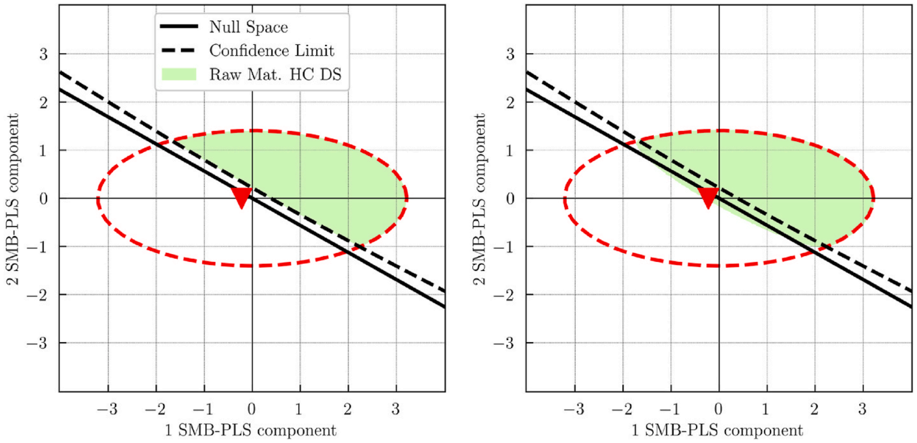


Fig. 10. Raw Material High-Confidence Design Space prior to knowing T_a without improved control (left) and under improved control (right) by showing the projection of the new raw material batch as a red triangle. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

respect to the past and, hence, the confidence limits are calculated disregarding the disturbance block. Thus, Fig. 10 shows the Raw Material HC DS prior to knowing T_a without improved control (Fig. 10a), and by considering the possibility to modify process conditions (Fig. 10b), using Eq. (9) and Eq. (10), respectively.

Fig. 10 shows that the Raw Material HC DS is slightly expanded when considering the possibility to modify process conditions for compensating raw material variations. Indeed, the new raw material batch illustrated in Subsection 7.5 would be on the border of the Raw Material HC DS, since there is no control action that allows to be within the HC DS. This happens because only 3.27% of the response variability can be inferred as the effect of 8.28% of the variation in **X** not related to **Z** and. The latter may not be sufficient to carry out effective improvement in the control action.

8. Conclusions

In this paper, we propose a novel approach for defining analytically the multivariate raw material specification region by considering the possibility to modify process conditions to compensate for raw material properties variations. This methodology is based on the SMB-PLS model inversion where prediction uncertainty is back-propagated. The most remarkable advantages of the proposal approach are.

- It can be used with historical data (i.e., daily production data not coming from any experimental design but with varying raw material properties, typical from Industry 4.0 environment).
- It considers a multivariate approach providing much insight into the correlated nature of raw material properties and process conditions. Besides, the SMB-PLS does identify the variation in process conditions uncorrelated with raw material properties and known disturbances, which is crucial to implement an effective process control system attenuating most raw material variations.
- It allows expanding the multivariate raw material specification when considering the possibility to modify process conditions and, hence, one may potentially be able to accept lower cost raw materials that will yield products with perfectly satisfactory quality properties.

Definitely, this methodology takes advantage of the variation in process conditions uncorrelated with raw material properties and known disturbances to expand the raw material specification. However, this variation may result insufficient to carry out effective improvement. In such a case, process excitation would be needed by running design of experiments on process operating conditions.

Author statement

Conceptualization Ideas; formulation or evolution of overarching research goals and aims. Alberto and Carl.

Methodology Development or design of methodology; creation of models Alberto, Carl and Joan.

Software Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components. Joan.

Validation Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs. Carl and Joan.

Formal analysis Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data. Joan.

Investigation Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection. Alberto, Carl and Joan.

Resources Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools. Alberto, Carl and Joan.

Data Curation Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse. Carl and Joan.

Writing – original draft preparation Creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation). Joan.

Writing – review and editing Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages. Alberto and Carl.

Visualization Preparation, creation and/or presentation of the published work, specifically visualization/data presentation. Joan.

Supervision Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team. Alberto and Carl.

Project administration Management and coordination responsibility for the research activity planning and execution. Alberto.

Funding **acquisition** of the financial support for the project leading to this publication. Alberto.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

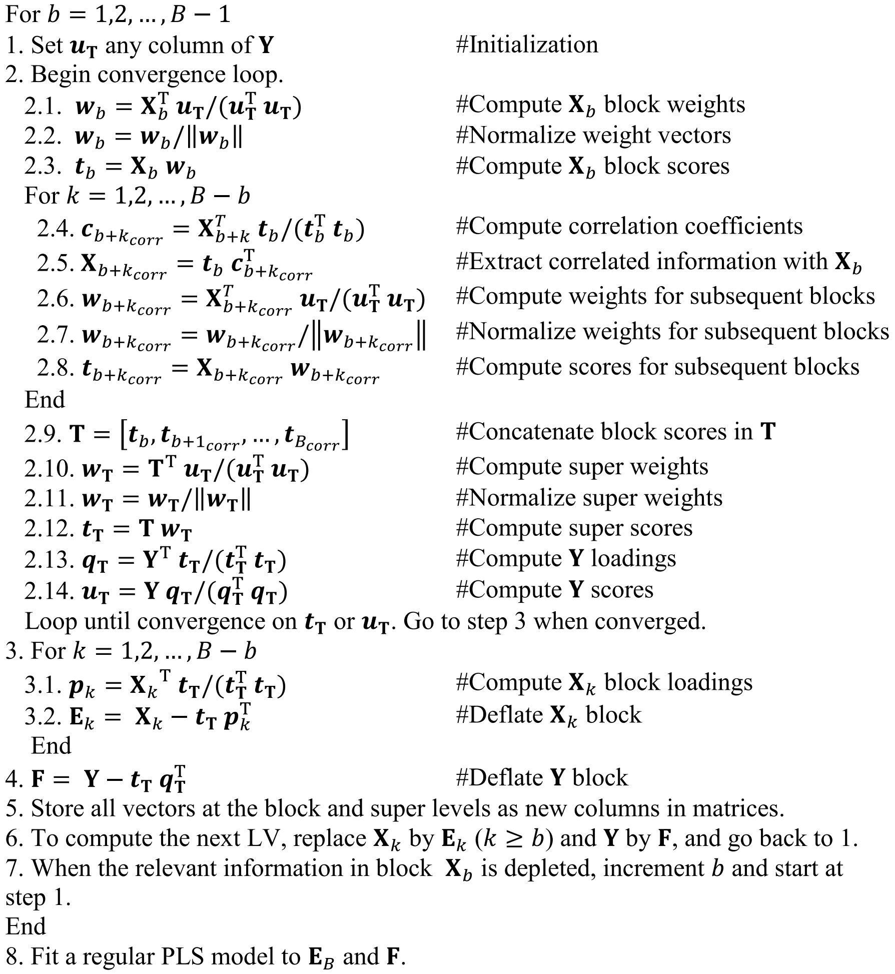
Data will be made available on request.

Acknowledgements

This work was partially supported by the Spanish Ministry of Science and Innovation (PID2020-119262RB-I00), the Generalitat Valenciana (AICO/2021/111) and the European Social Fund (ACIF/2018/165).

Appendix A. Pseudocode for the SMB-PLS

The pseudocode of the SMB-PLS assuming a process with *B* blocks is similarly as in Ref. [14].



```
For $b = 1,2,\dots,B-1$
1. Set $\boldsymbol{u}_{\mathbf{T}}$ any column of $\mathbf{Y}$          #Initialization
2. Begin convergence loop.
  2.1. $\boldsymbol{w}_b = \mathbf{X}_b^{\mathrm{T}}\,\boldsymbol{u}_{\mathbf{T}}/(\boldsymbol{u}_{\mathbf{T}}^{\mathrm{T}}\,\boldsymbol{u}_{\mathbf{T}})$          #Compute $\mathbf{X}_b$ block weights
  2.2. $\boldsymbol{w}_b = \boldsymbol{w}_b/\|\boldsymbol{w}_b\|$          #Normalize weight vectors
  2.3. $\boldsymbol{t}_b = \mathbf{X}_b\,\boldsymbol{w}_b$          #Compute $\mathbf{X}_b$ block scores
  For $k = 1,2,\dots,B-b$
   2.4. $\boldsymbol{c}_{b+k_{corr}} = \mathbf{X}_{b+k}^{T}\,\boldsymbol{t}_b/(\boldsymbol{t}_b^{\mathrm{T}}\,\boldsymbol{t}_b)$          #Compute correlation coefficients
   2.5. $\mathbf{X}_{b+k_{corr}} = \boldsymbol{t}_b\,\boldsymbol{c}_{b+k_{corr}}^{\mathrm{T}}$          #Extract correlated information with $\mathbf{X}_b$
   2.6. $\boldsymbol{w}_{b+k_{corr}} = \mathbf{X}_{b+k_{corr}}^{T}\,\boldsymbol{u}_{\mathbf{T}}/(\boldsymbol{u}_{\mathbf{T}}^{\mathrm{T}}\,\boldsymbol{u}_{\mathbf{T}})$ #Compute weights for subsequent blocks
   2.7. $\boldsymbol{w}_{b+k_{corr}} = \boldsymbol{w}_{b+k_{corr}}/\|\boldsymbol{w}_{b+k_{corr}}\|$ #Normalize weights for subsequent blocks
   2.8. $\boldsymbol{t}_{b+k_{corr}} = \mathbf{X}_{b+k_{corr}}\,\boldsymbol{w}_{b+k_{corr}}$          #Compute scores for subsequent blocks
  End
  2.9. $\mathbf{T} = [\boldsymbol{t}_b, \boldsymbol{t}_{b+1_{corr}}, \dots, \boldsymbol{t}_{B_{corr}}]$          #Concatenate block scores in $\mathbf{T}$
  2.10. $\boldsymbol{w}_{\mathbf{T}} = \mathbf{T}^{\mathrm{T}}\,\boldsymbol{u}_{\mathbf{T}}/(\boldsymbol{u}_{\mathbf{T}}^{\mathrm{T}}\,\boldsymbol{u}_{\mathbf{T}})$          #Compute super weights
  2.11. $\boldsymbol{w}_{\mathbf{T}} = \boldsymbol{w}_{\mathbf{T}}/\|\boldsymbol{w}_{\mathbf{T}}\|$          #Normalize super weights
  2.12. $\boldsymbol{t}_{\mathbf{T}} = \mathbf{T}\,\boldsymbol{w}_{\mathbf{T}}$          #Compute super scores
  2.13. $\boldsymbol{q}_{\mathbf{T}} = \mathbf{Y}^{\mathrm{T}}\,\boldsymbol{t}_{\mathbf{T}}/(\boldsymbol{t}_{\mathbf{T}}^{\mathrm{T}}\,\boldsymbol{t}_{\mathbf{T}})$          #Compute $\mathbf{Y}$ loadings
  2.14. $\boldsymbol{u}_{\mathbf{T}} = \mathbf{Y}\,\boldsymbol{q}_{\mathbf{T}}/(\boldsymbol{q}_{\mathbf{T}}^{\mathrm{T}}\,\boldsymbol{q}_{\mathbf{T}})$          #Compute $\mathbf{Y}$ scores
  Loop until convergence on $\boldsymbol{t}_{\mathbf{T}}$ or $\boldsymbol{u}_{\mathbf{T}}$. Go to step 3 when converged.
3. For $k = 1,2,\dots,B-b$
  3.1. $\boldsymbol{p}_k = \mathbf{X}_k{}^{\mathrm{T}}\,\boldsymbol{t}_{\mathbf{T}}/(\boldsymbol{t}_{\mathbf{T}}^{\mathrm{T}}\,\boldsymbol{t}_{\mathbf{T}})$          #Compute $\mathbf{X}_k$ block loadings
  3.2. $\mathbf{E}_k = \mathbf{X}_k - \boldsymbol{t}_{\mathbf{T}}\,\boldsymbol{p}_k^{\mathrm{T}}$          #Deflate $\mathbf{X}_k$ block
   End
4. $\mathbf{F} = \mathbf{Y} - \boldsymbol{t}_{\mathbf{T}}\,\boldsymbol{q}_{\mathbf{T}}^{\mathrm{T}}$          #Deflate $\mathbf{Y}$ block
5. Store all vectors at the block and super levels as new columns in matrices.
6. To compute the next LV, replace $\mathbf{X}_k$ by $\mathbf{E}_k$ ($k \geq b$) and $\mathbf{Y}$ by $\mathbf{F}$, and go back to 1.
7. When the relevant information in block $\mathbf{X}_b$ is depleted, increment $b$ and start at
step 1.
End
8. Fit a regular PLS model to $\mathbf{E}_B$ and $\mathbf{F}$.
```

Appendix B

This appendix is applicable to blocks, $\mathbf{Z}$ and $\mathbf{X}_{corr}$ (hereinafter called $\mathbf{B}$).

The weights matrix, $\mathbf{W_B}$, do not directly relate the matrix $\mathbf{B}$ to the score matrix $\mathbf{T_B}$, as $\mathbf{B}$ is deflated after each component by the loading matrix $\mathbf{P_B}$. However, the weights, $\mathbf{W_B}$, can be transformed to $\mathbf{W_B^*}$ by $\mathbf{M}$ (Eq. (B.1)) and, thus, $\mathbf{W_B^*}$ does directly relate $\mathbf{B}$ to $\mathbf{T_B}$, (Eq. (B.2)).

$$\mathbf{W_B^*} = \mathbf{W_B}\cdot\mathbf{M} \tag*{Eq. (B.1)}$$

$$\mathbf{T_B} = \mathbf{B}\cdot\mathbf{W_B^*} = \mathbf{B}\cdot\mathbf{W_B}\cdot\mathbf{M} \tag*{Eq. (B.2)}$$

If multiplying both sides of Eq. (B.2) by the transpose of the super score matrix, $\mathbf{T_T}$, the $\mathbf{M}$ matrix can be expressed as Eq. (B.3).

$$\mathbf{M} = \left(\mathbf{T_T^T}\cdot\mathbf{B}\cdot\mathbf{W_B}\right)^{-1}\cdot\left(\mathbf{T_T^T}\cdot\mathbf{T_B}\right) \tag*{Eq. (B.3)}$$

On the other hand, $\mathbf{T_T}$ are good “summaries” of $\mathbf{B}$ according to the loading matrix $\mathbf{P_B}$ (Eq. (B.4)).

$$\mathbf{B} = \mathbf{T_T}\cdot\mathbf{P_B}^{\mathrm{T}} + \mathbf{E_B} \tag*{Eq. (B.4)}$$

where $\mathbf{E_B}$ is the residual matrix. Then, multiplying both sides of Eq. (B.4) by the transpose of $\mathbf{T_T}$, Eq. (B.5) is obtained.

$$\mathbf{T_T^T}\cdot\mathbf{B} = \mathbf{T_T^T}\cdot\mathbf{T_T}\cdot\mathbf{P_B}^{\mathrm{T}} + \mathbf{T_T^T}\cdot\mathbf{E_B} = \mathbf{T_T^T}\cdot\mathbf{T_T}\cdot\mathbf{P_B}^{\mathrm{T}} \tag*{Eq. (B.5)}$$

Note that, the super scores columns vectors of $\mathbf{T_T}$ are orthogonal to $\mathbf{E_B}$. Substituting Eq. (B.3) and (B.5) in Eq. (B.1), the relation between $\mathbf{W_B}$ and $\mathbf{W_B^*}$ is obtained according to Eq. (B.6).

$$\mathbf{W_B^*} = \mathbf{W_B}\cdot\left(\mathbf{T_T^T}\cdot\mathbf{T_T}\cdot\mathbf{P_B}^{\mathrm{T}}\cdot\mathbf{W_B}\right)^{-1}\cdot\left(\mathbf{T_T^T}\cdot\mathbf{T_B}\right) \tag*{Eq. (B.6)}$$

Note that, regarding the block $\mathbf{X}_{corr}$, the matrix $\mathbf{A} = \mathbf{T_T^T}\cdot\mathbf{T_T}\cdot\mathbf{P}_{\mathbf{X}_{corr}}^{\mathrm{T}}\cdot\mathbf{W}_{\mathbf{X}_{corr}}$ may be rank-deficient as more latent variables could be extracted than the rank of $\mathbf{X}_{corr}$ and, hence, $\mathbf{A}$ would not be invertible. In such a case, $\mathbf{X}_{corr}$ and $\mathbf{T}_{\mathbf{X}_{corr}}$ cannot be directly related by $\mathbf{W}_{\mathbf{X}_{corr}}^*$.

References

[1] J.A. De Smet, Development of Multivariate Specification Limits Using Partial Least Squares Regression, McMaster University, 1993.

[2] A. Paris, C. Duchesne, É. Poulin, Establishing multivariate specification regions for incoming raw materials using projection to latent structure models: comparison between direct mapping and model inversion, Front. Anal. Sci. 1 (2021) 1–15, https://doi.org/10.3389/frans.2021.729732.

[3] C. Duchesne, J.F. MacGregor, Establishing multivariate specification regions for incoming materials, J. Qual. Technol. 36 (2004) 78–94, https://doi.org/10.1080/00224065.2004.11980253.

[4] S. García-Muñoz, Establishing multivariate specifications for incoming materials using data from multiple scales, Chemometr. Intell. Lab. Syst. 98 (2009) 51–57, https://doi.org/10.1016/j.chemolab.2009.04.008.

[5] J.F. MacGregor, Z. Liu, M.J. Bruwer, B. Polsky, G. Visscher, Setting simultaneous specifications on multiple raw materials to ensure product quality and minimize risk, Chemometr. Intell. Lab. Syst. 157 (2016) 96–103, https://doi.org/10.1016/j.chemolab.2016.06.021.

[6] K. Azari, J. Lauzon-Gauthier, J. Tessier, C. Duchesne, Establishing multivariate specification regions for raw materials using SMB-PLS, IFAC-PapersOnLine 48 (2015) 1132–1137, https://doi.org/10.1016/j.ifacol.2015.09.120.

[7] S. García-Muñoz, S. Dolph, H.W. Ward, Handling uncertainty in the establishment of a design space for the manufacture of a pharmaceutical product, Comput. Chem. Eng. 34 (2010) 1098–1107, https://doi.org/10.1016/j.compchemeng.2010.02.027.

[8] P. Facco, F. Dal Pastro, N. Meneghetti, F. Bezzo, M. Barolo, Bracketing the design space within the knowledge space in pharmaceutical product development, Ind. Eng. Chem. Res. 54 (2015) 5128–5138, https://doi.org/10.1021/acs.iecr.5b00863.

[9] D. Palací-López, P. Facco, M. Barolo, A. Ferrer, New tools for the design and manufacturing of new products based on Latent Variable Model Inversion, Chemometr. Intell. Lab. Syst. 194 (2019), https://doi.org/10.1016/j.chemolab.2019.103848.

[10] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, A. Ferrer, Defining multivariate raw material specifications in industry 4.0, Chemometr. Intell. Lab. Syst. 225 (2022), https://doi.org/10.1016/j.chemolab.2022.104563.

[11] ICH Harmonised Tripartite, Guidance for Industry Q8(R2), Pharmaceutical Development, 2009.

[12] 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, https://doi.org/10.1021/ie301214c.

[13] C.M. Jaeckle, J.F. MacGregor, Product design through multivariate statistical analysis of process data, AIChE J. 44 (1998) 1105–1118, https://doi.org/10.1016/0098-1354(96)00182-2.

[14] J. Lauzon-Gauthier, P. Manolescu, C. Duchesne, The Sequential Multi-block PLS algorithm (SMB-PLS): comparison of performance and interpretability, Chemometr. Intell. Lab. Syst. 180 (2018) 72–83, https://doi.org/10.1016/j.chemolab.2018.07.005.

[15] T. Kourti, J.F. MacGregor, Multivariate SPC methods for process and product monitoring, J. Qual. Technol. 28 (1996) 409–428, https://doi.org/10.1080/00224065.1996.11979699.

[16] P. Nomikos, J.F. MacGregor, Multivariate SPC charts for batch monitoring processes, Technometrics 37 (1995) 41–59, https://doi.org/10.2307/1269152.

[17] N.D. Tracy, J.C. Young, R.L. Mason, Multivariate control charts for individual observations, J. Qual. Technol. 24 (1992) 88–95, https://doi.org/10.1080/00224065.1992.12015232.

[18] A. Ferrer, Multivariate statistical process control based on principal component analysis (MSPC-PCA): some reflections and a case study in an autobody assembly process, Qual. Eng. 19 (2007) 311–325, https://doi.org/10.1080/08982110701621304.

[19] C.M. Jaeckle, J.F. MacGregor, Industrial applications of product design through the inversion of latent variable models, Chemometr. Intell. Lab. Syst. 50 (2000) 199–210, https://doi.org/10.1016/S0169-7439(99)00058-1.

[20] F. Yacoub, J.F. MacGregor, Product optimization and control in the latent variable space of nonlinear PLS models, Chemometr. Intell. Lab. Syst. 70 (2004) 63–74, https://doi.org/10.1016/J.CHEMOLAB.2003.10.004.

[21] 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, https://doi.org/10.1021/ie060314g.

[22] 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, https://doi.org/10.1021/ie3032839.

[23] D. Palací-López, J. Borràs-Ferrís, L. Thaise da Silva de Oliveria, Multivariate Six Sigma: A Case Study in Industry 4.0, Processes, vol. 8, 2020, pp. 1–20, https://doi.org/10.3390/pr8091119.

[24] J.F. MacGregor, M.-J. Bruwer, A framework for the development of design and control spaces, J. Pharm. Innov. 3 (2008) 15–22, https://doi.org/10.1007/s12247-008-9023-5.

[25] T. Kourti, J.F. MacGregor, Process analysis, monitoring and diagnosis, using multivariate projection methods, Chemometr. Intell. Lab. Syst. 28 (1995) 3–21, https://doi.org/10.1016/0169-7439(95)80036-9.